On-line Learning -
Methods and Open Problems

Manfred K. Warmuth
UC Santa Cruz, USA

Dagstuhl, Germany
July 26, 2001

Early contributors:
Nick Littlestone
Volodya Vovk

Help with this tutorial:
Claudio Gentile
High-level Themes

- On-line versus Off-line

- Geometric versus information theoretic Kernels partial kernels
Why on-line

- Simple algs
- Data too large
- Superiour bounds
- Data inherently on-line
- The game MAFIA
Overview

- Differential motivation of updates
- Motivation with Bregman divergences
- Bregman divergences as loss functions
- Pythagorean Theorem
- Rotation invariance and kernel trick
- Proving relative loss bounds
- Bregman divergences and the exponential family
On-line Linear Regression

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

- Get instance $x_t \in \mathbb{R}^n$
- Predict $\hat{y}_t = w_t \cdot x_t$
- Get label $y_t \in \mathbb{R}$
- Incur loss $L_t(w_t) = (y_t - \hat{y}_t)^2$
- 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 $u$
Examples of Updates

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

\[
\begin{align*}
    w_{t+1} &= w_t - \eta \nabla L_t(w_t) \\
    &= w_t - \eta (w_t \cdot x_t - y_t) x_t
\end{align*}
\] [WH]

Exponentiated Gradient Algorithm [KW]
\( (w \text{ is probability vector}) \)

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

Gradient Descent

$w \in \mathbb{R}^n$

\[ \dot{w}_t = -\eta \nabla wL_t(w_t) \]

Unnormalized Exponentiated Gradient Alg.

$w \geq 0$

\[ \log(w_t) = -\eta \nabla wL_t(w_t) \]

[WJ]
Characterization of algs.
i.t.o. link function \( f = \nabla F \)  \[WJ, MW\]

\( F(w) \) 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} ) ( w \in [0, 1]^{n-1}, |w|_1 \leq 1 )</td>
<td></td>
</tr>
<tr>
<td>BEG</td>
<td>( f(w) = \ln \frac{w}{1-w} ) ( w \in [0, 1]^n )</td>
<td></td>
</tr>
</tbody>
</table>

\( f \) is barrier function
Discretization

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

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

We use \( h = 1 \)

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

Conjecture: **Forward Euler** better:

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

**GD**

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

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

**EG**

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

\[ = w_{t,i} \exp \left[ -\eta \left( w_{t+1} \cdot x_t - y_t \right) x_{t,i} \right] / \text{normaliz} \]
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>$\sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{l,i}}$</td>
<td>Exponentiated Gradient Alg.</td>
<td>expert algs Normalized Winnow “AdaBoost”</td>
</tr>
</tbody>
</table>
Families of update algorithms (cont)

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


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


any Bregman divergence

Members of different families exhibit different behavior
Bregman Divergences \[ [Br, CL, Cs] \]

For any differentiable convex function \( F \)

\[
\Delta_F(\tilde{w}, w) = F(\tilde{w}) - \text{supporting hyperplane through } (w, F(w)) \\
= F(\tilde{w}) - F(w) - (\tilde{w} - w) \cdot \nabla_w F(w) \\
\]

\[
F(w) - (\tilde{w} - w) \cdot f(w) 
\]
Bregman Divergences, Simple Properties

1. $\Delta_F(\tilde{w}, w)$ is convex in $\tilde{w}$

2. $\Delta_F(\tilde{w}, w) \geq 0$
   If $F$ convex equality holds iff $\tilde{w} = w$

3. $\nabla_{\tilde{w}} \Delta_F(\tilde{w}, w) = f(\tilde{w}) - f(w)$

4. Usually not symmetric: $\Delta_F(\tilde{w}, w) \neq \Delta_F(w, \tilde{w})$

5. Linearity (for $a \geq 0$):
   $\Delta_{F+aH}(\tilde{w}, w) = \Delta_F(\tilde{w}, w) + a \Delta_H(\tilde{w}, w)$

6. Unaffected by linear terms ($a \in \mathbb{R}$, $b \in \mathbb{R}^n$):
   $\Delta_{H+a\tilde{w}+b}(\tilde{w}, w) = \Delta_H(\tilde{w}, w)$

