LECTURE 9

EXPERT SETTING

SO FAR:

\[ t = 1, \ldots, \]

\[
\text{Pick expert } i \text{ according to prob. vector } W_t, i \quad \# \text{of experts}
\]

\[
\text{Receive loss vector } L_t = [0, 1]^n
\]

\[
\text{Incur loss } L_t, i
\]

\[
\text{Or expected loss } W_t \cdot L_t
\]

\[
W_t, i \sim w_t, i e^{-\eta L_t, i}
\]

\[
\text{Bound: } \sum_{t=1}^{\eta} W_t \cdot L_t \leq \frac{-\sum_{t=1}^{\eta} L_t, i}{1 - \beta}
\]

\[
\text{Tuning}
\]

\[
\text{Loss of all } \leq \text{Loss of best}
\]

\[
\frac{1}{2} \text{Loss of best} \ln n + \ln n
\]

TODAY: SPECIFIC LOSS FUNCTIONS THAT GIVE BOUND OF THE FORM

\[
\text{Loss of Alg} - \text{Loss of Best} = O(\ln n)
\]

LATER: OPTIMAL ALG.
MORE ON EXPERT SETTING WITH DIFFERENT LOSSES

<table>
<thead>
<tr>
<th>TRIALS</th>
<th>$x_{t,1}$</th>
<th>$x_{t,2}$</th>
<th>$x_{t,n}$</th>
<th>PREDICTION</th>
<th>TRUE LABEL</th>
<th>LOSS</th>
</tr>
</thead>
<tbody>
<tr>
<td>$E_1$</td>
<td>$E_2$</td>
<td>$E_n$</td>
<td></td>
<td>$\hat{y}_t$</td>
<td>$y_t$</td>
<td>$L(y_t, \hat{y}_t)$</td>
</tr>
</tbody>
</table>

TODAY: $\hat{y}_t, y_t \in [0,1]$

---

**SQUARE LOSS**

$$L(y, \hat{y}) = (y - \hat{y})^2$$

---

**RELATIVE ENTROPY LOSS**

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

**SPECIAL CASE**

$y, \hat{y} \in [0,1]$  
LABEL OF A COIN  
\$\hat{y}\$ IS PROBABILITY OF COIN  

$$\begin{align*}
L(1, \hat{y}) &= -\ln(\hat{y}) \\
L(0, \hat{y}) &= -\ln(1-\hat{y})
\end{align*}$$

CALLED LOG LOSS WHEN $y, \hat{y} \in [0,1]$

---

**HELLINGER LOSS**

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

---

**ABSOLUTE LOSS**

$$L(y, \hat{y}) = |y - \hat{y}|$$

"UNUSUAL LOSS"

SQUARE ROOT TERM NECESSARY
\[ S = (\bar{x}_1, y_1), ..., (\bar{x}_t, y_t), ..., (\bar{x}_T, y_T) \]

SEQUENCE OF EXAMPLES

WANT BOUNDS OF THE FORM

\[ L_A(S) \leq L_{E_i}(S) + \alpha L \ln n \]

\[ \text{# OF EXPERTS} \]

\[ \frac{1}{T} \sum_{t=1}^{T} L(y_t, \hat{y}_t) \]

\[ \text{DEPENDS ON LOSS L} \]

\[ \frac{1}{T} \sum_{t=1}^{T} L(y_t, x_t, i) \]

\[ \text{UNNORMALIZED WEIGHTS} \]

\[ w_{t,i} = w_{t-1,i} e^{-\eta \sum_{n=1}^{T} L(y_t, x_t, i)} \]

\[ \text{FOR SIMPLEST CASE} \quad \eta = \frac{1}{2L} \quad \text{(NOT POSSIBLE FOR ABSOLUTE LOSS)} \]

\[ v_{t,i} = \frac{w_{t,i}}{\sum_{j=1}^{n} w_{t,j}} \]

\[ \text{NORMALIZED WEIGHTS} \]

