Lecture 9

- Sols to HW2 & 3
- Status of open problems

HW2:

1) Weighted Median Bound
   - as LHM
   \[ \log_2 n \] Adv. Strategy against any DET. Expert Alg. under assumption that one expert is consistent:
   - if \( q \) consistent experts are left
     let \( \lfloor \frac{q}{2} \rfloor \) predict 0;
     \( \lceil \frac{q}{2} \rceil \) predict 1
   - choose \( y \neq \hat{y} \)

No matter how \( \hat{y} \) is chosen
\[ \lfloor \frac{q}{2} \rfloor \] consistent experts left

After 2 trials:
\[ \lfloor \frac{\lfloor \frac{q}{2} \rfloor}{2} \rfloor \]

After \( m \) trials when starting with \( m \) one expert left
\[ \lfloor \frac{m}{2^m} \rfloor = 1 \]
\text{SINCE RHS INTEGER}

\[
\frac{n}{2^m} < 2
\]

\[
m > (\log_2 n) - 1
\]

\[
\uparrow
\]

\text{INTEGER}

\[
m \geq \lceil \log_2 n \rceil
\]

3) \text{RUN PREVIOUS STRATEGY FOR}

\[\lceil \log_2 n \rceil - 1 \text{ STEPS}\]

\[
\Rightarrow 2 \text{ CONSISTENT EXPERTS LEFT}
\]

\[
\text{- LET ONE PREDICT 1}
\]

\[
\text{OTHER = 0}
\]

\[
2k + 1 \text{ MORE TRIAL IN WHICH ADV. CHOOSE } y \neq y'
\]

\[
\text{- ONE EXPERT LEFT WITH } \leq k \text{ MISTAKES}
\]

\[
\lceil \log_2 n \rceil - 1 + 2k + 1 \text{ IN TOTAL}
\]

\text{SOLUTIONS TO } 2 \times 3 \text{ CAN ALSO BE FOUND IN}

\text{WM PAPER, THM 8.1}
EC: SUFFICES TO SHOW THAT

\[
\min_{\sum u_i = 1} \left( \Delta (u_i, w_i) - \Delta (u_i, w_{T+1}) + \eta u_i \cdot L_{T+1} \right) \leq \min_{\sum u_i = 1} \left( \Delta (u_i, w_i) + \eta u_i \cdot L_{T+1} \right)
\]

Any min of LHS is \( w_{T+1,i} = \frac{w_{i,T} e^{-\eta L_{T+1,i}}}{z_{T+1}} \) \( \text{NORMALIZE.} \)

PLUGGING \( u = w_{T+1} \) INTO LHS WE FIND THAT VALUE OF LHS IS \(-\ln z_{T+1}\)

REWRITE RHS:

\[
\sum u_i \ln \frac{w_{i,T}}{w_{i,i}} - \sum u_i \ln \frac{w_{i,i}}{w_{T+1,i}} + \eta u_i \cdot L_{T+1}
\]

\[
= \sum u_i \ln \frac{w_{T+1,i}}{w_{i,i}} + \eta u_i \cdot L_{T+1}
\]

\[
= \sum u_i \ln \frac{w_{T+1,i} e^{-\eta L_{T+1,i}}}{w_{T+1,i} z_{T+1}} + \eta u_i \cdot L_{T+1}
\]

\[
= \eta u_i \cdot L_{T+1} - \left( \sum u_i \right) \ln z_{T+1} + \eta u_i \cdot L_{T+1}
\]

\[
= -\ln z_{T+1}
\]

\( \Rightarrow \) LHS SAME FOR ALL \( u \) s.t. \( u_i > 0, \sum u_i = 1 \)

\( \Rightarrow \) MIN OF LHS IS \(-\ln z_{T+1}\)

NICE CHOICE OF \( u \) IS \( u_{T+1} \)

IN THIS CASE SECOND REL. ENTROPY IN LHS:

\( \Delta (w_{T+1}, w_{T+1}) = 0 \)
HW3: \[ M \quad \tilde{M} \]

1) \[ \begin{array}{ccc}
1 & 1 & \text{sink} \\
2 & \text{horn} & a \quad b
\end{array} \]

A permutation \( f \) is assigned product of factors

\[ \prod_{i} M_{i} f(i) \]

Each Sinkhorn step preserves ratios between different permutations

\[ \frac{\prod_{i} M_{i} f(i)}{\prod_{i} M_{i} f'(i)} = \frac{\prod_{i} \tilde{M}_{i} f(i)}{\prod_{i} \tilde{M}_{i} f'(i)} \]

