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

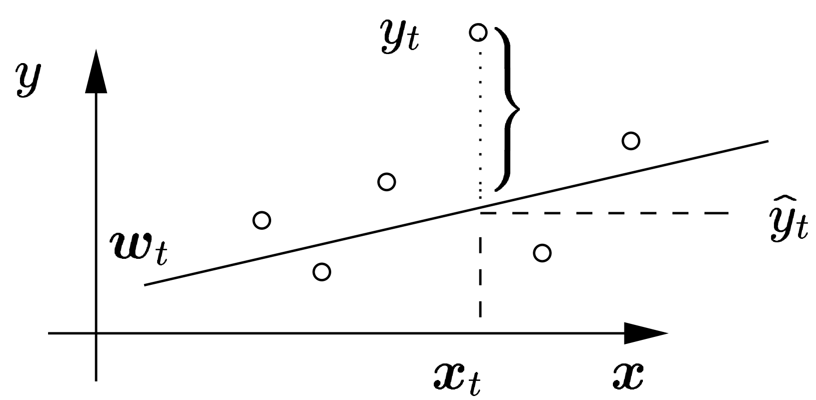
Overview

- Differential motivation of updates
- Motivation with Bregman divergences
- Bregman divergences as loss functions
- Pythogorean 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	$\boldsymbol{x}_t \in \mathbf{R}^n$
Predict	$\widehat{y}_t = \boldsymbol{w}_t \cdot \boldsymbol{x}_t$
Get label	$y_t \in \mathbf{R}$
Incur loss	$L_t(\boldsymbol{w}_t) = (y_t - \widehat{y}_t)^2$
Update	$\boldsymbol{w}_t$ to $\boldsymbol{w}_{t+1}$



Comparison class is a set of linear predictors

What if no comparator consistent?

Sequence of examples $S = (\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)$

- $L_A(S)$ be the total loss of alg. A
- $L_{\boldsymbol{u}}(S)$ be the total loss of linear weight vector $\boldsymbol{u}$

Want bounds of the form:

$$\forall S: \quad L_A(S) \leq \min_{\boldsymbol{u}} \left(L_{\boldsymbol{u}}(S) + \text{additional}\right)$$

Bounds loss of algorithm relative to loss of best linear predictor $\boldsymbol{u}$

Examples of Updates

Gradient descent

$(\boldsymbol{w} \in \mathbf{R}^n)$

$$\boldsymbol{w}_{t+1} = \boldsymbol{w}_t - \eta \nabla L_t(\boldsymbol{w}_t)$$

$$= \boldsymbol{w}_t - \eta(\boldsymbol{w}_t \cdot \boldsymbol{x}_t - y_t)\, \boldsymbol{x}_t \qquad \text{[WH]}$$

Exponentiated Gradient Algorithm [KW]

($\boldsymbol{w}$ is probability vector)

$$\boldsymbol{w}_{t+1,i} = w_{t,i} \ \exp\left[-\eta \frac{\partial L_t(\boldsymbol{w}_t)}{\partial w_{t,i}}\right] / \text{ normaliz.}$$

Continous Updates

Gradient Descent

$\boldsymbol{w} \in \mathbf{R}^n$

$$\dot{\boldsymbol{w}}_t = -\eta \nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_t)$$

Unnormalized Exponentiated Gradient Alg.

$\boldsymbol{w} \geq \mathbf{0}$

$$\dot{\log}(\boldsymbol{w}_t) = -\eta \nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_t)$$

[WJ]

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

$F(\boldsymbol{w})$ convex

$$\underbrace{\dot{\overline{f(\boldsymbol{w}_t)}}}_{\theta_t} = -\eta \nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_t)$$

Alg.	$f(\boldsymbol{w})$	Domain of $\boldsymbol{w}$
GD	$f(\boldsymbol{w}) = \boldsymbol{w}$	$\boldsymbol{w} \in \mathbf{R}^n$
EGU	$f(\boldsymbol{w}) = \log \boldsymbol{w}$	$\boldsymbol{w} \in [0, \infty)^n$
EG	$f(\boldsymbol{w}) = \ln \frac{\boldsymbol{w}}{1 - \|\|\boldsymbol{w}\|\|_1}$	$\boldsymbol{w} \in [0,1]^{n-1}$, $\|\|\boldsymbol{w}\|\|_1 \leq 1$
BEG	$f(\boldsymbol{w}) = \ln \frac{\boldsymbol{w}}{\mathbf{1} - \boldsymbol{w}}$	$\boldsymbol{w} \in [0,1]^n$

f is barrier function

Discretization

$$\frac{\boldsymbol{f}(\boldsymbol{w}_{t+h}) - \boldsymbol{f}(w_t)}{h} = -\eta \nabla L_t(\boldsymbol{w}_t)$$

$$w_{t+h} = \boldsymbol{f}^{-1}\left(\boldsymbol{f}(w_t) - \eta\, h \nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_t)\right)$$

