RECALL HEDGE UPDATE

EXPERT MODEL WHERE EXPERTS PREDICT
8 LOSS BOUNDS FOR OTHER LOSSES
CONDITIONAL PROBS vs BAYES RULE
HOW DOES BAYES RULE FIT INTO
EXPERT FRAMEWORK
STREAMLINE SETUP (NO LABELS)

FOR $t = 1$ TO $T$ DO

CHOOSE AN EXPERT $i$

GET LOSS VECTOR $L_t \in [0,1]^N$

INCUR LOSS $L_t,i$

GOAL: ACHIEVE SMALL REGRET

TOTAL LOSS OF ALG - TOTAL LOSS OF BEST

ALG I: FOLLOW THE LEADER

- ALWAYS CHOOSE THE BEST EXPERT
  (BRAKE TIES ARBITRARILY)

ADVERSARY:

- CHOSEN EXPERT 1 UNIT OF LOSS
  - ALL OTHERS LOSS 0

LOSS OF ALG

$T \left\lfloor T/n \right\rfloor$

FACTOR OF $n$ OFF
ALGORITHM: HEDGE ALGORITHM
(SIMILAR TO RANDOMIZED WEIGHTED MAJORITY ALGORITHM)

PROBABILISTIC CHOICE OF EXPERT

\( \overline{w}_t \): PROBABILITY VECTOR USED AT TRIAL \( t \)
\( w_{t,i} \): "BELIEVE" AT TRIAL \( t \) THAT \( i \) IS BEST

\( \overline{w}_t = \left( \frac{1}{m}, \ldots, \frac{1}{m} \right) \)

FOR \( t = 1 \) TO \( T \) DO

CHOOSE EXPERT \( i \) WITH PROB. \( w_{t,i} \)
GET LOSS VECTOR \( L_t \)
INCE RLOSS \( L_{t,i} \) OR
EXPECTED LOSS \( \overline{w}_t \cdot L_t = \sum_{i} w_{t,i} L_{t,i} \)

\( -\eta L_{t,i} \)

\( w_{t+1,i} = \frac{w_{t,i} e^{-L_{t,i}}}{Z_t} \)

\( \uparrow \) NORMALIZATION

\( \eta > 0 \): LEARNING RATE

\( e^{-M} \approx \beta \) \( \rightarrow \) \( e^{-\infty} = 0 \)
\[ w_{t+1,i} = \frac{e^{-\eta L_t,i}}{Z_t} \]

As \( \eta \to \infty \), all weight placed on best & hedge becomes "follow the leader".

\[ w_{t+1,i}^\eta = \frac{w_{t,i} e^{-\eta L_t,i}}{Z_t} = \frac{w_{t,i} e^{-\gamma L_{t,i}}}{Z_{w,i} e^{-\gamma L_{t,i}}} \]

\( \eta = 0 \): weights unchanged

\( \eta > 0 \): gradually move weight to experts with low loss

"Soft Min"

\( \eta < 0 \): \( \to \) high loss

In this class
- How to derive updates? 2
- Prove bound!
- Lots of open problems
- Many philosophical issues and open problems
Motivation of Hedge Update

\[ \overline{w}_{t+1} = \inf \left( \frac{1}{\eta} \Delta(\overline{w}, \overline{w}_t) + \overline{w} \cdot \overline{l}_t \right) \]

Updated weight vector

Relative entropy to last weight vector

Loss in last trial

Lagrangian ...

\[ = \frac{\sum w_{t,i} e^{-\eta \overline{l}_{t,i}}}{\sum w_{t,i} e^{-\eta \overline{l}_{t,i}}} \]

Unraveled ...

\[ = \frac{w_{t,i} e^{-\eta \overline{l}_{t,i}}}{\sum_{j} w_{t,j} e^{-\eta \overline{l}_{t,i,j}}} \]

Unraveled motivation: \( u(\overline{w}) \)

\[ \overline{w}_{t+1} = \inf \left( \frac{1}{\eta} \Delta(\overline{w}, \overline{w}_t) + \overline{w} \cdot \overline{l}_t \right) \]

\[ = \frac{w_{t,i} e^{-\eta \overline{l}_{t,i}}}{\sum_{j} w_{t,j} e^{-\eta \overline{l}_{t,i,j}}} \]
POTENTIAL:

Plug solution into objective

\[ U_t(\bar{w}_{t+1}) = -\frac{1}{\eta} \ln \sum_j w_{t+1, j} e^{-\eta L_{t+1, j}} \]

\[ =: P_{t+1}, \text{ Potential} \]

- Log partition function

\[ P_{t+1} \rightarrow \lim_{\eta \rightarrow \infty} \min_j L_{t+1, j} \]
\[
D_{t+1} - P_t = \frac{1}{\eta} \left( \Delta(u, w_t) - \Delta(u, w_{t+1}) \right) + u \cdot L_t \\
\]

\[
= \frac{1}{\eta} \Delta(w_{t+1}, w_t) + w_{t+1} \cdot L_t \\
\]

\[
P_{t+1} = \min_{u} \left\{ \frac{1}{\eta} \Delta(u, w_t) + u \cdot L \leq t \right\} \\
= \frac{1}{\eta} \Delta(w_{t+1}, w_t) + w_{t+1} \cdot L \leq t \\
\]

\[
= \frac{1}{\eta} \left( \Delta(u, w_t) - \Delta(u, w_{t+1}) \right) + u \cdot L \leq t \\
= \frac{1}{\eta} \Delta(u, w_t) + u \cdot L_t \\
= \frac{1}{\eta} \Delta(w_{t+1}, w_t) + w_{t+1} \cdot L_t \\
= \min_{u} \left\{ \frac{1}{\eta} \Delta(u, w_t) + u \cdot L \right\} \\
\sum u_i = 1 \\
\]

SUBOPTIMAL CHOICE \[ u \]
CORRECTION TERM
BOUND:

\[ P_{T+1} - P_t \]

\[ \sum_{t} P_{t+1} - P_t \]

Lecture 1:

\[ \frac{1 - \epsilon^{-\eta}}{\eta} \sum_{t} w_t \cdot L_t \]

TOTAL LOSS OF ALL

\[ \sum_{t} w_t \cdot L_t \leq \frac{1}{1 - \epsilon^{-\eta}} \sum_{t} P_{t+1} \]

\[ \frac{\Delta(w_{T+1}, w_1) + \eta \cdot w_{T+1} \cdot L_t}{1 - \epsilon^{-\eta}} \]

\[ \frac{\Delta(u, w_1) - \Delta(u, w_{T+1}) + \eta \cdot u \cdot L_t}{1 - \epsilon^{-\eta}} \]

EXISTS u s.t.

\[ u \cdot L_{T+1} \leq L^x, \quad \Delta(u, w_1) \leq R^y \]

\[ \Rightarrow \text{TUNED } \eta \]

\[ \leq u \cdot L_{T+1} + \Delta(u, w_1) \]

\[ + \sqrt{2 L^x R^y} \]
MORE SOPHISTICATED EXPERT MODEL

EXPERTS & MASTER ALL. PREDICT

<table>
<thead>
<tr>
<th>PREDICTIONS OF EXPERTS</th>
<th>MASTER PREDICTION</th>
<th>TRUE LABEL</th>
<th>LOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>E₁, E₂, ..., Eₙ</td>
<td>₵ₜ</td>
<td>yₜ</td>
<td>L(yₜ, ̂yₜ)</td>
</tr>
<tr>
<td>TRIAL L x₁₁, x₁₂, x₁ₙ</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

TODAY:  ̂yₜ, yₜ ∈ [0,1]

SQUARE LOSS OF MASTER:

\[ L(y, ̂y) = (y - ̂y)^2 \]

... OF EXPERT i:

\[ L(y, x_i) = (y - x_i)^2 \]

RELATIVE ENTROPY LOSS

\[ L(y, ̂y) = (1-y) \ln \frac{1-\hat{y}}{1-y} + y \ln \frac{\hat{y}}{y} \]

SPECIAL CASE

y ∈ {0,1}, LABEL OF A COIN
\( \hat{y} \) IS PROBABILITY OF COIN

\[ L(1, ̂y) = -\ln( ̂y) \]
\[ L(0, ̂y) = -\ln(1- ̂y) \]  \{ CALLED LOG LOSS \}

WHEN y ∈ {0,1}

HELLINGER LOSS

\[ L(y, ̂y) = \frac{1}{2} \left( (\sqrt{y} - \sqrt{ ̂y})^2 - (\sqrt{1-y} - \sqrt{1- ̂y})^2 \right) \]