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 \( \hat{y}_t = v_t \cdot x_t \). \[ \text{SPECIAL FORM OF PREDICTION} \]
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
How can we prove bounds that hold for arbitrary sequences of \((x_t, y_t) \in \{0,1\}^n \times \{0,1\}\)

\[ P_t = -c \ln W_t \]  
\[ W_t = \sum_i w_{t,i} \]  

**Potential**

**Key Inequality We Need**

\[ L(y_t, \tilde{y}_t) \leq P_{t+1} - P_t \]  
Whenever \((x_t, y_t) \in \{0,1\}^n \times \{0,1\}\)
\[ \tilde{w}_t \in \{0,\ldots,\infty\}^n \]

**Assume we have inequality**

**By summing over trials we get**

\[ L_A(s) = \frac{1}{T} \sum_{t=1}^{T} L(y_t, \tilde{y}_t) \leq \frac{1}{T} \sum_{t=1}^{T} P_{t+1} - P_t \]

\[ = P_{T+1} - P_1 \]
\[ L_A(s) \leq P_{T+1} - P_I \]
\[ \leq -\alpha \ln \sum_{i=1}^{n} w_{i} e^{-\frac{1}{2} L E_i(s)} + \alpha \ln \frac{\mathcal{W}}{2} \]
\[ \leq -\alpha \ln \frac{1}{n} e^{-\frac{1}{2} L E_i(s)} \]
\[ = L E_i(s) + \alpha \ln n \]

**Proof of key inequality:**

\[ L(y_t, v_t, x_t) \leq -\alpha \ln \sum_{i=1}^{n} v_{t,i} e^{-\frac{1}{2} L(y_t, x_{t,i})} \]

\[ \iff \quad e^{\frac{1}{2} L(y_t, v_t, x_t)} \geq \sum_{i=1}^{n} v_{t,i} e^{-\frac{1}{2} L(y_t, x_{t,i})} \]

*With* \( f_y(x) = e^{-\frac{1}{2} L(y, x)} \)

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

SUFFICES TO SHOW THAT \( f_y(x) \) CONCAVE
**Digression:**

**Jensen's Inequality**

**Definition:** A function $f(x)$ is said to be **convex** over an interval $(a, b)$ if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda f(x_1) + (1 - \lambda)f(x_2).$$  
(2.72)

A function $f$ is said to be **strictly convex** if equality holds only if $\lambda = 0$ or $\lambda = 1$.

**Definition:** A function $f$ is **concave** if $-f$ is convex.

![Graphs of Convex and Concave Functions](image)

**Convex**

- $f(x) = x^2$
- $f(x) = e^x$

**Concave**

- $f(x) = \log x$
- $f(x) = \sqrt{x}, x \geq 0$
\[ x_0 = \lambda x_1 + (1-\lambda) x_2 \]

\[ \forall \ 0 \leq \lambda \leq 1 : \ f(\lambda x_1 - (1-\lambda) x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2) \]

For whole line \( \mathbb{R} \)

For segment \( x_1 \leq r \leq x_2 \)
Theorem 2.6.1: If the function $f$ has a second derivative which is non-negative (positive) everywhere, then the function is convex (strictly convex).

Proof: We use the Taylor series expansion of the function around $x_0$, i.e.,

$$f(x) = f(x_0) + f'(x_0)(x - x_0) + \frac{f''(x^*)}{2}(x - x_0)^2 \tag{2.73}$$

where $x^*$ lies between $x_0$ and $x$. By hypothesis, $f''(x^*) \geq 0$, and thus the last term is always non-negative for all $x$.

$$f(x) \geq f(x_0) + f'(x_0)(x - x_0)$$

We let $x_0 = \lambda x_1 + (1 - \lambda)x_2$ and take $x = x_1$ to obtain

$$f(x_1) \geq f(x_0) + f'(x_0)[(1 - \lambda)(x_1 - x_2)]. \tag{2.74}$$

Similarly, taking $x = x_2$, we obtain

$$f(x_2) \geq f(x_0) + f'(x_0)[\lambda(x_2 - x_1)]. \tag{2.75}$$

Multiplying (2.74) by $\lambda$ and (2.75) by $1 - \lambda$ and adding, we obtain (2.72). The proof for strict convexity proceeds along the same lines. \Box
Let $E$ denote expectation. Thus $EX = \Sigma_{x \in x} p(x)x$ in the discrete case and $EX = \int xf(x) \, dx$ in the continuous case.

