I) ENTROPY
   RELATIVE ENTROPY
   - MEASURE OF PROGRESS
     IN ANALYSIS OF EXPERT ALGS

II) LINEAR REGRESSION
   - DERIVING UPDATES WITH
     SQUARED EUCLIDEAN DISTANCE
     AND RELATIVE ENTROPY AS
     REGULARIZER

III) LEARNING W. LINEAR THRESHOLD FUNCTIONS

IV) LOGISTIC REGRESSION

WANT TO SENT SYMBOL X ON CHANNEL

\[
\begin{align*}
X & \quad P(X=x_i) & \quad -\log P(x_i) \\
 x_1 & \quad \frac{1}{2} & \quad 1 \\
x_2 & \quad \frac{1}{4} & \quad 2 \\
x_3 & \quad \frac{1}{8} & \quad 3 \\
x_4 & \quad \frac{1}{8} & \quad 3
\end{align*}
\]

\{ \text{BITS} \}

\[\uparrow\]

\text{MEASURE OF SURPRISE}
\[
- \log 1 = 0 \quad \text{NO SURPRISE}
\]
\[
- \log 0 = \infty \quad \text{INFINITE I}\]
\[
- \log 2^i = i \quad \text{BITS}
\]
Entropy equals expected surprise

\[ H(X) = \sum_{i} p(x_i) \log_2 \frac{1}{p(x_i)} \]

\[ = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 3 \]

\[ = 1 \frac{3}{4} \]

Huffman code

\[
\begin{array}{c|c}
X & 1 \\
X_1 & \frac{1}{2} \\
X_2 & \frac{1}{4} \\
X_3 & \frac{1}{6} \\
X_4 & \frac{1}{6} \\
\end{array}
\]

Loop

Pick smallest two
Combine both into one
Sum their probs

Entropy = expected code length

\[ H(X) = \log_2 3 = 1.58 \text{ bits} \]

Expected code length

\[ = \frac{1}{3} (1 + 2 + 2) \]

\[ = \frac{5}{3} \]

\[ = 1.66 \text{ bits} \]
**Code:** Assigns symbols a bitstring (codeword)

- Any sequence of codewords must be uniquely decodable

\[ L(C) = \sum_i p(x_i) L_C(x_i) \quad \text{expected code length} \]

**Optimal Code** \( C^* \)

- Minimum \( L(C) \)

**Thm:** \( H(X) \leq L(C^*) \leq H(X) + 1 \)

**Thm:** Huffman codes are optimal

**More Info:**

First five chapters of Cover & Thomas
\[
\Delta (\vec{P}, \vec{q}) = \sum_i p_i \ln \frac{p_i}{q_i}
\]

**Symbol Used for Divergences**

\[
\sum_i p_i \ln \frac{1}{q_i} - \sum p_i \ln \frac{1}{p_i}
\]

**Expected Code Length of Best Codebook for \(\vec{q}\)**

**Expected Code Length of Best Codebook for \(\vec{p}\)**

**Expectations Are WRT \(\vec{p}\)**

\[
\Delta (\vec{p}, \frac{1}{n}) = \sum_i p_i \ln \frac{p_i}{\frac{1}{n}}
\]

\[
= \sum_i p_i \ln p_i + \sum p_i \ln n
\]

\[
= \ln n - H(\vec{p})
\]

\[
\geq 0
\]

\[
= 0 \text{ at corners of simplex}
\]

**Maple Plots**

\[
\Delta (\vec{P}, \vec{q}) \quad \text{not too steep at boundary}
\]

\[
\Delta (\vec{P}, \vec{q}) \quad \text{very steep}
\]

BARRIERS FOR SIMPLEX
DUAL USE OF RELATIVE ENTROPY

1) AS REGULARIZER IN MOTIVATION OF UPDATE

2) AS MEASURE OF PROGRESS IN ANALYSIS

STREAMLINED SETUP OF EXPERT SETTING

\[ \bar{w}_t = \left( \frac{1}{N}, \ldots, \frac{1}{N} \right) \quad \text{NORMALIZED WEIGHT VECTOR} \]

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

CHOOSE EXPERT \( i \) WITH PROBABILITY \( w_{t,i} \)

GET LOSS VECTOR \( \overline{l}_t \in \{0,1\}^N \)