ABSOLUTE LOSS

\[ L(y, ̂y) = |y - ̂y| \]

"UNUSUAL LOSS"

SQUARE ROOT TERM NECESSARY
$S = (x_1, y_1), \ldots, (x_t, y_t), \ldots, (x_T, y_T)$

**SEQUENCE OF EXAMPLES**

**WANT BOUNDS OF THE FORM**

$$L_A(S) \leq L_E(S) + c \ln n$$

**# OF EXPERTS**

$$\sum_{t=1}^{T} L(y_t, \hat{y}_t) \leq \sum_{t=1}^{T} L(y_t, x_t, i)$$

**DEPENDS ON LOSS L**

**MAINTAIN ONE WEIGHT PER EXPERT**

$$w_{t,i} = w_{t-1,i} e^{-\eta \sum_{t=1}^{T} L(y_t, x_t, i)} \text{ UNNORMALIZED WEIGHTS}$$

**FOR SIMPLEST CASE**

$$\eta = \frac{1}{L} \text{ NOT POSSIBLE FOR ABSOLUTE LOSS}$$

$$v_{t,i} = \frac{w_{t,i}}{\sum_{j} w_{t,j}} \text{ NORMALIZED WEIGHTS}$$

Initialize the weights to some probability vector $v_{1,i}$; set the parameter $c$ to some positive value.

**Repeat** for $t = 1, \ldots, T$:

1. Receive the instance $x_t$.
2. Output the prediction $\hat{y}_t = v_t \cdot x_t$. A SPECIAL FORM OF PREDICTION
3. Receive the outcome $y_t$.
4. Update the weights by the loss update defined as follows:

$$v_{t+1,i} = v_{t,i} \exp(-L(y_t, x_t, i)/c)/\text{norm}_t$$

where

$$\text{norm}_t = \sum_{i=1}^{n} v_{t,i} \exp(-L(y_t, x_t, i)/c)$$

**Fig. 1.** The Weighted Average Algorithm (WAA) for combining expert predictions
\[ \tilde{V}_{t+1} = \arg \min \sum_{i=1}^{\tilde{V}_{t+1}} c_i \Delta(v, v_t) + \sum_{i=1}^{\tilde{V}_{t+1}} \sum_{q=1}^{y_t} L(y_{t+1} \mid x_t,i) \]

\[ = \frac{V_{t+1} e^{-\frac{1}{T} \sum_{i=1}^{\tilde{V}_{t+1}} L(y_{t+1} \mid x_t,i)}}{\sum_{i=1}^{\tilde{V}_{t+1}}} \]

\[ = V_{t+1} \]

**Alternate Motivation**

\[ \hat{V}_{t+1} = \arg \min \sum_{i=1}^{\hat{V}_{t+1}} \lambda \Delta(v, v_t) + \sum_{i=1}^{\hat{V}_{t+1}} L(y_{t+1} \mid x_t,i) \]

\[ = \frac{\hat{V}_{t+1} e^{-\frac{1}{T} \sum_{i=1}^{\hat{V}_{t+1}} L(y_{t+1} \mid x_t,i)}}{\sum_{i=1}^{\hat{V}_{t+1}}} \]

\[ = V_{t+1} \]

**How can we prove bounds that hold for arbitrary sequences of \((x_t, y_t) \in [0,1]^n \times [0,1]^n \)?**

**Potential:** \(P_{t+1} = U(\hat{V}_{t+1}) = -c_1 \log Z_t\)

**Key Inequality We Need**

\[ L(y_t, y_t) \leq P_{t+1} - P_t \]

Whenever \((x_t, y_t) \notin [0,1]^n \times [0,1]^n \)

\[ \tilde{w}_t \notin (0, \infty)^n \]

**Assume we have inequality**

**By summing over trials we get**

\[ L_{\lambda}(S) = \sum_{t=1}^{T} L(y_t, y_t) \leq \sum_{t=1}^{T} P_{t+1} - P_t \]

\[ = P_{T+1} - P_1 \]
\[ L_A(S) \leq P_{T+1} - \beta_1 \]

\[ \leq -C \ln \sum_{i=1}^{\infty} w_{i,i} e^{-\frac{1}{\theta} L_{E_i}(S)} + C \ln \frac{\sum w_{i,i}}{n} \]

\[ \leq -C \ln \frac{1}{n} e^{-\frac{1}{\theta} L_{E_i}(S)} \]

\[ = L_{E_i}(S) + C \ln n \]