7. $\Delta_F(w_1, w_2) + \Delta_F(w_2, w_3)$
   $= \Delta_F(w_1, w_3) + (w_1 - w_2) \cdot (f(w_3) - f(w_2))$
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

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

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

\[ \Delta_F(\tilde{w}, w) = \sum_i \left( \tilde{w}_i \ln \frac{\tilde{w}_i}{w_i} + w_i - \tilde{w}_i \right) \]
Bregman Divergences
Lead to Good Loss Functions

\[ \hat{y} = h(w \cdot x) \]

- Sigmoid function \( h(z) = \frac{1}{1+e^{-z}} \)

- For a set of examples \( (x_1, y_1), \ldots, (x_T, y_T) \)
  total loss \( \sum_{t=1}^{T} (h(w \cdot x) - y_t)^2 / 2 \)
  can have exponentially many minima in weight space \[ [Bu, AHW] \]
Want loss that is convex in $w$
Bregman Divergences
Lead to Good Loss Functions (cont)

\[ (h = \nabla H) \]

\[ \int_{h^{-1}(y)}^{w \cdot x} (h(z) - y) \, dz \]

\[ = H(w \cdot x) - H(h^{-1}(y)) - (w \cdot x - h^{-1}(y)) \cdot y \]

\[ = \Delta_H(w \cdot x, h^{-1}(y)) \]
Use $\Delta_H(w \cdot x, h^{-1}(y))$ as loss of $w$ on $(x, y)$

Called matching loss for $h$ [AHW, HKW]

Matching loss is convex in $w$

<table>
<thead>
<tr>
<th>transfer f. $h(z)$</th>
<th>$H(z)$</th>
<th>match. loss $d_H(w \cdot x, h^{-1}(y))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$z$</td>
<td>$\frac{1}{2}z^2$</td>
<td>$\frac{1}{2}(w \cdot x - y)^2$ square loss</td>
</tr>
<tr>
<td>$\frac{e^z}{1 + e^z}$</td>
<td>$\ln(1 + e^z)$</td>
<td>logistic loss</td>
</tr>
<tr>
<td>$\text{sign}(z)$</td>
<td>$</td>
<td>z</td>
</tr>
</tbody>
</table>
For transfer function $h(z) = \text{sign}(z)$

$$H(z) = |z|$$

Matching loss is hinge loss [GW]

$$HL(w \cdot x, h^{-1}(y)) = \max\{0, -y w \cdot x\}$$

Convex in $w$ but not differentiable
Motivation of linear threshold algs

Gradient descent with Hinge Loss

Expon. gradient with Hinge Loss

Perceptron

Normalized Winnow

Known linear threshold algs for $\pm 1$-class are gradient-based algs with hinge loss
Trade-off between two divergences \([KW]\)

\[
  w_{t+1} = \arg\min_w \left( \Delta_F(w, w_t) + \eta \Delta_H(w \cdot x_t, h^{-1}(y_t)) \right)
\]

- parameter divergence
- matching loss

Both divergences are convex in \(w\)

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

Generalization of the “delta”-rule
Projections

\[ w_{t+1} = \arg \min_w (\Delta_F(w, w_t) + \eta (w \cdot x_t - y_t)^2) \]

When \( \eta \) is large then \( w_{t+1} \) is projection of \( w_t \) onto plane \( w \cdot x_t = y_t \)

\[ w_{t+1} = \arg \min_{\{w : w \cdot x_t = y_t\}} \Delta_F(w, w_t) \]

The AdaBoost update of the probability vector \( w_t \) on the examples is a projection w.r.t. divergence \( \Delta_F(w, w_t) = \sum_i w_i \ln \frac{w_i}{w_{t,i}} \)

[La,KW]
**A Pythagorean Theorem**  [Br, Cs, A, HW]

$w^*$ is projection of $w$ onto convex set $\mathcal{W}$ w.r.t. Bregman divergence $\Delta_F$:

$$w^* = \operatorname{argmin}_{u \in \mathcal{W}} \Delta_F(u, w)$$

**Th:**

$$\Delta_F(u, w) \geq \Delta_F(u, w^*) + \Delta_F(w^*, w)$$
Kernel trick

• Prediction determined by $w \cdot x$

• $w$ is linear combination of past examples

\[ x \rightarrow \Phi(x) \]
\[ \left( \sum_t \alpha_t \Phi(x_t) \right) \cdot \Phi(x) = \sum_t \alpha_t \frac{\Phi(x_t) \cdot \Phi(x)}{K(x_t, x)} \]

When trick applicable?