The next inequality is one of the most widely used in mathematics and one that underlies many of the basic results in information theory.

**Theorem 2.6.2 (Jensen's inequality):** If $f$ is a convex function and $X$ is a random variable, then

$$Ef(X) \geq f(EX).$$

Moreover, if $f$ is strictly convex, then equality in (2.76) implies that $X = EX$ with probability 1, i.e., $X$ is a constant.

**Proof:** We prove this for discrete distributions by induction on the number of mass points. The proof of conditions for equality when $f$ is strictly convex will be left to the reader.

For a two mass point distribution, the inequality becomes

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2),$$

which follows directly from the definition of convex functions. Suppose the theorem is true for distributions with $k - 1$ mass points. Then writing $p'_i = p_i/(1 - p_k)$ for $i = 1, 2, \ldots, k - 1$, we have

$$\sum_{i=1}^{k} p_i f(x_i) = p_k f(x_k) + (1 - p_k) \sum_{i=1}^{k-1} p'_i f(x_i)$$

\[ \overset{\text{INDUCTION}}{\geq} p_k f(x_k) + (1 - p_k) f \left( \sum_{i=1}^{k-1} p'_i x_i \right) \]

\[ \geq f \left( p_k x_k + (1 - p_k) \sum_{i=1}^{k-1} p'_i x_i \right) \]

\[ = f \left( \sum_{i=1}^{k} p_i x_i \right) \]

**Continuous Case** proven using continuity arguments! \[ \Box \]
RETURN TO PROOF

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

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

\[ f'_y(x) = -\frac{1}{\xi} L'_y(x) e^{-\frac{x}{\xi} L_y(x)} \]

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

THEN

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

CONCAVE

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

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

LABEL EXPERT

\[ \overline{c_L} = \sup_{0 < y, x < 1} \frac{4(y - x)^2}{2} = 2 \]
FANLY PRED. $\hat{y}_t = \bar{y}_t \cdot x_t$

<table>
<thead>
<tr>
<th>L</th>
<th>CL</th>
<th>$\hat{C}_L$</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>HELLINGER</td>
<td>0.71</td>
<td>1</td>
</tr>
</tbody>
</table>

OUTLINE OF HOW TO GET BETTER CONSTANT $C_L^2$

$\Delta(y) = P_t - P_{t+1}$

$= -c \ln W_{t+1} + c \ln W_t$

$= -c \ln \sum_{i=1}^{N} e^{-\frac{1}{c} L(y, x_{t,i})}$

$C_L$ 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 $y \in [0, 1]^3$

NOW SHOW THAT KEY INEQUALITY HOLDS FOR WHOLE INTERVAL $y \in [0, 1]^3$
\[ P_t = -\frac{1}{1-\beta} \ln W_t \]

\[ = -\frac{1}{1-\beta} \ln \sum_i W_{t,i} e^{-\frac{1}{2} \gamma (y_{t,i} - x_{t,i})} \]

\[ \text{NOT INVERSES} \]

**KEY INEQ.**

\[ |y_t - \bar{y}_t| \leq P_{t+1} - P_t \]

\[ = -\frac{1}{1-\beta} \ln W_{t+1} e^{-\frac{1}{2} \gamma (y_{t+1} - x_{t+1})} \]

\[ \frac{1}{T} \sum_{t=1}^{T} |y_t - \bar{y}_t| \leq P_{T+1} - P_1 \]

\[ \leq \frac{\ln \frac{1}{\beta}}{1-\beta} \sum_{t=1}^{T} |y_{t} - x_{t,i}| + \frac{\ln n}{1-\beta} \]