**Proof of Key Inequality:**

\[ L(y_t, v_t, x_t) \leq -C \ln \sum_{i=1}^{n} v_{t,i} e^{-\frac{1}{\theta} L(y_t, x_t, i)} \]

\[ \Rightarrow \quad \frac{1}{C} L(y_t, v_t, x_t) \geq \sum_{i=1}^{n} v_{t,i} e^{-\frac{1}{\theta} L(y_t, x_t, i)} \]

**With**

\[ f_y(x) = e^{-\frac{1}{\theta} L(y, x)} \]

\[ \Rightarrow \quad f_{y_t}(\sum v_{t,i} x_{t,i}) \geq \sum v_{t,i} f_y(x_{t,i}) \]

SUFFICES TO SHOW THAT \( f_y(x) \) CONCAVE
RETURN TO PROOF

\[ f_y(x) = e^{-\frac{1}{c} L_y(x)} \]

NEED TO SHOW THAT \( f_y(x) \) CONCAVE

\[ f'_y(x) = -\frac{1}{c} L'_y(x) e^{-\frac{1}{c} L_y(x)} \]
\[ f''_y(x) = \left( -\frac{1}{c} L'_y(x) - \frac{1}{c^2} L''_y(x) \right) e^{-\frac{1}{c} L_y(x)} \geq 0 \]

THEN

\[ f''_y(x) \leq 0 \quad \text{IFF} \quad c \geq \frac{(L'_y(x))^2}{L''_y(x)} \]

\[ \tilde{c}_L := \sup_{0 < y, x < 1} \frac{(L'_y(x))^2}{L''_y(x)} \]

\[ L_y(x) = (y-x)^2 \quad L'_y(x) = 2(x-y) \quad L''_y(x) = 2 \]

LABEL EXPERT

\[ \tilde{c}_L = \sup_{0 < y, x < 1} \frac{4(y-x)^2}{2} = 2 \]
\[ \hat{y}_t = \bar{y}_t \times \hat{x}_t \]

<table>
<thead>
<tr>
<th>( L )</th>
<th>( C_L )</th>
<th>( \varepsilon^*_L )</th>
</tr>
</thead>
<tbody>
<tr>
<td>REL. ENTR.</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>SQUARE</td>
<td>( \frac{1}{2} )</td>
<td>2</td>
</tr>
<tr>
<td>HEUINGER</td>
<td>0.71</td>
<td>1</td>
</tr>
</tbody>
</table>

Outline of how to get better constant \( C_L \)?

\[ \Delta(y) = \tilde{P}_{t+1} - \tilde{P}_t \]

\[ = -C \ln W_{t+1} + C \ln W_t \]

\[ = \frac{1}{C} \ln \sum_{i=1}^{N} v_{t,i} e^{-\frac{1}{C} L(y,x_{t,i})} \]

- \( C_L \) is \( \max \ C \) s.t.
  
  There always exist \( \hat{y}_t \) for which

\[ L(0, \hat{y}_t) \leq \Delta(0) \]

\[ L(1, \hat{y}_t) \leq \Delta(1) \]

Thus key inequality holds for \( y \in [0, 13] \)

- Now show that key inequality holds for whole interval \( y \in [0, 13] \)
Absolute loss (proofs in WMC paper)

When prediction is $\hat{y}_t = \mathbf{v}_t \cdot \mathbf{x}_t$

$$P_t = -\frac{1}{1-\beta} \ln W_t$$

$$=-\frac{1}{1-\beta} \ln \sum_i w_{t,i} e^{-\ln \frac{1}{\beta} \sum_{i=1}^{t-1} |y_{t,i} - y_{t,i}|}$$

Not Inverses

Key Inequality

$$|y_t - \hat{y}_t| \leq P_{t+1} - P_t$$

$$= -\frac{1}{1-\beta} \ln W_{t+1} e^{-\ln \frac{1}{\beta} \sum_{i=1}^{t-1} |y_{t,i} - y_{t,i}|}$$

$$\sum_{t=1}^{T} |y_t - \hat{y}_t| \leq P_{T+1} - P_1$$

$$\leq \frac{\ln \frac{1}{\beta}}{1-\beta} \sum_{t=1}^{T} |y_{t} - y_{t-1}| + \frac{\ln{m}}{1-\beta}$$

Hedge bound for WMC WMR.