INCUR LOSS \( L_{t,i} \) OR

EXPECTED LOSS \( \overline{w}_t \cdot \overline{l}_t \)

\[
\begin{align*}
W_{t+1,i} &= W_{t,i} e^{ -\eta L_{t,i} } \\
\frac{1}{Z_t} \\
e^{-\eta} = \beta
\end{align*}
\]
Motivation of update:

\[ \overline{w}_{t+1} = \underset{\Sigma w_i = 1}{\arg \min} \left( \Delta (\overline{w}, \overline{w}_t) + \eta \overline{w} \cdot L_t \right) \]

Updated weight vector

Relative entropy to last weight vector

Loss in last trial

Minimizing convex function subject to linear constraints

Constraint

\[ L (\overline{w}, \lambda) = \Sigma w_i \ln \frac{w_i}{\overline{w}_{t,i}} + \eta \overline{w} \cdot L_t + \lambda \left( \Sigma w_i - 1 \right) \]

Lagrangian

Lagrangian multiplier

Find optimal solution by setting all derivatives of \( L \) to zero

\( \frac{\partial L}{\partial \lambda} = \Sigma \lambda w_i - 1 = 0 \) (\(^\ast\))

Our constraint

\( \frac{\partial L}{\partial w_i} = \ln \frac{w_i}{\overline{w}_{t,i}} + w_i \frac{1}{\overline{w}_i} + \eta L_{t,i} + \lambda = 0 \)

\[ \ln \frac{w_i}{\overline{w}_{t,i}} = -\eta L_{t,i} - \lambda - 1 \]

\[ w_i = \overline{w}_{t,i} e^{-\eta L_{t,i}} \]

Enforce (\(*\)): \[ \lambda = \frac{1}{\Sigma i \overline{w}_{t,i} e^{-\eta L_{t,i}}} \]

Final result: \[ \overline{w}_{t+1,i} = \frac{w_{t,i} e^{-\eta L_{t,i}}}{\Sigma j \overline{w}_{t,j} e^{-\eta L_{t,j}}} \]
\[ \text{UNRAVEL: } w_{t+1, i} = \frac{w_{t, i} e^{-\eta L_{t, i}}}{\sum_j w_{t, j} e^{-\eta L_{t, j}}} \]

Also: \[ \bar{w}_{t+1} = \bar{w}_{t} + \eta \nabla_{w_t} f (\Delta (\bar{w}, \bar{w}_t) + \eta \bar{w} \cdot L_{t} ) \]

\[ \sum_{w_i = 1} \]

\[ \text{II \hspace{1cm} PROGRESS TOWARDS } \bar{w} \]

\[ \Delta(\bar{u}, \bar{w}_t) - \Delta(\bar{u}, \bar{w}_{t+1}) \]

\[ = \sum_i u_i \ln \frac{w_{t+1, i}}{w_{t, i}} \]

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

\[ = -\eta \sum_i u_i \bar{L}_{t, i} - \ln Z_t \]

\[ = -\eta \bar{u} \cdot \bar{L}_t - \ln Z_t \frac{\sum_i w_{t, i} e^{-\eta L_{t, i}}}{q} \]

\[ - \ln q \geq 1 - q, \text{ FOR } q \in (0, 1] \]

\[ \geq -\eta \bar{u} \cdot \bar{L}_t + 1 - \sum_i w_{t, i} e^{-\eta L_{t, i}} \]

\[ e^{-\eta L} \leq 1 - (1 - e^{-\eta}) L, \text{ FOR } L \in [0, 1] \]

\[ \geq -\eta \bar{u} \cdot \bar{L}_t + 1 - \sum_i w_{t, i} (1 - (1 - e^{-\eta}) L_{t, i}) \]

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

LOSS OF COMPARATOR

LOSS OF ALL
\[
\Delta(\tilde{u}, \tilde{w}_1) - \Delta(\tilde{u}, \tilde{w}_{T+1}) = \sum_{t=1}^{T} \Delta(\tilde{u}, \tilde{w}_t) - \Delta(\tilde{u}, \tilde{w}_{t+1}) \quad \text{TELESCOPING}
\]

\[
\geq -\eta \sum_{t=1}^{T} \tilde{u}_t \cdot L_t + (1 - e^\eta) \sum_{t=1}^{T} \tilde{w}_t \cdot L_t
\]

\[
\sum_{t=1}^{T} \tilde{w}_t \cdot L_t \leq \frac{\eta \sum_{t=1}^{T} \tilde{u}_t \cdot L_t + \Delta(\tilde{u}, \tilde{w}_1) - \Delta(\tilde{u}, \tilde{w}_{T+1})}{1 - e^{-\eta}} \tag{*}
\]

holds for all \(\tilde{L}_t \in \{0, 1\}^N\), \(\tilde{u} \in \mathbb{R}^N\)

