LECTURE 1: 290C
ADVANCED ML W07

BATCH VS ON-LINE

1) - GET BATCH OF DATA PTS (TRAINING SET)
   - FIT MODEL
   - GET TEST PT. & PREDICT W. MODEL

PROBLEM: IN MANY PRACTICAL SETTINGS
DATA IS CONTINUOUSLY CHANGING

2) IN THIS CLASS
   LOOP
   - GET NEXT PT.
   - PREDICT BASED ON CURRENT MODEL
   - UPDATE MODEL
BATCH:
- Training and test data generated by same distribution
- If model class not too complex and enough examples, model that does best on training data not too much worse on test data

ON-LINE:
- All is in flux
- No statistical assumptions
- Still can bound "regret" =

\[
\text{Total loss of on-line} - \text{Total loss of best off-line chosen in hind sight}
\]

- Bounds hold for arbitrary sequences of examples
WHAT YOU WILL LEARN

- TECHNIQUES FOR DERIVING & ANALYSING
  ON-LINE LEARNING ALGS
  - BREGMAN DIVERGENCES
  - BREGMAN PROJECTIONS

  - HOW TO PROVE REGRET BOUNDS
  OR RELATIVE LOSS BOUNDS

RECURRING THEME

  - HOW TO COMBINE MANY RULES OF THUMB
    - EXPERT SETTING
    - BOOSTING
    - BUG MACHINE :-)}
OUTLINE:

TODAY:
- EXPERT SETTING
- VARIOUS METHODS FOR PROVING RELATIVE LOSS BOUNDS

LECTURE 2:
- APPLICATIONS
- DISK SPIN DOWN CACHING
- HOW TO MEASURE ON-LINENESS
- SHIFTING EXPERT SETTING
  - LONG TERM MEMORY
- HW1 (PRACTICAL)

LECTURE 3:
- ANALYSIS OF SHIFTING
  - VERSIONS OF RELATIVE ENTROPY
  - VICINITY LEMMAS

LECTURE 4:
- HW1 IS DUE
- HW2 (THEORETICAL)
On-Line Learning

<table>
<thead>
<tr>
<th>experts</th>
<th>(E_1)</th>
<th>(E_2)</th>
<th>(E_3)</th>
<th>(E_n)</th>
<th>prediction</th>
<th>true label</th>
<th>loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>day 1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>day 2</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>day 3</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>day (t)</td>
<td>(x_{t,1})</td>
<td>(x_{t,2})</td>
<td>(x_{t,3})</td>
<td>(x_{t,n})</td>
<td>(\hat{y}_t)</td>
<td>(y_t)</td>
<td>(</td>
</tr>
</tbody>
</table>

Protocol of the Master Algorithm

For \(t = 1\) To \(T\) Do

Receive \(x_t \in \{0, 1\}^n\)
Predict \(\hat{y}_t \in \{0, 1\}\)
Get label \(y_t \in \{0, 1\}\)
Incur loss \(|y_t - \hat{y}_t| \in \{0, 1\}\)
CASE 1: THERE IS A CONSISTENT EXPERT

GIVEN SEQUENCE \((x_t, y_t)\) s.t

\[ x_{t,r} = y_t \text{ for all } t \]

LOSS OF OFF-LINE COMPARATOR IS ZERO

NOISE-FREE CASE
Halving Algorithm

- Predicts with majority

- If mistake then number of consistent experts is halved
A run of the Halving Algorithm

<table>
<thead>
<tr>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
<th>$E_5$</th>
<th>$E_6$</th>
<th>$E_7$</th>
<th>$E_8$</th>
<th>majority</th>
<th>true label</th>
<th>loss</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>0</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>1</td>
<td>x</td>
<td>x</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>x</td>
<td>x</td>
<td>x</td>
<td>↑</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

consistent

For any sequence with a consistent expert, HA makes $\leq \log_2 n$ mistakes

GAME AGAINST NATURE (ADVERSARY)

WHICH CHOOSES THE $x_t$ & $y_t$

IF THERE IS ONE CONSISTENT EXPERT
THEN ALG. $\leq \log_2 n$ MISTAKES
What if no expert is consistent?