Discrete loss also special

WMC ALG.
WHAT HAVE WE LEARNED?

- **Amortized Analysis for Proving Relative Loss Bounds With Best Expert as Comparator**
  
  - Potential
  - Relative Entropy as Measure of Progress
  - Motivation of Loss Update

\[
\tilde{w}_{t+1} = \min_{\sum w_i = 1} \left( \frac{1}{\eta} \Delta(\tilde{w}, \tilde{w}_t) + \sum_{t} L_{t,i}(\tilde{w}) \right) \\
\tilde{w}_{t+1,i} = \tilde{w}_{t,i} e^{-\eta L_{t,i}}
\]

\[ u_t(w_{t+1}) = p_{t+1} = \frac{1}{\eta} \ln \sum w_{i+1} \cdot e^{-\eta L_{t,i}} \]

**Potential**

\[
= \frac{1}{\eta} \left[ \Delta(\tilde{u}, \tilde{w}_t) - \Delta(\tilde{u}, w_{t+1}) \right] + u \cdot L_{t,i}
= u_t(\tilde{u}) - \frac{1}{\eta} \| \tilde{u}, \tilde{w}_{t+1} \|
\]

Questions - Review of Conditional Probabilities - How Does Bayesian Analysis Fit Into This?
PROBABILITY THEORY

FINITE SET $S$ OF ELEMENTARY EVENTS

$S = \{ (1, 6), (2, 6), (3, 6), (4, 6) \}$

PROBABILITY DISTRIBUTION

$P : S \rightarrow [0, 1]$

- $P(s_i) \geq 0$
- $\sum_{i} P(s_i) = 1$

- EVENT $A$ IS ANY SUBSET OF $S$

$P(A) = \sum_{s_i \in A} P(s_i)$

SUM OVER ELEMENTARY EVENTS IN $A$

- AXIOMS:
- $P(S) = 1$
- $P(A \cup B) = P(A) + P(B)$
  \[ \uparrow \]
  \[ \text{DISJOINT UNION} \]
- $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
A ball is selected from an urn containing two black balls, numbered 1 and 2, and two white balls, numbered 3 and 4. The number and color of the ball is noted, so the sample space is \{(1, b), (2, b), (3, w), (4, w)\}. Assuming that the four outcomes are equally likely, find \(P[A \mid B]\) and \(P[A \mid C]\), where \(A\), \(B\), and \(C\) are the following events:

- \(A = \{(1, b), (2, b)\}\), "black ball selected,"
- \(B = \{(2, b), (4, w)\}\), "even-numbered ball selected," and
- \(C = \{(3, w), (4, w)\}\), "number of ball is greater than 2."

\[
\begin{align*}
P(A \cap B) &= P((2, b)) = 0.25 \\
P(A \cap C) &= P(\emptyset) = 0 \\
P(A \mid B) &= \frac{P(A \cap B)}{P(B)} = \frac{0.25}{0.5} = 0.5 = P(A) \\
P(A \mid C) &= \frac{P(A \cap C)}{P(C)} = \frac{0}{0.5} = 0 
eq P(A)
\end{align*}
\]
In the first case, knowledge of $B$ did not alter the probability of $A$. In the second case, knowledge of $C$ implied that $A$ had not occurred. ■

If we multiply both sides of the definition of $P[A \mid B]$ by $P[B]$ we obtain

$$P[A \cap B] = P[A \mid B]P[B].$$

(2.25a)

Similarly we also have that

$$P[A \cap B] = P[B \mid A]P[A].$$

(2.25b)

---

**INDEPENDENCE OF EVENTS**

If knowledge of the occurrence of an event $B$ does not alter the probability of some other event $A$, then it would be natural to say that event $A$ is independent of $B$. In terms of probabilities this situation occurs when

$$P[A] = P[A \mid B] = \frac{P[A \cap B]}{P[B]}.$$

The above equation has the problem that the right-hand side is not defined when $P[B] = 0$.