Tuning \(\eta\) as function of \(\min \sum_{t=1}^{T} L_t = L^*\) and \(n\)

\[
\sum_{t=1}^{n} \tilde{w}_t \tilde{L}_t \leq L^* + \sqrt{2L^* \ln n} + \ln N
\]
HOW IS THIS RELATED TO POTENTIAL PROOF?

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

WE SHOWED IN LECTURE 1:

\[ P_{t+1} - P_t \geq (1 - e^{-\eta}) \bar{w}_t \cdot \bar{L}_t \quad (*) \]

MOTIVATION:

\[ \bar{w}_{t+1} = \min_{\sum w_{t,i}} \left( \Delta(\bar{w}, \bar{w}_i) + \eta \bar{w} \cdot \bar{L}_t \right) \]

\[ = \frac{w_{t,i} e^{-\eta L_{t,i}}}{z_t} \]

CLAIM: \( U_t(\bar{w}_{t+1}) = P_{t+1} = - \ln z_t \)

PROOF:

\[ \sum_{i} \frac{w_{t,i} e^{-\eta L_{t,i}}}{z_t} + \eta \bar{w}_{t+1} \bar{L}_t = \]

\[ = - \eta \bar{w}_{t+1} \bar{L}_t + \sum w_{t,i} \ln \frac{1}{z_t} + \eta \bar{w}_{t+1} \bar{L}_t \]

\[ = - \ln z_t \]

\[ \square \]
POTENTIAL BASED PROOF:

**Sum (y)**

\[ \Delta(\bar{u}, \bar{w}) + \eta \ u \cdot L_{\leq T} \]

\[ \geq \ \text{min} \left( \Delta(\bar{u}, \bar{w}) + \eta \ \bar{w} \cdot L_{\leq T} \right) \]

\[ \Sigma w_i = 1 \]

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

\[ \rightarrow 0 \]

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

TELES COPING

**Lecture 1**

\[ \geq \left(1 - e^{-\gamma}\right) \Sigma_{t=1}^{T} \bar{w}_t \cdot L_t \]

\[ \Sigma_{t=1}^{T} \bar{w}_t \cdot L_t \]

\[ \leq \frac{\Delta(\bar{u}, \bar{w}) + \eta \ u \cdot L_{\leq T}}{1 - e^{-\gamma}} \]

SAME BOUND (x)

ON PAGE 8
LINEAR REGRESSION

\[ w \cdot x \]

ON-LINE MODEL

INITIALIZE \( \tilde{w} \).

FOR \( t = 1, 2, \ldots \) DO

RECEIVE INSTANCE \( \tilde{x}_t \)

PREDICT \( \hat{y}_t = \tilde{w} \cdot \tilde{x}_t \)

GET LABEL \( y_t \)

INCUR LOSS \( (\hat{y}_t - y_t)^2 \)

UPDATE \( \tilde{w}_t \) TO \( \tilde{w}_{t+1} \).

GOAL:

THE TOTAL LOSS OF THE ON-LINE ALG.

SHOULD NOT BE MUCH LARGER THAN THE
TOTAL LOSS OF THE OFF-LINE ALG.

\[ \forall S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_T, y_T) \} \]

\[ \frac{1}{T} \sum_{t=1}^{T} (\tilde{w}_t \tilde{x}_t - y_t)^2 \] - \[ \inf_{w} \frac{1}{T} \sum_{t=1}^{T} (w \tilde{x}_t - y_t)^2 \]

TOTAL LOSS OF ON-LINE ALG

TOTAL LOSS OF OFF-LINE ALG.
HOW TO DERIVE UPDATES:

\[ w_{t+1} = \text{INF}_{\tilde{w}} \left( \frac{1}{2} \| \tilde{w} - w_t \|^2 + \eta \frac{1}{2} (\tilde{w} \cdot x_t - y_t)^2 \right) \]

