LECTURE 2 a) ANALYSIS OF HEDGE

\[ P_t = -\ln \sum_i w_{t,i} e^{-\eta L_{t-1,i}} \]

\[ \uparrow \text{DUE TO NORMALIZATION} \]

\[ P_{t+1} - P_t = -\ln \sum_i w_{t,i} e^{-\eta L_{t+1,i}} + \ln \sum_i w_{t,i} e^{-\eta L_{t-1,i}} \]

\[ = -\ln \frac{\sum_i w_{t,i} e^{-\eta L_{t+1,i}} e^{-\eta L_{t,i}}}{\sum_i w_{t,i} e^{-\eta L_{t-1,i}}} \]

\[ = -\ln \sum_i w_{t,i} e^{-\eta L_{t,i}} \]

\[ \geq -\ln \sum_i w_{t,i} (1 - (1-e^{-\eta}) L_{t,i}) \]

\[ e^{\eta x} \leq 1 - (1-e^{-\eta}) x \]

\[ x \in [0,1] \]

\[ \ln (1-x) \leq x \]

\[ \geq (1-e^{-\eta}) \bar{w}_t \cdot \bar{L}_t \]

DROP OF POTENTIAL

\[ \geq (1-e^{-\eta}) \text{ LOSS OF ALG.} \]
Summing over $t$

\[ \sum_{t=1}^{T} P_{t+1} - P_t \geq (1 - e^{-\eta}) \sum_{t=1}^{T} W_t \cdot L_t \]

Lower bound

\[ \sum_{t=1}^{T} P_{t+1} - P_t = P_{T+1} - P_1 \]

\[ = - \ln \sum_{i=1}^{N} w_{i,t} e^{-\eta L_{t+1} \hat{i}} \]

\[ \leq - \ln w_{i,t} e^{-\eta L_{t+1} \hat{i}} \]

\[ = - \ln w_{i,t} + \eta L_{t+1} \hat{i} \]

Upper bound

\[ \sum_{t=1}^{T} W_t \cdot L_t \leq \frac{\ln \sum_{i=1}^{N} w_{i,t} + \eta L_{t+1} \hat{i}}{1 - e^{-\eta}} \]

L_{\text{ALG}}

If $w_{i,t} = (\frac{1}{n}, \ldots, \frac{1}{n})$ THEN $\ln \frac{1}{w_{i,t}} = \ln n$

- CAN HANDLE LOTS OF EXPERTS

\[ L_{\text{ALG}} \leq \frac{\ln n}{1 - e^{-\eta}} + \frac{\eta L_{t+1} \hat{i}}{1 - e^{-\eta}} \]

\[ \eta = 1.58 \ln n + 1.58 L_{t+1} \hat{i} \]

↑ WANT 1

If $\eta$ TUNED AS FUNCTION OF $n$ & $\hat{l}$ THEN

\[ \sum_{t=1}^{T} W_t L_t \leq \operatorname{ms}L_{t+1} \hat{i} + \sqrt{2 \hat{l} \ln \frac{1}{\ln n}} + \ln n \]

\[ \hat{l}^{*} \leq \hat{l} \]

IF $\hat{l}^{*} \leq \hat{l}$ REGRET BOUND
BIG PICTURE
- WE USED EXPONENTIAL WEIGHTS AND SOFTMIN TO ACHIEVE REGRET BOUNDS
- EXPECTED LOSS BOUNDS HOLD FOR ARBITRARY SEQUENCES
- EXPECTATION WRT INTERNAL RANDOMIZATION OF ALG
- LOGARITHMIC DEPENDANCE ON # OF EXPERTS Y.
  TYPICAL FOR "MULTIPICATIVE" UPDATES

QUESTIONS:
- LOWER BOUNDS ²?
- MOTIVATION OF UPDATES ²?
- WHERE DID THE POTENTIAL COME FROM ²?
- WHAT ABOUT OTHER LOSS FUNCTIONS ²?
- COMPARE AGAINST BEST LINEAR COMBINATION OF EXPERTS ².
Lots of "stupid" experts are "specialized" combined to something better

Later: Boosting
  - iteratively builds
  small linear combination
  of weak hypothesis

For fun: Bug Machine

Many stupid bugs better
than one smart bug

- Variety is asset in changing environment
LECTURE 2) EXPERT SETTING

SO FAR:

\[ t = 1, \ldots, T \]

PICK EXPERT \( i \) ACCORDING TO
PROB. VECTOR \( w_{t,i} \)

RECEIVE LOSS VECTOR \( L_t = [0, 1]^n \)

