LECTURE 16

HOW TO DERIVE ON-LINE UPDATES

- STARTING WITH A DIFFERENTIAL EQUATION
- USING BREGMANN DIVERGENCES
On-line Linear Regression

For $t = 1, \ldots, T$ do

Get instance $x_t \in \mathbb{R}^n$

Predict $\tilde{y}_t = w_t \cdot x_t$

Get label $y_t \in \mathbb{R}$

Incur loss $L_t(w_t) = (y_t - \tilde{y}_t)^2$

Update $w_t$ to $w_{t+1}$

Comparison class is a set of linear predictors
What if no comparator consistent?

Sequence of examples $S = (x_1, y_1), \ldots, (x_T, y_T)$

- $L_A(S)$ be the total loss of alg. $A$

- $L_u(S)$ be the total loss of linear weight vector $u$

Want bounds of the form:

$$\forall S : L_A(S) \leq \min_u (L_u(S) + \text{additional})$$

Bounds loss of algorithm relative to loss of best linear predictor $u$

Off-line alg. can choose best $u$ based on all data
Examples of Updates

Gradient descent
($w \in \mathbb{R}^n$)

$$w_{t+1} = w_t - \eta \nabla L_t(w_t)$$

$$= w_t - \eta (w_t \cdot x_t - y_t) x_t \quad [WH]$$

Exponentiated Gradient Algorithm
($w$ is probability vector)

$$w_{t+1,i} = w_{t,i} \exp \left[ -\eta \frac{\partial L_t(w_t)}{\partial w_{t,i}} \right] / \text{normaliz.} \quad [KW]$$
Continuous Updates

Gradient Descent
\[ w \in \mathbb{R}^n \]
\[ \dot{w}_t = -\eta \nabla_w L_t(w_t) \]

Unnormalized Exponentiated Gradient Alg.
\[ w \geq 0 \]
\[ \log(w_t) = -\eta \nabla_w L_t(w_t) \]

[WJ]
Characterization of algs.
i.t.o. link function $f = \nabla F$  \[ [\text{WJ, MW}] \]

$F(w)$ convex

$$\frac{\dot{f}(w_t)}{\theta_t} = -\eta \nabla_w L_t(w_t)$$

<table>
<thead>
<tr>
<th>Alg.</th>
<th>$f(w)$</th>
<th>Domain of $w$</th>
</tr>
</thead>
<tbody>
<tr>
<td>GD</td>
<td>$f(w) = w$</td>
<td>$w \in \mathbb{R}^n$</td>
</tr>
<tr>
<td>EGU</td>
<td>$f(w) = \log w$</td>
<td>$w \in [0, \infty)^n$</td>
</tr>
<tr>
<td>EG</td>
<td>$f(w) = \ln \frac{w}{1 - |w|_1}$</td>
<td>$w \in [0, 1]^{n-1}$, $|w|_1 \leq 1$</td>
</tr>
<tr>
<td>BEG</td>
<td>$f(w) = \ln \frac{w}{1-w}$</td>
<td>$w \in [0, 1]^n$</td>
</tr>
</tbody>
</table>

$f$ is barrier function
IN GAME THEORY:

\[ w_i \] FRACTION OF INDIVIDUALS OF SPECIES \( i \);
I.E. \( \bar{w} \in (0, \infty)^m \), E.G.

WHEN \( |w_i| = 1 \), I.E. \( w_i \) MIXTURE COEFFICIENTS
THEN E.G. NOW \( \bar{w} \) ON SIMPLEX

WHEN \( \bar{w} \in [0, 1]^m \), THEN \( \bar{w} \) IN BOX, E.G.
Discretization

\[ \frac{f(w_{t+h}) - f(w_t)}{h} = -\eta \nabla L_t(w_t) \]

\[ w_{t+h} = f^{-1} (f(w_t) - \eta h \nabla wL_t(w_t)) \]

We use \( h = 1 \)

\[ w_{t+1} = f^{-1} (f(w_t) - \eta \nabla L_t(w_t)) \]

Conjecture: Forward Euler better:

Replace \( \nabla_w L_t(w_t) \) by \( \nabla_w L_t(w_{t+h}) \)
Alternate Motivation \[ [KW] \]

GD

\[ w_{t+1} = \arg\min_w \left( \|w - w_t\|_2^2/2 + \eta(y_t - w \cdot x_t)^2/2 \right) \]

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

EG

\[ w_{t+1} = \arg\min_{w \in \text{simplex}} \left( \sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + \eta(y_t - w \cdot x_t)^2/2 \right) \]

\[ = w_{t,i} \exp \left[ -\eta \left( \frac{w_{t+1} \cdot x_t - y_t}{\approx w_t \cdot x_t} \right) x_{t,i} \right] / \text{normaliz} \]
### Families of update algorithms