For any sequence $S = (x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T)$
- $L_A(S)$ is total loss of alg. A and
- $L_i(S)$ is the total loss of expert $E_i$

**Relative Loss**
Want bounds of the form:

$$\forall S : \ L_A(S) \leq a \ \min_i L_i(S) + b \ \log(n)$$

where $a, b$ are constants

Bounds loss of algorithm relative to
loss of best expert

$$a = 1$$

$$L_A(S) - \min_i L_i(S) \ \text{CALLED REGRET}$$
Can't wipe out experts!
One weight per expert

**Weighted Majority Algorithm**  [LW]

- Predicts with larger side

- Weights of wrong experts are multiplied by $\beta \in (0, 1]$

- $\beta$ is *fitness factor*

- $HA: \beta = 0$
Number of mistakes of the WM algorithm

\[ M_{t,i} = \# \text{ of mistakes of } E_i \text{ before trial } t \]

\[ w_{t,i} = \beta^{M_{t,i}} \text{ weight of } E_i \text{ at trial } t \]

\[ W_t = \sum_{i=1}^{n} w_{t,i} \text{ total weight at trial } t \]

Minority \( \leq \frac{1}{2} W_t \)

Majority \( \geq \frac{1}{2} W_t \)

If no mistake then

\[ W_{t+1} \leq 1 \ W_t \]
If mistake then
majority multiplied by $\beta$

$$W_{t+1} \leq \frac{1}{2} W_t + \beta \frac{1}{2} W_t$$

$$= \frac{1 + \beta}{2} W_t$$

$$W_{T+1}^{\text{total final weight}} \leq \left( \frac{1 + \beta}{2} \right)^M W_1$$

$$W_{T+1} = \sum_{j=1}^{n} w_{T+1,j} = \sum_{j=1}^{n} \beta^{M_j} \geq \beta^{M_i}$$

$$\left( \frac{1 + \beta}{2} \right)^M \frac{W_1}{n} \geq \beta^{M_i}$$
\[ M \leq \frac{-\ln \beta}{\ln \frac{2}{1+\beta}} M_i + \frac{1}{\ln \frac{2}{1+\beta}} \ln n \]

\[ M \leq \frac{2.63}{\ln \frac{2}{1+\beta}} \min_{i \in \mathcal{M}^*} M_i + \frac{2.63}{\ln n} \ln n \]

\[ \beta = \frac{1}{e} \min_{i \in \mathcal{M}^*} M_i + \frac{2.63}{\ln n} \ln n \]

For all sequences, loss of the master algorithm is comparable to the loss of the best expert.

Relative loss bounds \[ [F] \]

*With fancy choice of \( \beta \) that depends on \( n, M^* \):

\[ M \leq 2M^* + 2\sqrt{M^* \ln(n)} + \log_2 n \]

\( \uparrow \)

*Necessary for deterministic prediction*
SUMMARY OF ANALYSIS METHOD

\[ w_{t,i} = 1 \]
\[ w_{t+1,i} = w_{t,i} \beta^M_{t,i} \quad \text{UNNORMALIZED WEIGHTS} \]

UNNORMALIZED POTENTIAL:

\[ P_{t+1} = - \sum_{i} \beta^{M_{t+1,i}} \]

\[ \frac{P_{t+1}}{P_t} = \begin{cases} 
\frac{1 + \beta}{2} & \text{IF MISTAKE IN TRIAL} \\
1 & \text{IF NO MISTAKE}
\end{cases} \]
STREAMLINE SETUP (NO LABELS)

FOR $t = 1$ TO $T$ DO

CHOOSE AN EXPERT $i$

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

INCRUR 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 $i$ UNIT OF LOSS
- ALL OTHERS LOSS $0$