**BOUND FOR WMC WMR ALG.**

**DISCRETE LOSS ALSO SPECIAL**

**WM ALG.**
WHAT HAVE WE LEARNED?

- AMORTIZED ANALYSIS FOR PROVING RELATIVE LOSS BOUNDS

- POTENTIAL

- RELATIVE ENTROPY AS MEASURE OF PROGRESS

- MOTIVATION OF LOSS UPDATE

\[
\overline{w}_{t+1} = \min_{\sum w_i = 1} \left( \Delta(\overline{w}, \overline{w}_t) + \eta \sum_{t} \overline{L}_{x_t, \overline{w}} \right)
\]

\[
w_{t+1, i} = \frac{w_{t, i} e^{-\eta \overline{L}_{x_t, i}}}{2t}
\]

\[
U_t(w_{t+1}) = p_{t+1} = -\ln \sum w_{t+1, i} e^{-\eta \overline{L}_{x_t, i}}
\]

- POTENTIAL

NEXT:

- REVIEW OF CONDITIONAL PROBABILITIES

- HOW DOES BAYESIAN ANALYSIS FIT INTO THIS?

- PROJECTION METHODS
PROBABILITY THEORY

FINITE SET $S$ OF ELEMENTARY EVENTS

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

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:
  - $P(S) = 1$
  - $P(A \cup B) = P(A) + P(B)$
    \[ \text{DISJOINT UNION} \]
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
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."

\[
P(A \cap B) = P(\{(2, b)\}) = .25
\]
\[
P(A \cap C) = P(\emptyset) = 0
\]
\[
P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{.25}{.5} = .5 = P(A)
\]
\[
P(A \mid C) = \frac{P(A \cap C)}{P(C)} = \frac{0}{.5} = 0 \neq P(A)
\]
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

$$P(A \cap B) = P(A \mid B)P(B). \quad (2.25a)$$

Similarly we also have that

$$P(A \cap B) = P(B \mid A)P(A). \quad (2.25b)$$

---

**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

$$P(A \cap B) = P(A)P(B). \quad (2.28)$$

Equation (2.28) then implies both

$$P(A \mid B) = P(A) \quad (2.29a)$$

and

$$P(B \mid A) = P(B) \quad (2.29b)$$

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. $\blacksquare$
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 $P(B_i)$ lets 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

$$P(B_j | A) = \frac{P(A \cap B_j)}{P(A)} = \frac{P(A | B_j)P(B_j)}{\sum_{k=1}^{n} P(A | B_k)P(B_k)} \tag{2.27}$$

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

$P(B_j)$ PRIOR PROBABILITIES

EXPERIMENT PERFORMED AND
A OCCURRED

$P(B_j | A)$ POSTERIOR PROBABILITIES
GIVEN ADDITIONAL INFORMATION
MODEL 2

1. ONE EXPERT $E_i$ GENERATES THE LABELS $y_1, y_2, \ldots, y_T$
2. PRIOR PROBABILITY OF EXPERT $E_i$ IS $P(E_i)$

$y \in Y$ FINITE

IMPORTANT SPECIAL CASE:

$y_1, y_2, \ldots, y_T$ ARE GENERATED INDEPENDENTLY AT RANDOM ACCORDING TO $P(y_1|E_i)$

Thus $P(y_1, y_2, \ldots, y_T|E_i) = \prod_{t=1}^{T} P(y_t|E_i)$

FOR EXAMPLE: EXPERTS ARE COINS $Y = \{0, 1\}$

<table>
<thead>
<tr>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(1</td>
<td>E_i)$</td>
<td>.1</td>
<td>.2</td>
</tr>
<tr>
<td>$P(E_i)$</td>
<td>.2</td>
<td>.4</td>
<td>.3</td>
</tr>
</tbody>
</table>

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

$P(E_i|\bar{y}_3) = \frac{P(\bar{y}_3|E_i)P(E_i)}{P(\bar{y}_3)} = \frac{P(111|E_i)^3 (1-P(111|E_i)) P(E_i)}{P(\bar{y}_3)}$