INCURR LOSS \( L_{t,i} \)
OR EXPECTED LOSS \( w_{t} \cdot L_t \)

\[ w_{t,i} \sim \text{Unif}(\frac{1}{n}) \cdot L_{t,i} \]

BOUND:

\[
\frac{1}{T} \sum_{t=1}^{T} w_{t,\cdot} L_{t,\cdot} \leq \frac{-\ln \frac{1}{n} + \ln \frac{1}{\beta}}{1-\beta} \]

\[ \downarrow \text{TUNING} \]

\[ \text{LOSS OF ALL} \leq \text{LOSS OF BEST} \]

\[ + \sqrt{2 \ln n} \]

\[ + \ln n \]

LATER: OPTIMAL ALGO.

TODAY: SPECIFIC LOSS FUNCTIONS
THAT GIVE BOUNDS OF THE FORM

\[ \text{LOSS OF ALL} - \text{LOSS OF BEST} = O \left( \ln n \right) \]
MORE ON EXPERT SETTING WITH DIFFERENT LOSSES

\[ \begin{array}{cccccc}
E_1 & E_2 & \cdots & E_n & \text{PREDICTION} & \text{TRUE LABEL} & \text{LOSS} \\
\text{TRIAL T} & x_{t1} & x_{t2} & \cdots & x_{tn} & \hat{y}_t & y_t & L(y_t, \hat{y}_t)
\end{array} \]

TODAY: \( \hat{y}_t, y_t \in \{0,1\} \)

\[ \text{SQUARE LOSS:} \]
\[ L(y, \hat{y}) = (y - \hat{y})^2 \]

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

SPECIAL CASE
\( y \in \{0,1\} \)
\( \hat{y} \) is probability of coin
\[ \begin{align*}
L(0, \hat{y}) &= -\ln(\hat{y}) \\
L(1, \hat{y}) &= -\ln(1-\hat{y})
\end{align*} \]
\( \text{Called log loss when } y \in \{0,1\} \)

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

\[ \text{ABSOLUTE LOSS} \]
\[ L(y, \hat{y}) = |y - \hat{y}| \]
\("\text{UNUSUAL LOSS}\"
\(\text{SQUARE ROOT TERM NECESSARY}\)
$S = (x_{11}, y_1), \ldots, (x_{1T}, y_T), \ldots, (x_{T1}, y_T)$

SEQUENCE OF EXAMPLES

WANT BOUNDS OF THE FORM

\[
L_A(S) \leq L_{E_i}(S) + c \ln n \quad \# \text{OF EXERTS}
\]

\[
= \sum_{t=1}^T L(y_t, \hat{y}_t) + \sum_{t=1}^T L(y_t, x_{t,i}) \quad \text{DEPENDS ON LOSS L}
\]

MENTAIN ONE WEIGHT PER EXPERT

\[
w_{t,i} = w_{1,i} e^{-\eta \sum_{t'=1}^{t-1} l(y_{t'}, x_{t',i})} \quad \text{UNNORMALIZED WEIGHTS}
\]

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

\[
v_{t,i} = \frac{w_{t,i}}{\sum_j w_{t,j}} \quad \text{NORMALIZED WEIGHTS}
\]

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 \( \hat{y}_t = v_t \cdot x_t \).
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 \((x_t, y_t) \in [0,1]^n \times [0,1]\)

\[
P_t = -\lambda \ln W_t \quad \text{potential}
\]

\[
W_t = \sum_i w_{t,i}
\]

\text{Key inequality we need}

\[
L(y_t, y_t') \leq P_{t+1} - P_t
\]

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

\[
L_A(s) = \sum_{t=1}^T L(y_t, y_t') \leq \sum_{t=1}^T P_{t+1} - P_t
\]

\[
= P_{T+1} - P_1
\]
\[ L_A(s) \leq P_{T+1} - \beta_i \]
\[ = -c \ln \sum_{i=1}^{n} w_i, i e^{-\frac{c}{L_{E_i}(s)}} + c \ln \frac{W_i}{\hat{w}_i} \]
\[ \leq -c \ln \frac{1}{n} e^{-\frac{c}{L_{E_i}(s)}} \]
\[ = L_{E_i}(s) + c \ln n \]

**Proof of Key Inequality:**

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

*With* \[ f_y(x) = e^{-\frac{c}{L(y, x)}} \]
\[ \Rightarrow \quad f_y(\sum v_{t,i} x_{t,i}) \geq \sum v_{t,i} f_y(x_{t,i}) \]

*Suffices to show that* \[ f_y(x) \] *concave*
**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.