<table>
<thead>
<tr>
<th>parameter divergence</th>
<th>name of family</th>
<th>update algs.</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\sum_{i=1}^{n} w_i \log \frac{w_i}{w_{t,i}}$</td>
<td>Exponentiated Gradient Alg.</td>
<td>expert algs Normalized Winnow &quot;AdaBoost&quot;</td>
</tr>
</tbody>
</table>
Families of update algorithms (cont)

\[ \sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + w_{t,i} - w_i \]

Unnormalized Winnow
Exp. Grad. Alg.

\[ \sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + (1 - w_i) \ln \frac{1-w_i}{1-w_{t,i}} \]

Binary
Exp. Grad. Alg.

any
Bregman divergence

Members of different families exhibit different behavior
Bregman Divergences

For any differentiable convex function $F$

\[
\Delta_F(\tilde{w}, w) = F(\tilde{w}) - F(w) - (\tilde{w} - w) \cdot \nabla w F(w) / f(w)
\]

\[F(\tilde{w})\]

\[\Delta_F(\tilde{w}, w)\]

\[F(w) - (\tilde{w} - w) \cdot f(w)\]
Examples

Squared Euclidean Distance

\[ F(w) = \|w\|_2^2/2 \]

\[ f(w) = w \]

\[ \Delta_F(\tilde{w}, w) = \|\tilde{w}\|_2^2/2 - \|w\|_2^2/2 - (\tilde{w} - w) \cdot w \]

\[ = \|\tilde{w} - w\|_2^2/2 \]

(Unnormalized) Relative Entropy

\[ F(w) = \sum_i (w_i \ln w_i - w_i) \]

\[ f(w) = \ln w \]

\[ \Delta_F(\tilde{w}, w) = \sum_i \left( \tilde{w}_i \ln \frac{\tilde{w}_i}{w_i} + w_i - \tilde{w}_i \right) \]

Bregmann divergences also lead to good loss functions.
THE PREVIOUS GAME THEORETIC APPLICATION

$w_i$: MIXTURE COEFFICIENT FOR
PURE STRATEGY $i$

$L_{t,i}$: LOSS OF STRATEGY $i$ AT TRIAL $t$

FOR $t=1$ TO $T$ DO

RECEIVE LOSSES $L_{t,i}$

ALG. INCURS LOSS $\bar{w}_t \cdot L_t$

UPDATE $\bar{w}_t$ TO $\bar{w}_{t+1}$.

DERIVATION OF UPDATE

$\bar{w}_{t+1} = \text{ARG INF}_{\bar{w} \in \text{SIMPLEX}} \left( \sum w_i L_{t,i} \frac{w_i}{\bar{w}_i} + \eta \bar{w} \cdot L_t \right)$

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

$\frac{1}{T} \sum w_{t} \cdot L_{t} \leq \text{INF} \left( \frac{1}{T} \sum w \cdot L_{t,i} + \Delta_{t,m} \right)$

TOTAL LOSS OF ALG
$\Delta_{t,m} = \sqrt{2m} \frac{m^{-1}}{T} + 6m \frac{n}{T}$

NEEDS: $L_{t,i} \in [0,1]$ & $\eta = \ln \left( 1 + \sqrt{\frac{2m}{T}} \right)$
PROBLEM:

$\eta$ depends on $T$

Not all $L_t, i$ might be known

More powerful model (Multiarmed Bandit Problem):

Pick $i$ according to $\omega, i$

Only told $L_t, i$

(need to estimate $L_t, j$ for $j \neq i$)
REACHING TOWARDS FUTURE
- USE FUTURE GRADIENT
- ESTIMATE FUTURE LOSS/label

RIDGE REGRESSION:

\[
\tilde{w}_{t+1} = \underset{w}{\text{argmin}} \left( \alpha \tilde{w}^2 + \sum_{s=1}^{t} (w \cdot x_s - y_s)^2 \right)
\]

INERTIA TERM VERSUS LOSS OF ALL PAST EXAMPLES

\[
\tilde{w}_{t+1} = \left( \alpha I + \sum_{s=1}^{t} x_s x_s^T \right)^{-1} \sum_{s=1}^{m \times 1} y_s x_s
\]

BEFORE:

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

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

WIDROW HOFF
Better forward update \([V, KW]\):

\[
\hat{w}_{t+1} = \text{arg inf}_w \left\{ a \bar{w}^2 + \sum_{s=1}^{t} (w^s x_s - y_s)^2 + (w x_{t+1} - 0)^2 \right\}
\]

\[
\uparrow \quad \text{ENCORPORATED}
\]

Better bounds for

\[
\frac{1}{t} \sum_{t=1}^{T} (y_t - w_t x_t)^2 \leq \text{inf}_w \left( \frac{1}{t} \sum_{t=1}^{T} (y_t - w x_t)^2 + a \|w\|^2 \right)
\]

Why should \(y_t\) be guessed as 0?