We use $h = 1$

$$\boldsymbol{w}_{t+1} = \boldsymbol{f}^{-1}\left(\boldsymbol{f}(\boldsymbol{w}_t) - \eta_t \nabla L_t(\boldsymbol{w}_t)\right)$$

Conjecture: Forward Euler better:

Replace $\nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_t)$ by $\nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_{t+h})$

Alternate Motivation [KW]

GD

$$\boldsymbol{w}_{t+1} = \text{argmin}_{\boldsymbol{w}} \left(||\boldsymbol{w} - \boldsymbol{w}_t||_2^2/2 + \eta(y_t - \boldsymbol{w} \cdot \boldsymbol{x}_t)^2/2 \right)$$

$$= \boldsymbol{w}_t - \eta(\underbrace{\boldsymbol{w}_{t+1} \cdot \boldsymbol{x}_t}_{\approx \boldsymbol{w}_t \cdot \boldsymbol{x}_t} - y_t)\, \boldsymbol{x}_t$$

EG

$$\boldsymbol{w}_{t+1} = \text{argmin}_{\boldsymbol{w} \in .} \left(\sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + \eta(y_t - \boldsymbol{w} \cdot \boldsymbol{x}_t)^2/2 \right)$$

$$= w_{t,i} \; \exp \left[-\eta \,(\underbrace{\boldsymbol{w}_{t+1} \cdot \boldsymbol{x}_t}_{\approx \boldsymbol{w}_t \cdot \boldsymbol{x}_t} - y_t)\, x_{t,i} \right] / \text{ normaliz}$$

Families of update algorithms

parameter divergence	name of family	update algs.
$\|\|\boldsymbol{w} - \boldsymbol{w}_t\|\|_2^2$	Grad. Desc.	Widrow Hoff (LMS) Lin. Least Squ. Backprop. Perceptron Alg. kernel based algs.,...
$\sum_{i=1}^n w_i \ln \frac{w_i}{w_{t,i}}$	Exponentiated Gradient Alg.	expert algs Normalized Winnow "AdaBoost"

Families of update algorithms (cont)