Example: \( f = \begin{pmatrix} 1 & 2 \\ 2 & 1 \end{pmatrix} \), \( f' = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \)

\[ \begin{array}{cc}
1 & 2 \\
1 & 1
\end{array} = \frac{a^2}{b^2} \]

\[ a = \sqrt{2} \cdot b \]

Also: \( a + b = 1 \)

\[ b = \frac{1}{1 + \sqrt{2}} \quad a = \frac{\sqrt{2}}{1 + \sqrt{2}} \]

SOLUTION SATISFIES (x)

IRRATIONAL

Sinkhorn steps only lead to rational matrices

Convergence takes infinitely many steps
1 1 \rightarrow a \quad b
1 4 \rightarrow b \quad a

\frac{1 \cdot 4}{1 \cdot 1} = \frac{a^2}{b^2}

\alpha = 2 b
\alpha + b = 1 \quad \Rightarrow \quad b = \frac{1}{3} \quad \alpha = \frac{2}{3}

SEEMS TO ALSO TAKE INFINITELY MANY STEPS

2\) \quad k = 2

ADVERSARY: - LET ALG COMMIT TO CACHE
- PICK HIT SO THAT \( \geq \left\lceil \frac{k}{2} \right\rceil \) LEFT

\begin{align*}
\text{if ODD} & \quad \begin{cases}
1 \text{ LEFT} \\
2 \text{ LEFT}
\end{cases} \\
\text{if EVEN} & \quad 0
\end{align*}

\text{LOWER BOUND OF } \lg 2n

\( k > 2 \): SIZE OF CACHE IS \( n \). AFTER EACH MISS, CERTAIN NEW POSITIONS MUST BE IN ANY CONSISTENT CACHE. \( L(q) \)

\# OF UNDET. POS.

INITIAL: \( q = n \)
To show: \( L(n) \geq C_k \log_2 \frac{n}{k} \)

\[ q = n - \sum_{i} s_i \]

Size of cache is \( n \)

\[ \sum_{i} s_i + g_i = n \]
\[ \sum_{i} g_i = n - \sum_{i} s_i = q \]

Adv. choose hit s.t. \( q \) reduces the least

\[ L(q) \geq 1 + L(q - \frac{q}{k}) \]
\[ = 1 + L(q(1 - \frac{1}{k})) \]

Cond.: \( L(q) \geq k \ln n \)

\[ k \ln q \geq 1 + k \ln (q(1 - \frac{1}{k})) \]
\[ = 1 + k \ln q + k \ln (1 - \frac{1}{k}) \]
\[ \geq 1 \]

Details to be worked out
Anindya had some more insights
3) Assume $L(y, x)$ not convex. That is, $\exists x_1, x_2 \in \mathbb{R}$:

\[
L(y, \alpha x_1 + (1-\alpha)x_2) \geq \alpha L(y, x_1) + (1-\alpha)L(y, x_2)
\]

\[
f_y(\alpha x_1 + (1-\alpha)x_2) = \exp \left( -\frac{1}{2} L(y, \alpha x_1 + (1-\alpha)x_2) \right) \]

**Ex monotone increasing**

\[
\leq \exp \left( -\frac{1}{2} \left( \alpha L(y, x_1) + (1-\alpha)L(y, x_2) \right) \right)
\]

**Ex convex**

\[
\leq \alpha \exp \left( -\frac{1}{2} L(y, x_1) \right) + (1-\alpha) \exp \left( -\frac{1}{2} L(y, x_2) \right)
\]

\[
f_y(x_1) \quad f_y(x_2)
\]

\[
\Rightarrow f \text{ not concave}
\]
STATUS OF OPEN PROBLEMS

- FIND APPLICATIONS OF LONG-TERM MEMORY

- CACHING (COMBINING LISTS)
  * \( k = 2 \) DET.: WEIGHTED MEDIAN
  * \( k > 2 \) DET. ALG. WITH \( 2M^k + c\sqrt{M^k} knn + O(klnn) \)
    Miss Bound?
    Generalize of Weighted Median?