Also occurs for Gaussian density estimator:

Predicting with mean \(\sum_{s=1}^{t-1} x_s\)

Better than predicting with

\[
\frac{1}{t-1} \sum_{s=1}^{t-1} x_s \quad \text{SHRINKAGE HELPS!}
\]

20
TEMPORAL DIFFERENCE LEARNING

AT TRIAL $t$
- LEARNER RECEIVES INSTANCE $\tilde{x}_t \in \mathbb{R}^n$
- MAKES PREDICTION $\tilde{w}_t \cdot \tilde{x}_t$
- RECEIVES REINFORCEMENT $r_t$
- INCURS LOSS $\left(\tilde{w}_t \cdot \tilde{x}_t - y_t\right)^2$

WHERE $y_t = y_t + \gamma y_{t+1} + \gamma^2 y_{t+1} + \ldots$

HOW DO UPDATES GENERALISE?

RECALL: $\arg\inf \left( \alpha \|\tilde{w}_t\| + \sum_{s=1}^{t} (\tilde{w}_t \cdot x_s - y_s)^2 \right)$

$\tilde{w}_t \cdot x_s$

$= A_t \sum_{s=1}^{t} y_s x_s \quad (\star)$
\[ y_s = \left( \sum_{k=s}^{t-1} x_k y_{s+k} \right) + \theta_{t-k} y_t \]

(4) BECOMES

\[ A_t^{-1} \left( \sum_{s=1}^{t} \sum_{k=s}^{t-1} \theta_{s-k} w_k x_s + y_t \sum_{s=1}^{t} \theta_t x_s \right) \]

UNKNOWN AT TIME OF PREDICTION AT TRIAL t

GUESS \( y_t = 0 \) AND USE

\[ w_{t+1} = A_t^{-1} \left( \sum_{s=1}^{t} \sum_{k=s}^{t-1} \theta_{s-k} w_k x_s \right) \]

TLS ALGORITHM [FWJ]
General Setup

- We hide some information from the learner

- The relative loss bound quantifies the price for hiding the information

- So far the future examples are hidden
  Off-line algorithm knows all examples
  On-line algorithm knows past examples
Minimax Algorithm for $T$ Trials

Gaussian

Learner against adversary

\[ \inf_{\theta_1} \sup_{x_1} \inf_{\theta_2} \sup_{x_2} \inf_{\theta_3} \sup_{x_3} \ldots \inf_{\theta_T} \sup_{x_T} \]

\[ \sum_{t=1}^{T} \frac{1}{2} (\theta_t - x_t)^2 - \inf_{\theta} \left( \sum_{t=1}^{T} \frac{1}{2} (\theta - x_t)^2 \right) \]

- total loss of on-line algorithm
- total loss of off-line algorithm

Instances must be bounded: $||x_t||_2 \leq X$

Minimax algorithm usually intractable

Bernoulli is another exception
**Gaussian**

Max likelihood \[ \theta_t = \frac{\sum_{q=1}^{t-1} x_q}{t-1} \]

**Forward Alg.** \[ \theta_t = \frac{\sum_{q=1}^{t-1} x_q}{t-1+1} \]

**Bound** \[ \frac{1}{2} X^2 (1 + \ln T) \]

**Minimax Alg.** \[ \theta_t = \frac{\sum_{q=1}^{t-1} x_q}{t+\ln(T)-\ln(t+O(\ln T))} \]

**Bound** \[ \frac{1}{2} X^2 (\ln T - \ln \ln T) + o(1) \]

Minimax alg. needs to know \( T \)

*Why shrinkage?*
Last-step Minimax

Assumes that current trial is last trial \([Fo, TW]\)

\[
\theta_t = \arg \inf_{\theta} \sup_{x_t} \left( \sum_{q=1}^{t} L_q(\theta_q) - \inf_{\theta} L_{1..t}(\theta) \right)
\]

\[
= \arg \inf_{\theta} \sup_{x_t} \left( L_t(\theta_t) - \inf_{\theta} L_{1..t}(\theta) \right)
\]

*NO DIVERGENCES?*

For Gaussian and linear regression
Last-step Minimax is same as Forward Alg.

For Bernoulli Last-step Minimax slightly better than Laplace Estimator

*GOOD LOSS BOUNDS FOR ENTIRE ONE-DIMENSIONAL EXPONENTIAL FAMILY OF DISTRIBUTIONS*

*CONJECTURE: LAST-STEP MINIMAX GOOD HEURISTIC FOR ANY "PROPER SCORING RULE"*