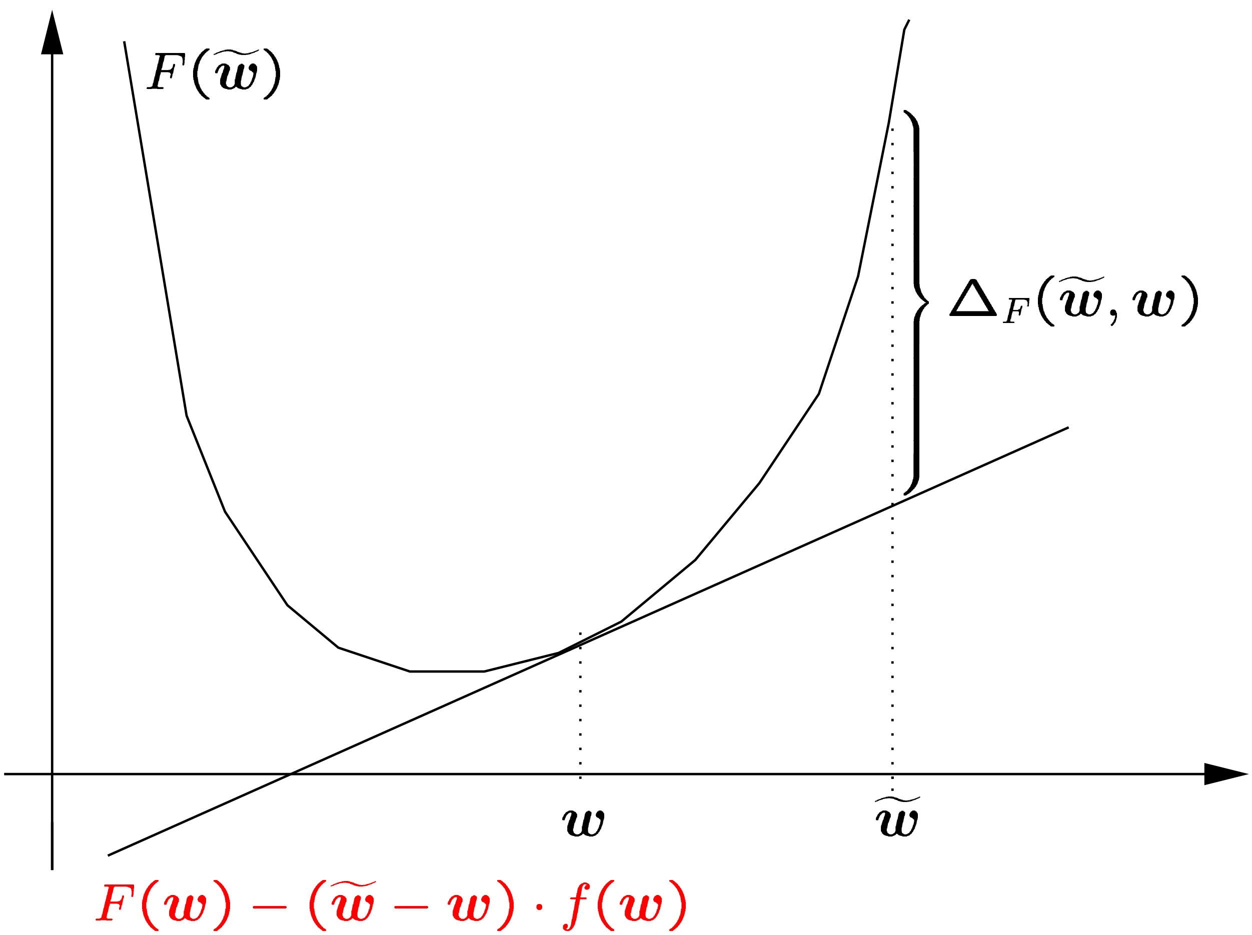
$\sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + w_{t,i} - w_i$	Unnormalized Exp. Grad. Alg.	Winnow
$\sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + (1 - w_i) \ln \frac{1-w_i}{1-w_{t,i}}$	Binary Exp. Grad. Alg.	
any Bregman divergence	-	-

Members of different families exhibit different behavior

Bregman Divergences [Br,CL,Cs]

For any differentiable convex function F

$$\begin{aligned}\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) &= F(\widetilde{\boldsymbol{w}}) - \text{supporting hyperplane through } (\boldsymbol{w}, F(\boldsymbol{w})) \\ &= F(\widetilde{\boldsymbol{w}}) - F(\boldsymbol{w}) - (\widetilde{\boldsymbol{w}} - \boldsymbol{w}) \cdot \underbrace{\nabla_{\boldsymbol{w}} F(\boldsymbol{w})}_{f(\boldsymbol{w})}\end{aligned}$$



Bregman Divergences, Simple Properties

[AW]

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

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

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

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

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

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

7. $\Delta_F(\boldsymbol{w}_1, \boldsymbol{w}_2) + \Delta_F(\boldsymbol{w}_2, \boldsymbol{w}_3)$
   $= \Delta_F(\boldsymbol{w}_1, \boldsymbol{w}_3) + (\boldsymbol{w}_1 - \boldsymbol{w}_2) \cdot (f(\boldsymbol{w}_3) - f(\boldsymbol{w}_2))$

Examples

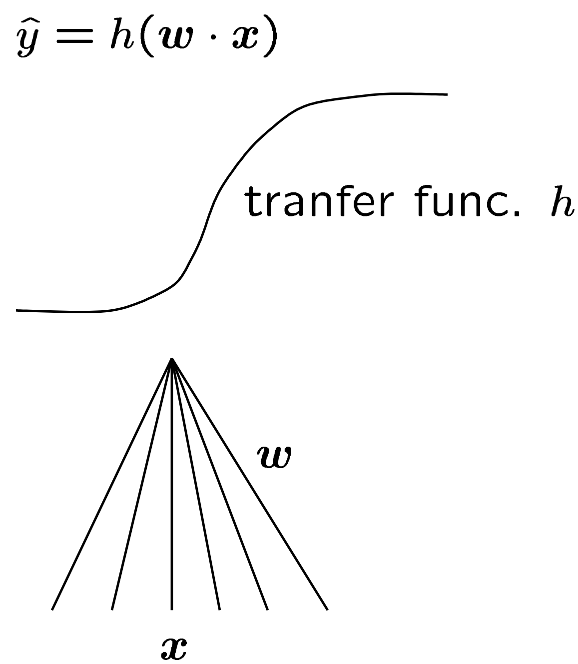
Squared Euclidean Distance

$$
\begin{aligned}
F(\boldsymbol{w}) &= ||\boldsymbol{w}||_2^2/2 \\
f(\boldsymbol{w}) &= \boldsymbol{w} \\
\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) &= ||\widetilde{\boldsymbol{w}}||_2^2/2 - ||\boldsymbol{w}||_2^2/2 - (\widetilde{\boldsymbol{w}} - \boldsymbol{w}) \cdot \boldsymbol{w} \\
&= ||\widetilde{\boldsymbol{w}} - \boldsymbol{w}||_2^2/2
\end{aligned}
$$