POSTERIOR

<table>
<thead>
<tr>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$P(E_i</td>
<td>\bar{y}_3)$</td>
<td>.1$^2$.9</td>
<td>.2$^2$.8</td>
</tr>
<tr>
<td>~</td>
<td>1.8</td>
<td>128</td>
<td>384</td>
</tr>
</tbody>
</table>
**Bayes Alg. Assumes Assumptions Are Correct and Predicts at Trial t with Distribution**

\[ P(y_t | \bar{y}_{t-1}) = \frac{P(y_t \cap \bar{y}_{t-1})}{P(\bar{y}_{t-1})} \]

*where* \( P(\bar{y}_0) = 1 \)

**Called The Predictive Distribution**

**Log Loss at Trial t**

**Alg:** \[- \log P(y_t | \bar{y}_{t-1}) \]

**E_i:** \[- \log P(y_t \cap E_i | \bar{y}_{t-1}) \rightarrow - \log P(y_t | E_i) \]

**Total Log Loss of Alg:**

\[ \sum_{t=1}^{T} - \log P(y_t | \bar{y}_{t-1}) = \sum_{t=1}^{T} - \log \frac{P(y_t)}{P(\bar{y}_{t-1})} \]

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

**Telescoping**

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

= \[ - \log P(y_T) \]
TOTAL LOSS OF EXPERT $E_i$:

$$\frac{1}{T} \sum_{t=1}^{T} -\log P(y_t | \bar{y}_{t-1}, E_i)$$

$$= \sum_{t=1}^{T} -\log \frac{P(y_t \wedge \bar{y}_{t-1}, 1 | E_i)}{P(\bar{y}_{t-1}, 1 | E_i)}$$

$$= \sum_{t=1}^{T} -\log \frac{P(\bar{y}_t | E_i)}{P(\bar{y}_{t-1} | E_i)}$$

$$= -\log P(\bar{y}_t | E_i) - \left( -\log \frac{P(\bar{y}_0 | E_i)}{1} \right)$$

$$= -\log P(\bar{y}_t | E_i)$$

TOTAL LOSS OF ALG.

$$-\log P(\bar{y}_1) = -\log \sum_{i} P(\bar{y}_1 | E_i)$$

$$= -\log \sum_{i} P(E_i) P(\bar{y}_1 | E_i)$$

$$\leq -\log P(\bar{y}_1 | E_i) + \log \frac{1}{P(E_i)}$$

TOTAL LOSS OF $i$-TH EXPERT
HOW IS THIS RELATED TO THE CANONICAL EXPERTS ALG?

\[ w_{t+1, i} = w_{t, i} e^{-\eta L_{t, i}} \quad \text{UNNORMALIZE WEIGHTS} \]

NOW \( \eta = 1 \), \( L_{t, i} = -\log p(y_t \mid E_i, Y_{t-1}) \)

SO \( e^{-\eta L_{t, i}} = p(y_t \mid E_i, Y_{t-1}) \)

\[ w_{t+1, i} = w_{t, i} p(y_t \mid E_i, Y_{t-1}) \]

\[ = w_{t, i} \prod_{q=1}^{t} \frac{p(y_q \mid E_i, Y_{q-1})}{p(E_i)} \]

\[ = w_{t, i} \prod_{q=1}^{t} \frac{p(y_q \cap E_i)}{p(Y_{q-1} \cap E_i)} \]

\[ = p(y_t \cap E_i) \]

NORMALIZED WEIGHTS

\[ v_{t+1, i} = \frac{w_{t+1, i}}{\sum_{j} w_{t+1, j}} = \frac{p(y_t \cap E_i)}{\sum_{j} p(y_t \cap E_j)} = \frac{p(y_t \cap E_i)}{p(y_t)} = p(E_i \mid y_t) \]

\[ \text{POSTERIOR} \]

ALSO:

\[ v_{t+1, i} = \frac{w_{t, i} p(y_t \mid E_i, Y_{t-1})}{\sum_{k} w_{t, k} p(y_t \mid E_k, Y_{t-1})} = \frac{v_{t, i} p(y_t \mid E_i, Y_{t-1})}{\sum_{k} v_{t, k} p(y_t \mid E_k, Y_{t-1})} \]