A linear predictions algorithm is rotation invariant when rotating the instances does not change the predictions

Rotation invariance $\Rightarrow w$ lin. comb. of ex.
When applicable (cont.)?

Bregman update:

\[
\begin{align*}
    f(w_T) &= f(w_1) + \eta \sum_t \nabla w L(y_t, h(w_t \cdot x_t)) \\
    &= f(w_1) + \text{lin. comb. of ex.}
\end{align*}
\]

<table>
<thead>
<tr>
<th>geometric</th>
<th>information theoretic</th>
</tr>
</thead>
<tbody>
<tr>
<td>rotation invariant</td>
<td>no</td>
</tr>
<tr>
<td>(f = id)</td>
<td>(f = \log)</td>
</tr>
<tr>
<td>kernels</td>
<td>???</td>
</tr>
</tbody>
</table>
How do we prove relative loss bounds?

Loss: \[ L_t(w) = L((x_t, y_t), w) \] convex in \( w \)

Divergence: \( \Delta_F(u, w) \)

Update: \[ f(w_{t+1}) - f(w_t) = -\eta \nabla_w L_t(w_t) \]

\[
L_t(u) \geq L_t(w_t) + (u - w_t) \cdot \nabla_w L_t(w_t)
\]

\[
= L_t(w_t) - \frac{1}{\eta} (u - w_t) \cdot (f(w_{t+1}) - f(w_t))
\]

Prop. 7 of \( \Delta_F \)

\[
= L_t(w_t)
+ \frac{1}{\eta} \left( \Delta_F(u, w_{t+1}) - \Delta_F(u, w_t) - \Delta_F(w_t, w_{t+1}) \right)
\]
Summing over $t$

$$\sum_t L_t(w_t) \leq \sum_t L_t(u)$$

$$+ \frac{1}{\eta} \sum_t \left( \Delta_F(u, w_t) - \Delta_F(u, w_{t+1}) \right) + \Delta_F(w_t, w_{t+1})$$

$$\leq \sum_t L_t(u)$$

$$+ \frac{1}{\eta} \left( \Delta_F(u, w_1) - \Delta_F(u, w_{T+1}) \right) \geq 0$$

$$+ \frac{1}{\eta} \sum_t \Delta_F(w_t, w_{t+1})$$

$$\sum_t L_t(w_t) \leq \sum_t L_t(u) + \frac{1}{\eta} \Delta_F(u, w_1)$$

$$+ \frac{1}{\eta} \sum_t \Delta_F(w_t, w_{t+1})$$

Any convex loss and any Bregman divergence!
Key step:

Relate $\Delta_F(w_t, w_{t+1})$ to loss $L_t(w_t)$

Loss & divergence dependent

Get $\Delta_F(w_t, w_{t+1}) \leq \text{const} L_t(w_t)$

Then solve for $\sum_t L_t(w_t)$

Yield bounds of the form

$$\sum_t L_t(w_t) \leq a \sum_t L_t(u) + b \Delta_F(u, w_1)$$

$a, b \text{ constants, } a > 1$.

Regret bounds ($a = 1$):

time changing $\eta$, subtler analysis [AW]
Some Bounds \textbf{[KW,GLS,GL]}
(Linear Regression with Square Loss)

Gradient Descent

\[
\sum_{t} L_t(w_t) \leq (1 + c) \sum_{t} L_t(u) + \frac{1+c}{c} X_2^2 U_2^2
\]

\[|x_t|_2 \leq X_2, \ ||u||_2 \leq U_2, \ c > 0\]

Scaled Exponentiated Gradient

\[
\sum_{t} L_t(w_t) \leq (1 + c) \sum_{t} L_t(u) + \frac{1+c}{c} \ln n X_\infty^2 U_1^2
\]

\[|x_t|_\infty \leq X_\infty, \ ||u||_1 \leq U_1, \ c > 0\]