(Unnormalized) Relative Entropy

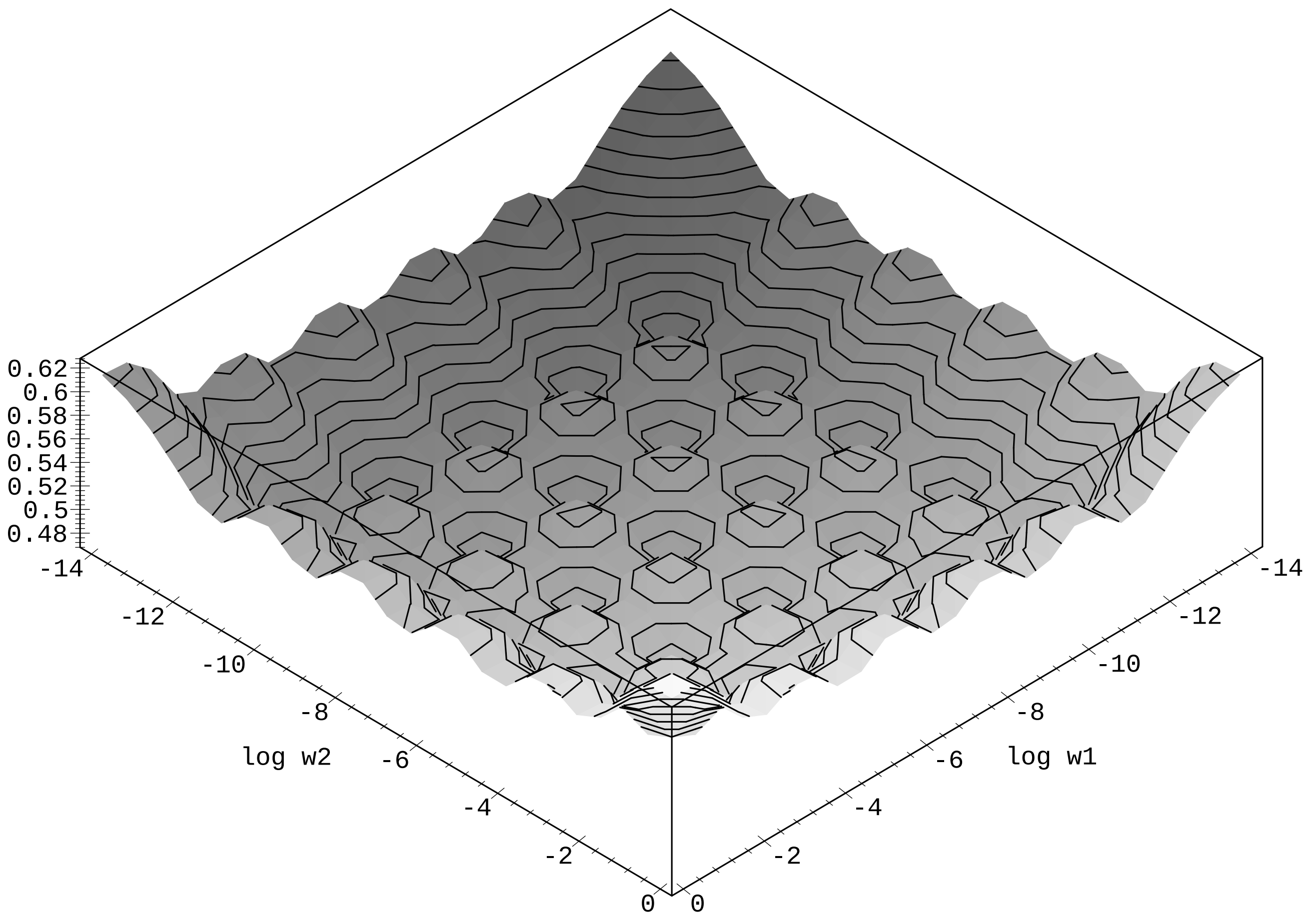
$$
\begin{aligned}
F(\boldsymbol{w}) &= \sum_i (w_i \ln w_i - w_i) \\
f(\boldsymbol{w}) &= \ln \boldsymbol{w} \\
\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) &= \sum_i \left(\widetilde{w_i} \ln \frac{\widetilde{w_i}}{w_i} + w_i - \widetilde{w_i} \right)
\end{aligned}
$$