$O$

LOSS OF ALG

$T$

LOSS OF BEST

$\left\lfloor \frac{T}{n} \right\rfloor$
ALG II: RANDOMIZED WEIGHTED MAJORITY

PRABABILISTIC CHOICE OF EXPERT

$\mathbf{w}_t$: PROBABILITY VECTOR USED AT TRIAL $t$

$w_{t,i}$ "BELIEVE" AT TRIAL $t$ THAT $i$ IS BEST

$\mathbf{w}_t = \left( \frac{1}{n}, \ldots, \frac{1}{n} \right)$

FOR $t = 1$ TO $T$ DO

CHOOSE EXPERT $i$ WITH PROB. $w_{t,i}$

GET LOSS VECTOR $L_t$

INCLUD LOSS $L_{t,i}$ OR

EXPECTED LOSS $\mathbf{w}_t \cdot L_t = \sum_i w_{t,i} L_{t,i}$

$-\eta L_{t,i}$

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

↑ NORMALIZATION

$\eta > 0$ LEARNING RATE

$e^{-\eta} = \beta \xrightarrow{} 1$

$e^{-\infty} = 0$

$0 \quad \eta$
\[ w_{t+1,i} = \frac{e^{-\eta \cdot L_{t,i}}}{Z_t} \]

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

\[ w_{t+1,i} = \frac{w_{t,i}e^{-\eta \cdot L_{t,i}}}{Z_t} = \frac{w_{t,i}e^{-\eta \cdot L_{t,i}}}{\sum w_{t,i}e^{-\eta \cdot L_{t,i}}} \]

\( \eta = 0 \): weights unchanged

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

"Soft Min"

\( \eta < 0 \): \(\Rightarrow\) high loss
ANALYSIS:

Potentials: \[ P_t = -\ln \sum_i w_{t,i} e^{-\eta \mathcal{L}_{t,i}} \]

Due to normalization:

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

\[ = -\ln \frac{\sum_i w_{t,i} e^{-\eta \mathcal{L}_{t,i}} e^{-\eta \mathcal{L}_{t,i}}}{\sum_i w_{t,i} e^{-\eta \mathcal{L}_{t-1,i}}} \]

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

\[ \geq -\ln \sum_i w_{t,i} \left( 1 - (1-e^{-\eta}) \mathcal{L}_{t,i} \right) \]

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

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

\[ \Rightarrow -\ln \left( 1 - (1-e^{-\eta}) \mathcal{L}_{t} \cdot \mathcal{L}_{t} \right) \]

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

\[ \Rightarrow (1-e^{-\eta}) \mathcal{L}_{t} \cdot \mathcal{L}_{t} \]

Drop of potential:

\[ \geq (1-e^{-\eta}) \text{ loss of alg.} \]
SUMMING OVER t

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

LOWER BOUND

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

\[ = 0 \]

= \[ - \ln \sum_{i} w_{1,i} e^{-\eta} L_{T,i} \]

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

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

UPPER BOUND

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

If \( w_{1,i} = (\frac{1}{n} - \frac{1}{n}) \) THEN \( \ln \frac{1}{w_{1,i}} = \ln n \)

CAN HANDLE LOTS OF EXPERTS

\( \eta = 1 \) GIVES BOUNDS OF THE FORM

\[ L_{ALL} \leq \eta \text{ LOSS OF BEST} + b \ln n \]

\( a, b > 1 \)

IF \( \eta \) TUNED AS FUNCTION OF \( n \) & \( \hat{L} \) THEN REGRET BOUND

\[ \sum_{t=1}^{T} w_t L_t \leq \min \left\{ L_{T,i} \right\} + \sqrt{2 \hat{L} \ln n} + \ln n \]

\( \hat{L} \)

IF \( L^* \leq \hat{L} \)
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 **
  TYPICAL FOR "MULTIPlicative" UPDATES

QUESTIONS:
- LOWER BOUNDS ?
- MOTIVATION OF UPDATES ?
- WHERE DID THE POTENTIAL COME FROM ?
- WHAT ABOUT OTHER LOSS FUNCTIONS ?
- COMPARE AGAINST BEST LINEAR COMBINATION OF EXPERTS ?
Master

Weighted Majority

E₁ E₂ E₃ E₄

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