\textit{p}-norm alg

\[
\sum_{t} L_t(w_t) \leq (1+c) \sum_{t} L_t(u) + \frac{1+c}{c} (p - 1) X_p^2 U_q^2
\]

\[|x_t|_p \leq X_p, \ ||u||_q \leq U_q, \ c > 0\]
Hadamard example

\[
\begin{pmatrix}
1 & 1 & 1 & 1 \\
1 & -1 & 1 & -1 \\
1 & 1 & -1 & -1 \\
1 & -1 & -1 & 1
\end{pmatrix}
= \begin{pmatrix}
0 \\
0 \\
1 \\
0
\end{pmatrix}
\]

\[H\]
instances

unit target

labels

Rotation invariant alg. \hspace{1cm} O(n) total loss

Information theoretic alg. \hspace{1cm} O(\log n) total loss

Rotating

\[
\begin{pmatrix}
n & 0 & 0 & 0 \\
0 & n & 0 & 0 \\
0 & 0 & n & 0 \\
0 & 0 & 0 & n
\end{pmatrix}
\begin{pmatrix}
\frac{1}{n} \\
\frac{1}{n} \\
\frac{1}{n} \\
\frac{1}{n}
\end{pmatrix}
= \begin{pmatrix}
1 \\
1 \\
-1 \\
-1
\end{pmatrix}
\]
Product of norms explains behavior

\[
\begin{align*}
\text{GD} & & \sqrt{n} & \left\| x \right\|_2^2 \left\| w \right\|_2^2 = n \\
\text{EG} & & \left\| x \right\|_\infty^2 \left\| w \right\|_1^2 \log n = \log n
\end{align*}
\]

\[F(w) = \frac{\left\| w \right\|_2^2}{2}\] only rotation invariant norm
Where do Bregman divergences come from?

- Exponential family of distributions
- Inherent duality
Exponential Family of Distributions

- Parametric density functions
  \[ P_G(x|\theta) = e^{\theta \cdot x - G(\theta)} P_0(x) \]

- \(\theta\) and \(x\) vectors in \(\mathbb{R}^d\)

- Cumulant function \(G(\theta)\) assures normalization
  \[ G(\theta) = \ln \int e^{\theta \cdot x} P_0(x) \, dx \]

- \(G(\theta)\) is convex function on convex set \(\Theta \subseteq \mathbb{R}^d\)

- \(G\) characterizes members of the family

- \(\theta\) is natural parameter
Bregman Divergences
as Relative Entropies
between Exponential Distributions

Let $P(x|\theta)$ and $P(x|\tilde{\theta})$ denote two distributions
with cumulant function $G$

$$\Delta_G(\tilde{\theta}, \theta)$$

$$= \int_x P_G(x|\theta) \ln \frac{P_G(x|\theta)}{P_G(x|\tilde{\theta})} \, dx$$

$$= G(\tilde{\theta}) - G(\theta) - (\tilde{\theta} - \theta) \cdot \frac{g(\theta)}{\tilde{w}}$$

$$F(w) = \theta \cdot w - G(\theta)$$

$$= F(w) - F(\tilde{w}) - (w - \tilde{w}) \cdot \frac{f(\tilde{w})}{\tilde{\theta}}$$

$$= \Delta_F(w, \tilde{w})$$

[A,BN,AW]
Area unchanged When Slide Flipped

\[ \Delta_G(\theta, \tilde{\theta}) = \Delta_F(\tilde{\mu}, w) \]
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 \( T \)
Last-step Minimax

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

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

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

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

For Bernoulli Last-step Minimax slightly better than Laplace Estimator
Open problem solved

Relative entropy is regularizer and barrier

Conjecture:

$$\arg\min_w \sum_i w_i \log \frac{w_i}{1/n}$$

subject to $|w|_1 = 1$

and

$$\arg\min_w \sum_i w_i^2$$

subject to $|w|_1 = 1$ and $w_i \geq 0$

behave similar!
Open problems

• When does shrinkage help?

• Kernel’s for other link functions

• Better conversions to off-line bounds [CG]

• Good on-line updates for learning rates?

• Sample size \[ \geq \frac{d}{\epsilon} \log \frac{1}{\epsilon} + \frac{\log \frac{1}{\delta}}{\epsilon} \]

  produces \((\epsilon, \delta)\)-good hypotheses

  When can the \(\log \frac{1}{\epsilon}\) factor be dropped?

  Is it true when class intersection closed and alg. uses smallest consistent concept?

• VC dim. \(d\)

  \(\Rightarrow\) compression scheme of size \(d\)