FOR LOG-LOSS THE \( e^{-\eta} \) LOSS UPDATE IS BAYES RULE
Algorithm predicts at trial $t$ with the distribution:

$$p(y_t | y_{t-1})$$

$$= \sum_i p(E_i \cap y_t | y_{t-1})$$

$$= \sum_i \frac{p(y_t | E_i, y_{t-1}) \cdot p(E_i | y_{t-1})}{\text{PRED. OF EXPERT $E_i$}} \cdot \text{POSTERIOR}$$

$$= \text{MEAN POSTERIOR DISTRIBUTION}$$

$$p(A \cap B) = p(A | B) \cdot p(B)$$

$$p(A \cap B | C) = p(A | B, C) \cdot p(B | C)$$
Potential used in online learning literature:

$$\text{Pot}_{t+1} = -\frac{1}{\eta} \log \sum_{i=1}^{T} w_{t+1,i}$$

$$= - \log \sum_{i=1}^{T} \hat{p}(y_t \wedge E_i)$$

$$= - \log \hat{p}(y_t)$$

**Key inequality**

**Loss of Alg. At Trial t**

$$-\log \hat{p}(y_t | \overline{y}_{t-1}) = - \log \frac{\hat{p}(y_t)}{\hat{p}(\overline{y}_{t-1})}$$

$$= - \log \hat{p}(y_t) - (- \log \hat{p}(\overline{y}_{t-1}))$$

$$= \text{Pot}_{t+1} - \text{Pot}_t$$

**Bound again:**

$$\sum_{t=1}^{T} -\log \hat{p}(y_t | \overline{y}_{t-1}) = -\log \hat{p}(\overline{y}_T) + \text{Pot}_{T+1}$$

$$\leq -\log \sum_{i} \hat{p}(\overline{y}_T | E_i) \hat{p}(E_i)$$

$$\leq -\log \hat{p}(\overline{y}_T | E_i) \hat{p}(E_i)$$

$$= - \log \hat{p}(\overline{y}_T | E_i) + \log n$$

$$= \sum_{t=1}^{T} -\log \hat{p}(y_t | E_i, \overline{y}_{t-1}) + \log n$$

**Total loss of $$E_i$$**
Motivation of Update \( \eta = 1 \)

\[
\bar{v}_{t+1} = \min_{\sum v_i = 1} \Delta(\bar{v}, P(E_i)) + \sum_i v_i (-\ln P(E_i | y_t))
\]

Prior \quad Loss of \( E_i \)

\[
U_t(\bar{v})
\]

\[
V_{t+1,i} = \frac{P(E_i) e^{-(-\ln P(E_i | y_t))}}{Z_t}
\]

\[
= \frac{P(E_i) P(E_i | y_t)}{\sum_j P(E_j) P(E_j | y_t)}
\]

\[
= \frac{P(E_i \cap y_t)}{P(y_t)}
\]

\[
= P(E_i | y_t)
\]

\[
U_t(P(E_i | y_t)) = -\text{Log}(y_t)
\]

\[
= -\text{Log} \sum_i P(E_i \cap y_t)
\]
\[
\Delta (\bar{v}, P(E_i)) - \Delta (\bar{v}, P(E_i | \bar{y}_t)) = \sum \nu_i \ln P(\bar{y}_t | E_i) \\
= \sum \nu_i \ln \frac{P(E_i | \bar{y}_t)}{P(E_i)} \\
= \sum \nu_i \ln \frac{P(E_i) P(\bar{y}_t | E_i)}{P(\bar{y}_t)} \\
= \sum \nu_i \ln P(\bar{y}_t | E_i) - \ln Z_t + \sum \nu_i \ln P(\bar{y}_t | E_i) \\
= -\ln Z_t \\
= -\ln \sum_i P(E_i) P(\bar{y}_t | E_i) \\
= -\ln P(\bar{y}_t) \\
= \text{Pot}_{t+1}
\]