We will define two events $A$ and $B$ to be **independent** if

$$P[A \cap B] = P[A]P[B].$$

(2.28)

Equation (2.28) then implies both

$$P[A \mid B] = P[A]$$

(2.29a)

and

$$P[B \mid A] = P[B]$$

(2.29b)

Note also that Eq. (2.29a) implies Eq. (2.28) when $P[B] \neq 0$ and Eq. (2.29b) implies Eq. (2.28) when $P[A] \neq 0$. 
$A = \{(1, b), (2, b)\}$, "black ball selected";
$B = \{(2, b), (4, w)\}$, "even-numbered ball selected"; and
$C = \{(3, w), (4, w)\}$, "number of ball is greater than 2."

Are events $A$ and $B$ independent? Are events $A$ and $C$ independent?
First, consider events $A$ and $B$. The probabilities required by Eq. (2.28)

$$P[A] = P[B] = \frac{1}{2},$$

and

$$P[A \cap B] = P[\{(2, b)\}] = \frac{1}{4}.$$

Thus

$$P[A \cap B] = \frac{1}{4} = P[A]P[B],$$

and the events $A$ and $B$ are independent. Equation (2.29b) gives more insight into the meaning of independence:

$$P[A \mid B] = \frac{P[A \cap B]}{P[B]} = \frac{P[\{(2, b)\}]}{P[\{(2, b), (4, w)\}]} = \frac{1/4}{1/2} = \frac{1}{2}$$

$$P[A] = \frac{P[A]}{P[S]} = \frac{P[\{(1, b), (2, b)\}]}{P[\{(1, b), (2, b), (3, w), (4, w)\}]} = \frac{1/2}{1}.$$

These two equations imply that $P[A] = P[A \mid B]$ because the proportion of outcomes in $S$ that lead to the occurrence of $A$ is equal to the proportion of outcomes in $B$ that lead to $A$. Thus knowledge of the occurrence of $B$ does not alter the probability of the occurrence of $A$.

Events $A$ and $C$ are not independent since $P[A \cap C] = P[\emptyset] = 0$ so

$$P[A \mid C] = 0 \neq P[A] = .5.$$

In fact, $A$ and $C$ are mutually exclusive since $A \cap C = \emptyset$, so the occurrence of $C$ implies that $A$ has definitely not occurred.
Let $B_1, B_2, \ldots, B_n$ be mutually exclusive events whose union equals the sample space $S$ as shown in Fig. 2.14. We refer to these sets as a partition of $S$. Any event $A$ can be represented as the union of mutually exclusive events in the following way:

$$A = A \cap S = A \cap (B_1 \cup B_2 \cup \cdots \cup B_n)$$
$$= (A \cap B_1) \cup (A \cap B_2) \cup \cdots \cup (A \cap B_n).$$

See Fig. 2.14. By Corollary 4, the probability of $A$ is

$$P[A] = P[A \cap B_1] + P[A \cap B_2] + \cdots + P[A \cap B_n].$$

By applying Eq. (2.25a) to each of the terms on the right-hand side, we obtain the theorem on total probability:

$$P[A] = P[A \mid B_1]P[B_1] + P[A \mid B_2]P[B_2] + \cdots + P[A \mid B_n]P[B_n].$$

**Knowledge of** $P(A \mid B_i)$

**And** $P(B_i)$

**Let us compute** $P(A)$
Bayes’ Rule

Let \( B_1, B_2, \ldots, B_n \) be a partition of a sample space \( S \). Suppose that event \( A \) occurs, what is the probability of event \( B_j \)? By the definition of conditional probability we have

\[
P(B_j | A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A | B_j)P(B_j)}{\sum_{k=1}^{n} P(A | B_k)P(B_k)},
\]

where we used the theorem on total probability to replace \( P(A) \). Equation (2.27) is called Bayes’ rule.

\[ P(B_j) \quad \text{PRIOR PROBABILITIES} \]

\[ \text{EXPERIMENT PERFORMED AND} \]

\[ A \quad \text{OCCURRED} \]

\[ P(B_j | A) \quad \text{POSTERIOR PROBABILITIES} \]

\[ \text{GIVEN ADDITIONAL INFORMATION} \]
BAYES

- N EXPERTS / MODELS Eᵢ

- IN EACH TRIAL T WE OBSERVE LABEL Yᵢ T

ASSUMPTION:

- ONE EXPERT Eᵢ GENERATED \( (y_{1}, y_{2}, \ldots, y_{T}) = \bar{y} \)
- PRIOR PROBABILITY OF EXPERT Eᵢ IS \( P(Eᵢ) \)

\( Y \subseteq Y \) FINE T

PROBABILITY OF DATA \( \bar{y} \) GIVEN Eᵢ GENERATED IT:

\[ P(\bar{y} | Eᵢ) \]

DATA LIKELIHOODS

IMPORTANT SPECIAL CASE:

\( y_{1}, y_{2}, \ldots, y_{T} \) ARE GENERATED INDEPENDENTLY AT RANDOM

Thus

\[ P(y_{1}, \ldots, y_{T} | Eᵢ) = \prod_{t=1}^{T} P(y_{t} | Eᵢ) \]

GENERAL CASE

\[ P(y_{1}, \ldots, y_{T} | Eᵢ) = \prod_{t=1}^{T} P(y_{t} | Eᵢ; y_{1}, \ldots, y_{t-1}) \]
For example: Experts are coins $Y = \{0, 1\}$

$P(1|E_i)$: $0.1$, $0.2$, $0.8$, $0.9$

$P(E_i)$: $0.2$, $0.4$, $0.3$, $0.1$

$\vec{y}_3 = (1, 1, 0)$

$P(E_i|\vec{y}_3) = \frac{P(\vec{y}_3|E_i)P(E_i)}{P(\vec{y}_3)}$

Posterior

$\quad \quad = \frac{P(1|E_i)^2 (1-P(1|E_i)) P(E_i)}{P(\vec{y}_3)}$

$P(E_i|\vec{y}_3) \sim 0.1^2 \cdot 0.2 \cdot 0.8 \cdot 0.3 \cdot 0.1$

$\sim 128 \quad 384 \quad 81$

For 1-heavy sequences, posterior will become $\approx \arg\max_i P(1|E_i)$ provided that all $P(E_i) > 0$
LOSS AT TRIAL t

\[ \text{ALG: } - \log p(y_t | y_{t-1}) \]
\[ E_i : - \log p(y_t | E_i, y_{t-1}) = - \log p(y_t | E_i) \]

DEF INDEPENDENCE

"LOG LOSS"

BAYES EXPERT ALG:

ALG MAINTAINS PROBABILITY VECTOR \( \overline{w}_t \)
\[ \overline{w}_t = P(E_0) \text{ PRIORS} \]
Motivation of Bayes Update:

\[ \hat{w}_{t+1} = \min_{\eta} \frac{1}{\eta} \Delta(w, P(E_i)) + \sum_{i} w_i \left( -\ln P(\bar{y}_t | E_i) \right) \]

\[ \eta = 1 \quad \text{Prior Loss of } E_i \]

\[ U_t(\hat{w}) \]

\[ W_{t+1,i} = \frac{P(E_i) e^{-\ln P(\bar{y}_t | E_i)}}{Z_t} \]

\[ = \frac{P(E_i) P(\bar{y}_t | E_i)}{\sum_j P(E_j) P(\bar{y}_t | E_j)} \quad \text{Bayes Rule} \]

\[ = \frac{P(E_i \cap \bar{y}_t)}{P(\bar{y}_t)} \]

\[ = P(E_i | \bar{y}_t) \quad \text{Posterior} \]

\[ \eta+1: \]

\[ w_{t+1,i} = \frac{P(E_i) P(\bar{y}_t | E_i)^\eta}{\sum_j P(E_j) P(\bar{y}_t | E_j)^\eta} \]

\[ \eta = 0 \quad \text{Follow the Leader} \]

\[ \text{All weight on } \text{argmax}_j P(\bar{y}_t | E_i) \]

\[ \eta < 0 \quad \text{Bayes Backwards in Time} \]
\[ U_t (W_{t+1}) = - \log \sum_j p(E_j) e^{-(\ln p(\bar{y}_t | E_j))} \]

\[ = - \log \sum_j p(E_j) p(\bar{y}_t | E_j) \]

\[ = - \log \sum_j p(\bar{y}_t \land E_j) \]

\[ = - \log p(\bar{y}_t) \leq \text{potential has probabilistic meaning} \]

\[ =: p_{t+1} \]
**Loss of Alg as Drop of Potential**

\[- \log P(y_{t+1} \mid y_t) \]

\[= - \log \frac{P(y_{t+1} \mid y_{t-1})}{P(y_t)} \]

\[= - \log P(y_t) - (- \log P(y_{t-1})) \]

\[= P_{t+1} - P_t \]