DIVERGENCE TO LAST WT

LOSS ON LAST EXAMPLE

TRADE-OFF PARAMETER \( \eta > 0 \)

\[ \frac{\partial}{\partial w} = (w - w_t) + \eta (\tilde{w} \cdot x_t - y_t) \cdot x_t = 0 \]

THUS \( \tilde{w}_{t+1} = \tilde{w}_t - \eta (\tilde{w}_{t+1} \cdot x_t - y_t) \cdot x_t \)

NEW OLD

\( \text{LEARNING RATE} \)

\( \approx \tilde{w}_t - \eta (w_{t+1} x_t - y_t) \cdot x_t \)

WIDROW-HOFF UPDATE

GRADIENT DESCENT
Let's use a different divergence

\[ w_{t+1,i} = \inf_{w \geq 0} \left( \sum_i w_i \ln \frac{w_i}{w_{t,i}} + w_{t,i} - w_i + \eta \frac{1}{2} (\bar{w} \cdot \bar{x}_t - y_t)^2 \right) \]

Relative Entropy
Generalized to arbitrary non-negative weights

\[ \frac{\partial}{\partial w_{t,i}} = \ln \frac{w_{t,i}}{w_{t+1,i}} + 1 - 1 + \eta (\bar{w} \cdot \bar{x}_t - y_t) x_{t,i} = 0 \]

\[ \ln w_{t+1,i} = \ln w_{t,i} - \eta (w_{t+1,i} x_t - y_t) x_{t,i} \]

\[ w_{t+1,i} = w_{t,i} e^{-\eta (w_{t+1,i} x_t - y_t) x_{t,i}} \]

\[ \approx w_{t,i} e^{-\eta (w_t x_t - y_t) x_{t,i}} \]

Unnormalized exponentiated gradient algo

EGLU
GENERAL LOSS

$$GD:$$

$$w_{t+1} = \text{argmin}_w \left( \frac{1}{2} \|w - w_{t}\|^2 + \eta \ L_{yt}(w, x_t) \right)$$

$$w_{t+1} = w_t - \eta \ L_{yt}'(w_{t+1}, x_t) x_t$$

$$\approx w_t - \eta \ L_{yt}'(w_t, x_t) x_t$$

EGU

$$w_{t+1} = \text{argmin}_w \left( \sum w_i \ln \frac{w_i}{w_{t,i}} + w_{t,i} - w_i + \eta \ L_{yt}(w, x_t) \right)$$

$$w > 0$$

$$\ln w_{t+1} = \ln w_t - \eta \ L_{yt}'(w_{t+1}, x_t) x_t$$

$$\approx \ln w_t - \eta \ L_{yt}'(w_t, x_t) x_t$$

FUNDAMENTAL:

IN GD WE LIN. COMB. OF PAST EXAMPLES
IN EGU \( \ln w_t \) \""
LEARNING WITH LINEAR THRESHOLD FUNCTIONS

PERCEPTRON ALG \((w, \eta, \theta)\)

FOR \(t = 1, 2, \ldots\)

GET \(x_t\)

\[
\hat{y}_t = \begin{cases} 
+1 & \text{IF } w_t x_t \geq \theta \\
-1 & \text{L} 
\end{cases}
\]

GET BINARY LABEL \(y_t \in \{\pm 1\}\)

IF \(\hat{y}_t = y_t\) THEN \(w_{t+1} = w_t\)
ELSE \(w_{t+1} = w_t + \eta y_t x_t\)

WINNOW \((w, \eta, \theta)\)

IF \(\hat{y}_t = y_t\) THEN \(w_{t+1} = w_t, \quad \eta y_t x_{t,i}\)
ELSE \(w_{t+1} = w_t, \quad \eta y_t x_{t,i}\)

UPDATE ONLY IF MISTAKE "CONSERVATIVE UPDATES"
LINEAR HINGE LOSS

\[ y_t = 1 \]

\[ y_t = -1 \]

GD = UPDATE OF PERCEPTRON

EGU = "WINNOW"

\[ L' y_t (w_t \cdot x_t) = 0 \] WHEN NO MISTAKE

\[ L' y_t (w_t \cdot x_t) = -y_t \] IF MISTAKE

WINNOW

\[ \ldots \] \[ w_{t+1,i} = w_{t,i} e^{-\eta (-y_t) x_t, i} \]