---

**Convex**

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

**Concave**

- \( f(x) = \log x \)
- \( f(x) = \sqrt{x}, \ x \geq 0 \)

**Positive 2nd Derivative**

**Negative 2nd Derivative**
\[ x_0 = \lambda x_1 + (1-\lambda) x_2 \]

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

For whole line \( \lambda \geq 1 \)

For segment \( \lambda \in (0,1) \)
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.,

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

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

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

Similarly, taking $x = x_2$, we obtain

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

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. □
Let $E$ denote expectation. Thus $EX = \sum_{x \in X} p(x)x$ 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

$$Ef(X) \geq f(EX). \quad (2.76)$$

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

$$p_1 f(x_1) + p_2 f(x_2) \geq f(p_1 x_1 + p_2 x_2), \quad (2.77)$$

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

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

$$\quad \overset{\text{INDUCTION}}{\geq p_k f(x_k) + (1 - p_k) f\left(\sum_{i=1}^{k-1} p'_i x_i\right)} \quad (2.79)$$

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

$$= f\left(\sum_{i=1}^{k} p_i x_i\right) \overset{\text{BINARY CASE}}{= f\left(\sum_{i=1}^{n} p_i x_i\right)}$$

CONTINUOUS CASE PROVEN USING CONTINUITY ARGUMENTS!

\[ \square \]

SEE EXAMPLES ON PAGE 5
$f_y(x) = e^{-\frac{1}{c} L_y(x)}$

NEED TO SHOW THAT $f_y(x)$ CONCAVE

$f'_y(x) = -\frac{1}{c} L'_y(x) e^{-\frac{1}{c} L_y(x)}$

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

THEN

$f''_y(x) \leq 0 \iff c \geq \frac{(L'_y(x))^2}{L''_y(x)}$

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

$L_y(x) = (y-x)^2$ \hspace{1cm} $L'_y(x) = 2(x-y)$ \hspace{1cm} $L''_y(x) = 2$

LABEL EXPERT

$\tilde{c}_L = \sup_{0 < y, x < 1} \frac{4(y-x)^2}{2} = 2$
Fancy Pred. $\frac{y_t}{x_t} = \bar{y} \cdot x_t$

<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 $C_L^2$;

\[ \Delta(y) = P_{thi} - P_e \]

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

\[ = -c \ln \sum_{i=1}^{N} v_{t,i} e^{-\frac{1}{2} L(y, x_{t,i})} \]

- $C_L$ 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 $y \in [0, 1]^3$
Absolute Loss (Proofs in WM Paper)
When prediction is $\hat{y}_t = \mathbf{v}_t \cdot \mathbf{x}_t$

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

$$
= -\frac{1}{1-\beta} \ln \sum_i w_{t,i} e^{-\ln \frac{1}{\beta} \sum_i^t 1 y - x_{t,i}}
$$

Not Inverses

Key Inequality

$$
|y_t - \hat{y}_t| \leq \rho_{t+1} - \rho_t
$$

$$
= -\frac{1}{1-\beta} \ln W_{t+1} e^{-\ln \frac{1}{\beta} \sum_i 1 y - x_{t,i}}
$$

$$
\sum_t |y_t - \hat{y}_t| \leq \rho_{T+1} - \rho_1
$$

$$
\leq \frac{\ln \frac{1}{\beta}}{1-\beta} \sum_t |y_t - x_{t,i}| + \frac{\ln m}{1-\beta}
$$

Discrete Loss also special WM Alg.

Hedge BOUND FOR WMC

WR ALG.
WHAT HAVE WE LEARNED?

- AMORTIZED ANALYSIS FOR PROVING RELATIVE LOSS BOUNDS

- POTENTIAL

- RELATIVE ENTROPY AS MEASURE OF PROGRESS

- MOTIVATION OF LOSS UPDATE

$$\tilde{w}_{t+1} = \min_{\sum w_{t+1}} \left( \frac{\Delta(\tilde{w}, \tilde{w}) + \eta \sum L_{t+1} \tilde{w}}{U_t(w)} \right)$$

$$w_{t+1,i} = w_{t,i} e^{-\eta L_{t+1} i \frac{2}{2t}}$$

- $$U_t(w_{t+1}) = p_{t+1} = -\ln \sum w_{t+1,i} e^{-\eta L_{t+1} i}$$

POTENTIAL

QUESTIONS - REVIEW OF CONDITIONAL PROBABILITIES

- HOW DOES BAYESIAN ANALYSIS FIT INTO THIS?

- PROJECTION METHODS