  * Easy to get Prob. Alg:
    WMR W, \( k \log_2 \frac{n}{k} \) Experts

- Algorithms that take Refetching into account
  * Generalize Arcing to \( k > 2 \)
    : Bounds for Arcing

- Matching lower bounds

- Co-Learning
  * Prove bounds with
    \[ k \ln n + m \ln \frac{n}{m} \]
    Complexity Term
  * Find Alg. Using Multiplic. weights that perform well in practice
    : Various versions of Share Updates?
PERMUTATIONS
- FIND GOOD APPL.
- MAKE ALL EFFICIENT
- ALTERNATE ALG W. ONE WEIGH PER LIST

STOCK MARKET:
- COMPARATOR THAT EXPLOITS "REVERSAL OF MEAN"
- WHAT IS EFFECT OF SHARING $\zeta \eta > 0$
  : STREAMLINED ALG

WHAT IS NEXT:

DERIVATION OF UPDATES
GRADIENT DESCENT
  VERSUS EXPONENTIALIZED GRADIENT
LINEAR REGRESSION

ON-LINE MODEL

INITIALIZE $\tilde{w}_1$

FOR $t = 1, 2, \ldots$ DO

RECEIVE INSTANCE $x_t$

PREDICT $\tilde{y}_t = \tilde{w}_t x_t$

GET LABEL $y_t$

INTRuire loss $(\tilde{y}_t - y_t)^2$

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.

\[
\forall S = \{ (x_1, y_1), (x_2, y_2), \ldots, (x_t, y_t) \} \exists
\]

\[
\inf \sum_{t=1}^{T} (\tilde{w}_t x_t - y_t)^2
\]

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

TOTAL LOSS OF ON-LINE ALG

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

\[ w_{th} = \inf_{\tilde{w}} \left( \frac{1}{2} ||\tilde{w} - \tilde{w}_t||^2 + \eta \frac{1}{2} (\tilde{w} \cdot \tilde{x}_t - y_t)^2 \right) \]

DIVERGENCE TO LAST W T

LOSS ON LAST EXAMPLE

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

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

\[ \begin{align*}
\text{THUS} & \quad \tilde{w}_{th} = \tilde{w}_t - \eta (\tilde{w}_{th} \cdot \tilde{x}_t - y_t) \cdot x_t \\
& \quad \text{NEW OLD} \quad \text{GRADIENT OF LOSS} \\
& \quad \text{LEARNING RATE} \quad \text{LEARNING RATE} \\
& \quad \approx \tilde{w}_t - \eta (w_t \cdot x_t - y_t) \cdot x_t \\
& \quad \text{WIDROW-HOFF UPDATE} \\
& \quad \text{GRADIENT DESCENT}
\end{align*} \]
**LET'S USE A DIFFERENT DIVERGENCE**

\[
W_{t+1,i} = \inf_{W > 0} \left( \sum_i w_i \ln \frac{w_i}{W_t,i} + w_{t,i} - w_{i} + \eta \frac{1}{2} (\bar{w} \cdot x_t - y_t)^2 \right)
\]

**REL. ENTROPY**

**GENERALIZED TO**

**ARBITRARY NON-NEG.**

**WEIGHTS**

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

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

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

**UNNORMALIZED EXPONENTIALIZED**

**GRADIENT ALG**

**EGU**
GENERAL LOSS

GD:

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

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

\[ \approx w_t - \eta \ \nabla L_{yt}(w_t \cdot x_t) x_t \]

EGU

\[ w_{t+1} = \min_{w>0} \left( \sum wi \ln \frac{w_i}{w_{i,t}} + w_{t,i} - w_i + \eta \ L_{yt}(w \cdot x_t) \right) \]

\[ \ln w_{t+1} = \ln w_t - \eta \ \nabla L_{yt}(w_{t+1} \cdot x_t) x_t \]

\[ \approx \ln w_t - \eta \ \nabla L_{yt}(w_t \cdot x_t) x_t \]

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 \(x_t\)

\[ y_t = \begin{cases} +1 & \text{if } w_t x_t \geq \theta \\ -1 & \text{else} \end{cases} \]

GET BINARY LABEL \(y_t \in \{+1,-1\}\)

IF \(y_t = y_t\) THEN \(w_{t+1} = w_t\)

ELSE \(w_{t+1} = w_t + \eta y_t x_t\)

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

\[\vdots\]

IF \(y_t = y_t\) THEN \(w_{t+1} = w_t, \eta y_t x_t, i\)

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

UPDATE ONLY IF MISTAKE
"CONSERVATIVE UPDATES"
linear hinge loss

\[ y_t = 1 \]

\[ y_t = -1 \]

GD = Update of Perceptron

EGBU = "Winnow"

\[ L' y_t (w_t \cdot x_t) = 0 \text{ when no mistake} \]

\[ L' y_t (w_t \cdot x_t) = -y_t \text{ if mistake} \]
\[ w_{t+1} = \text{INF} \left( \sum_{i} w_i \ln \frac{w_i}{w_{t,i}} + \eta \frac{1}{2} (\bar{w} \cdot x_t - y_t)^2 \right) \]