NEGATIVE GRADIENT
NOW wt PROB. VECTOR:

\[ w_{t+1} = \arg \min_w \left( \sum_{i} w_i \ln \frac{w_i}{w_{t,i}} + \eta \frac{1}{2} (\bar{w} \cdot x_t - y_t)^2 \right) \]

\[ L(w, \lambda) = \sum_{i} w_i \ln \frac{w_i}{w_{t,i}} + \eta \frac{1}{2} (\bar{w} \cdot x_t - y_t)^2 \]
\[ + \lambda \left( \sum_{i} w_i - 1 \right) \]

\[ \frac{\partial L}{\partial w_i} = \ln \frac{w_i}{w_{t,i}} + 1 + \eta (w \cdot x_t - y_t) x_{t,i} + \lambda = 0 \]

\[ \ln w_{t+1,i} = \ln w_{t,i} - \eta (w_{t,i} \cdot x_t - y_t) x_{t,i} - 1 - \lambda \]
\[ - \eta (w_{t+1,i} \cdot x_t - y_t) x_{t,i} - 1 - \lambda \]
\[ w_{t+1,i} = w_{t,i} e^{-\eta (w_{t+1,i} \cdot x_t - y_t) x_{t,i} - 1 - \lambda} \]

SINCE \( \sum_{i} w_{t+1,i} = 1 \)

\[ e^{-1-\lambda} \text{ is } 1/\text{NORMALIZATION} \]

\[ w_{t+1,i} = \frac{w_{t,i} e^{-\eta (w_{t+1,i} \cdot x_t - y_t) x_{t,i}}}{\sum_{j} w_{t,j} e^{-\eta (w_{t+1,j} \cdot x_t - y_t) x_{t,j}}} \]
\[ w_{t+1, i} = \frac{w_{t, i} e^{-\eta (w_t \cdot x_t - y_t) x_t, i}}{\sum_j w_{t, j} e^{-\eta (w_t \cdot x_t - y_t) x_t, j}} \]

**EXponentiated Gradient Alg.**

**FOR GENERAL LOSS** \( L_{yt}(w \cdot x_t) : \)

\[ w_{t+1, i} = \frac{w_{t, i} e^{-\eta L_{yt}(w_t \cdot x_t) x_t, i}}{\text{NORMALIZE}}. \]

\[ \Rightarrow \text{FOR HINGELOSS E6 BECOMES NORMALIZED VERSION OF WINNOW} \]

\[ \Rightarrow \text{DE} \ L_{yt}(w \cdot x_t) := \sum_i w_i L_{yt}(x_t, i) \uparrow \text{PROB. VECT.} \]

\[ \text{CONVEX COMBINATION OF LOSSES OF EXPERTS E_i} \]

**E6 BECOMES EXPERT UPDATE:**

\[ w_{t+1, i} = w_{t, i} e^{-\eta L_{yt}(x_t, i)} \text{ implicit} \]

\[ \text{NORMALIZE.} \]
How to get negative weights

Algorithm $\text{EG}_L^+(U,(s^+,s^-),\eta)$

**Parameters:**
- $L$: a loss function from $\mathbb{R} \times \mathbb{R}$ to $[0, \infty)$,
- $U$: the total weight of the weight vectors,
- $s^+$ and $s^-$: a pair of start vectors in $[0,1]^N$, with $\sum_{i=1}^{N} (s^+_i + s^-_i) = 1$, and
- $\eta$: a learning rate in $[0, \infty)$.

**Initialization:** Before the first trial, set $w^+_i = Us^+$ and $w^-_i = Us^-.$

**Prediction:** Upon receiving the $t$th instance $x_t$, give the prediction
\[
\hat{y}_t = (w^+_t - w^-_t) \cdot x_t.
\]

