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 is 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
### 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.

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

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>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</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 PREDICTION VECTOR $\vec{x}_t$ AND LABEL $y_t$

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 $E_i$

Want bounds of the form:

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

where $a, b$ are constants

Bounds loss of algorithm relative to loss of best expert

$$\alpha = 1$$

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

**Weighted Majority Algorithm**

- 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

\[ M_{t-1,i} = \text{# of mistakes of } E_i \text{ before trial } t \]

\[ w_{t-1,i} = \beta^{M_{t-1,i}} \text{ weight of } E_i \text{ at trial } t, \quad w_{0,i} = 1 \]

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

Minority \( \leq \frac{1}{2} W_{t-1} \)

Majority \( \geq \frac{1}{2} W_{t-1} \)

If no mistake then

minority multiplied by \( \beta \)

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

\[
W_t \leq 1 \left( \frac{1}{2} W_{t-1} \right)_{\text{minority}} + \beta \left( \frac{1}{2} W_{t-1} \right)_{\text{majority}}
\]

\[
= \frac{1 + \beta}{2} W_{t-1}
\]

\[
\frac{W_T}{\text{total final weight}} \leq \left( \frac{1 + \beta}{2} \right)^M W_0
\]

\[
W_T = \sum_{j=1}^{n} w_{T,j} = \sum_{j=1}^{n} \beta^{M_j} \geq \beta^{M_i}
\]

\[
\left( \frac{1 + \beta}{2} \right)^M \frac{W_0}{n} \geq \beta^{M_i}
\]
\[ M \leq \frac{-\ln \beta}{\ln \frac{2}{1+\beta}} M_i + \frac{1}{\ln \frac{2}{1+\beta}} \ln n \]

\[ M \leq \left( \frac{2.63}{a} \min_{i} M_i \right) + \frac{2.63 \ln n}{b} \]

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

Relative loss bounds

With fancy choice of \( \beta \) that depends on \( n, M_{\alpha} \):

\[ M \leq 2M_{\alpha} + 2\sqrt{M_{\alpha} \ln(N)} + \log_2 n \]

\( \uparrow \)

NECESSARY
FOR DETERMINISTIC
PREDICTION

[F]
STREAMLINE SETUP (NO LABELS)

FOR $t = 1$ TO $T$ DO

CHOOSE AN EXPERT $i$

GET LOSS VECTOR $\hat{L}_t \in [0,1]^n$

INCUR LOSS $L_{t,i}$

GOAL: ACHIEVE SMALL REGRET

TOTAL LOSS OF ALG - TOTAL LOSS OF BEST

ALG I: DETERMINISTIC FOLLOW THE LEADER

- ALWAYS CHOOSE AN EXPERT OF MINIMAL LOSS

ADVERSARY:

- CHOSEN EXPERT $i$ UNIT OF LOSS
- ALL OTHERS LOSS 0

(T IS # OF TRIALS)

\[
\begin{array}{c|c|c|c|c|c}
0 & 1 & 1 & 1 & 1 & 1 \\
1 & 1 & 1 & 1 & 1 & 1 \\
2 & 1 & 1 & 1 & 1 & 1 \\
\vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\
T & \frac{T}{n} & \frac{T}{n} & \frac{T}{n} & \frac{T}{n} & \frac{T}{n} \\
\end{array}
\]

LOSS OF ALG \sim n LOSS OF BEST

\[
\frac{\text{LOSS OF ALG}}{T} \leq \frac{\text{LOSS OF BEST}}{\lfloor T/n \rfloor}
\]
ALG III: PERTURB LOSSES OF EXPERTS
PREDICT W. PERTURBED LEADER

ALG IV: HEDGE ALGORITHM
(SIMILAR TO RANDOMIZED WEIGHTED MAJORITY ALGORITHM)

PROBABILISTIC CHOICE OF EXPERT

$W_{t-1}$: probability vector used at trial $t$

$w_{t-1,i}$ "believe" at trial $t$ that $i$ is best

$W_0, i = \left( \frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n} \right)$

$W_t, i = \frac{w_{t-1,i} e^{-\eta l_t,i}}{Z_t}$

$\uparrow$ NORMALIZATION

$\eta > 0$: LEARNING RATE

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

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

\( \eta = 0 \) weights unchanged

\( \eta > 0 \) gradually move weight to experts w. low loss "soft min"

\( \eta < 0 \) \( \to \) high loss "softmax"

Next class:

If \( \eta \) tuned as function of \( n \) & \( \hat{L} \) then

\[ \sum_{t=1}^{T} w_{t, i} L_t - \min_{\hat{L}} \sum_{t=1}^{T} i \leq V \] (1/n + 1/n)

\( \hat{L} \) if \( L^* \leq \hat{L} \)

Loss of alg. - loss of best

Regret