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

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

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

\[ w_{t+1,i} = w_{t,i} e^{-1 - \lambda} \]

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

\( e^{-1 - \lambda} \) is 1/NORMALIZATION

\[ w_{t+1,i} = \frac{w_{t,i} e^{-\eta (w_{t+1} \cdot x_t - y_t) x_{t,i}}}{\sum_{j} w_{t,j} e^{-\eta (w_{t} \cdot x_t - y_t) x_{t,j}}} \]
\[
\text{EXPO\textsc{NENTIATED GRADIENT ALG.}}
\]

For general loss: 
\[L_{yt}(w \cdot x_t) = -\eta \frac{L_{yt}(w \cdot x_t) x_{t,i}}{\sum_{j} w_{t,j} e^{-\eta (w_{t,j} \cdot x_t - y_t)}}\]

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

\[\rightarrow \text{FOR HINGELOSS EG BECOMES NORMALIZED VERSION OF WINNOW}\]

\[\rightarrow \text{DEF } L_{yt}(w \cdot x_t) := \sum_{i \in \text{PROB. VECT.}} w_i L_{yt}(x_{t,i}) \text{ CONVEX COMBINATION OF LOSSES OF EXPERTS E}_i\]

EG BECOMES EXPERT UPDATE:
\[W_{t+1,i} = W_{t,i} e^{-\eta L_{yt}(x_{t,i})} \text{ implicit}\]
\[\text{NORMALIZE.}\]
Algorithm $\text{EG}^\pm_L(U, (s^+, s^-), \eta)$

**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 $w^+_1 = Us^+$ and $w^-_1 = Us^-.$

**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

$$w^+_{t+1,i} = U \cdot \frac{w^+_{t,i} r^+_i}{\sum_{j=1}^{N}(w^+_{t,j} r^+_{t,j} + w^-_{t,j} r^-_{t,j})},$$

$$w^-_{t+1,i} = U \cdot \frac{w^-_{t,i} r^-_i}{\sum_{j=1}^{N}(w^+_{t,j} r^+_{t,j} + w^-_{t,j} r^-_{t,j})},$$

where

$$r^+_i = \exp \left(-\eta L'_{y_t}(\hat{y}_t) U x_{t,i} \right),$$

$$r^-_i = \exp \left(\eta L'_{y_t}(\hat{y}_t) U x_{t,i} \right) = \frac{1}{r^+_i}.$$

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

**IMPLICIT VS. EXPLICIT**

GD:

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

**IMPLICIT**

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

**EXPLICIT**

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

1. **ORDER TAYLOR OF**

$$L_{y_t}(w, x_t) \text{ AT } w = w_t$$

explicit
2nd Order Approximation:

\[ L(w) \approx L(w_t) + (w - w_t)^T \nabla L(w_t) \]

\[ + \frac{1}{2} (w - w_t)^T \nabla^2 L(w_t) (w - w_t) \]

\[ \tilde{L}(w) = \text{QUADRATIC} \]

NEwTON Step GOES TO MIN.

\[ \nabla \tilde{L}(w) = \nabla L(w_t) + \nabla^2 L(w_t) (w - w_t) \]

\[ \nabla \tilde{L}(w) \bigg|_{w = w_t} = 0 \]

\[ \nabla L(w_t) + \nabla^2 L(w_t) (w_{t+1} - w_t) = 0 \]

\[ w_{t+1} - w_t = -\left( \nabla^2 L(w_t) \right)^{-1} \nabla L(w_t) \]

\[ w_{t+1} = w_t - \left( \nabla^2 L(w_t) \right)^{-1} \nabla L(w_t) \]

\[ w_{t+1} = w_t - \left( \nabla^2 L(w_t) \right)^{-1} \nabla L(w_t) \]
SECOND VIEW OF NEWTON:

Finding root of $\nabla L(w) = f(w)$

$w_{t+1} = w_t - (\nabla f(w_t))^{-1} f(w_t)$

NEWTON WITH REGULARIZATION:

$\lambda (w - w_t) + \nabla L(w_t) + \nabla^2 L(w_t) (w - w_t)$

$\nabla \quad | w = w_{t+1} = 0$

$\nabla L(w_t) + (\nabla^2 L(w_t) + \lambda I) (w_{t+1} - w_t) = 0$

$w_{t+1} = w_t - (\nabla^2 L(w_t) + \lambda I)^{-1} \nabla L(w_t)$

END 2ND ORDER...