**Update:** Upon receiving the $t$th outcome $y_t$, update the weights according to the rules
\[
\begin{align*}
w^+_{t+1,i} &= U \cdot \frac{w^+_{t,i} r^+_i}{\sum_{j=1}^{N} (w^+_{t,j} r^+_j + w^-_{t,j} r^-_j)} \quad (3.8) \\
w^-_{t+1,i} &= U \cdot \frac{w^-_{t,i} r^-_i}{\sum_{j=1}^{N} (w^+_{t,j} r^+_j + w^-_{t,j} r^-_j)} \quad (3.9)
\end{align*}
\]

where
\[
\begin{align*}
r^+_t &= \exp\left(-\eta L'_y(\hat{y}_t) U x_{t,i}\right) \quad (3.10) \\
r^-_t &= \exp\left(\eta L'_y(\hat{y}_t) U x_{t,i}\right) = \frac{1}{r^+_t} \quad (3.11)
\end{align*}
\]

Figure 3: Exponentiated gradient algorithm with positive and negative weights $\text{EG}_L^+(U,(s^+,s^-),\eta)$. 
IMPLICIT VS. EXPLICIT

GD:

\[ w_{t+1} = \arg\min_w \left( \frac{1}{2} ||w - w_t||^2 + \eta L_y(w, x_t) \right) \]

IMPLICIT:

\[ w_{t+1} = w_t - \eta L_y(w, x_t) x_t \]

EXPLICIT:

\[ w_{t+1} = \arg\min_w \left( \frac{1}{2} ||w - w_t||^2 + \eta \left(L_y(w, x_t) + (w - w_t) L_y'(w, x_t) \right) \right) \]

1. ORDER TAYLOR OF \( L_y(w, x_t) \) AT \( w = w_t \)

\[ w_{t+1} = w_t - \eta L_y(w, x_t) x_t \]

EXPLICIT

IMPLICIT UPDATE MORE PRINCIPLED

OFTEN BETTER THAN EXPLICIT UPDATE

- OFTEN NO CLOSED FORM SOLUTION
CLOSED FORM OF IMPLICIT UPDATE
FOR GD & SQUARE LOSS

\[ w_{t+1} \]

\[ = \arg \min_w \left\{ \frac{1}{2} \| w - w_t \|^2 + \eta \ (w \cdot x_t - y_t)^2 \right\} \]

\[ = w_t - \eta \ (w_{t+1} \cdot x_t - y_t) \cdot x_t \]

\[ w_{t+1} \cdot x_t = w_t \cdot x_t - \eta \ (w_{t+1} \cdot x_t - y_t) \cdot x_t \cdot x_t \]

\[ w_{t+1} \cdot x_t (1 + \eta x_t^2) = w_t \cdot x_t + \eta \ y_t \ x_t^2 \]

\[ w_{t+1} \cdot x_t = \frac{1}{1 + \eta x_t^2} \ (w_t \cdot x_t + \eta \ y_t \ x_t^2) \]

SUBSTITUTE INTO \((*)\)

\[ w_{t+1} = w_t - \eta \ \left( \frac{w_t \cdot x_t + \eta \ y_t \ x_t^2}{1 + \eta x_t^2} - y_t \right) \cdot x_t \]

\[ = w_t - \eta \ \left( \frac{w_t \cdot x_t + \eta y_t x_t^2 - y_t - \eta y_t x_t^2}{1 + \eta x_t^2} \right) \cdot x_t \]

\[ = w_t - \frac{\eta}{1 + \eta x_t^2} \ (w_t \cdot x_t - y_t) \cdot x_t \]

IN THIS CASE IMPLICIT = EXPLICIT WITH
CHANGED LEARNING RATE

USUALLY IMPLICIT UPDATE REQUIRES LINESearch
**Logistic Regression**

**Estimating Probabilities w. A Single Neuron**

\[ \hat{y} = \sigma(\hat{\alpha}) \]

\[ \hat{\alpha} = w \cdot x \]

**Weight Vector**

**Feature Vector**
EXAMPLES $(\tilde{x}_t, y_t)$

$x_t$ FEATURE VECTOR

$y_t$ BINARY LABEL

0 CLICK

1 CLICK

OR PROBABILITY

ESTIMATES $\hat{y}_t = \hat{y}(y_t, x_t)$

ALSO PROBABILITIES

SIGMOID TRANSFERS LINEAR ACTIVATION INTO A PROBABILITY

\[
\text{loss}(y_t, \hat{y}_t) = y_t \ln \frac{y_t}{\hat{y}_t} + (1-y_t) \ln \frac{1-y_t}{1-\hat{y}_t}
\]

- SIMPLIFY
- COMPUTE DERIV.
- GD, EG, EGU