**Drop of Potential**

**Total Loss of Alg**

\[\sum_t - \log P(y_{t+1} \mid y_t) \]

\[= \sum_t P_{t+1} - P_t \]

\[= P_{T+1} - P_1 \]

\[= - \log (y_T) \]

**Two Kinds of Telescoping**

\[\log P(y_T) = \log P(y_T) - \log P(y_0) \]

\[P(y_T) = \frac{P(y_T)}{P(y_0)} \]

\[= \sum_{t=1}^T \log \frac{P(y_t)}{P(y_{t-1})} \]

**Additive**  

**Multiplicative**
TOTAL LOSS OF EXPERT $E_i$

$$\sum_{t=1}^{T} - \log p(y_t | y_{t-1}, E_i)$$

$$= \sum_{t=1}^{T} - \log \frac{p(y_t | y_{t-1}, E_i)}{p(y_{t-1} | E_i)}$$

$$= \sum_{t=1}^{T} - \log \frac{p(y_{t-1} E_i)}{p(y_{t-1} | E_i)}$$

$$= - \log p(y_T | E_i) - (- \log p(y_0 | E_i))$$

$$\therefore = - \log p(y_T | E_i)$$
SIMPLE BOUNDS

\[
\sum_{t} -\log p(y_t | y_{t-1}) = P_{T+1}
\]

\[
= -\log P(y_T)
\]

THM. OF TOTAL PROB.

\[
= -\log \sum_{i} P(y_T | E_i) P(E_i)
\]

DROP ALL BUT \(i\)

\[
\leq -\log p(y_T | E_i) P(E_i)
\]

\[
= -\log p(y_T | E_i) + \log n
\]

\[
= \sum_{t} -\log p(y_t | E_i, y_{t-1}) + \log n
\]

TOTAL LOSS OF EXPERT \(E_i\)

WORST-CASE BOUND
\[- \log P(\overline{y}_T) = \Delta (\overline{u}, P(E_i)) + \sum_i u_i \left( - \log P(\overline{y}_T | E_i) \right) \]

Choose \( u = \left( \begin{array}{c} 0 \\ 1 \end{array} \right) \) - i

\[- \log P(\overline{y}_T) = \log N - \log P(\overline{y}_T | E_i) - \Delta (\overline{u}, P(E_i | \overline{y}_T)) \]

\( \uparrow \) \text{LAST POSTERIOR}

\text{DROP LAST TERM}

\[ \leq \log N - \log P(\overline{y}_T | E_i) \]

\text{WORST CASE BOUND}
ALG. PREDICTS AT TRIAL t WITH THE DISTRIBUTION:

\[ p(y_t | \tilde{y}_{t-1}) \]

\[ = \sum_i \frac{1}{\lambda_i} P(E_i \cap y_t | \tilde{y}_{t-1}) \quad \text{THM OF TOTAL PROB.} \]

\[ = \sum_i \left( \frac{1}{\lambda_i} P(y_t | E_i, \tilde{y}_{t-1}) \cdot P(E_i | \tilde{y}_{t-1}) \right) \quad \text{PRED. OF EXPERT E_i} \]

\[ = \text{MEAN POSTERIOR (DISTRIBUTION)} \]

\[ p(A \cap B) = p(A | B) \cdot p(B) \]

\[ p(A \cap B | C) = p(A | B, C) \cdot p(B | C) \]

LOSS OF ALG. ≠ CONVEX COMBINATION OF LOSSES OF EXPERT

\[ - \log p(y_t | \tilde{y}_{t-1}) \]

\[ - \log \sum_i \frac{1}{\lambda_i} P(E_i | \tilde{y}_{t-1}) \cdot p(y_t | E_i, \tilde{y}_{t-1}) \]

\[ = \sum_i \left( \frac{1}{\lambda_i} P(E_i | \tilde{y}_{t-1}) \left( - \log p(y_t | E_i, \tilde{y}_{t-1}) \right) \right) \]