Bregman Divergences
Lead to Good Loss Functions



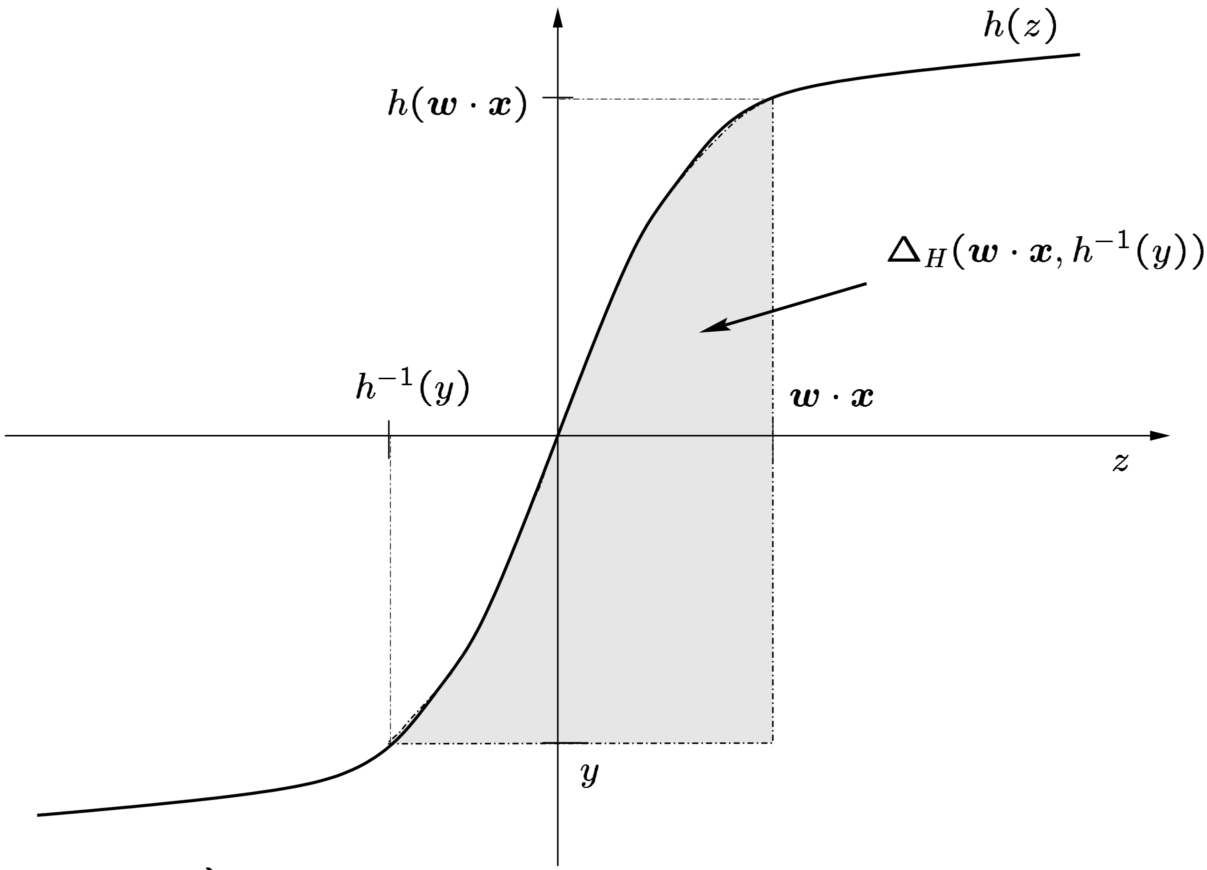
- Sigmoid function $h(z) = \frac{1}{1+e^{-z}}$

- For a set of examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)$
  total loss $\sum_{t=1}^{T}(h(\boldsymbol{w} \cdot \boldsymbol{x}) - y_t)^2/2$
  can have exponentially many minima
  in weight space [Bu,AHW]



Want loss that is convex in $\boldsymbol{w}$

Bregman Divergences
Lead to Good Loss Functions (cont)



$(h = \nabla H)$

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

$$= H(\boldsymbol{w}\cdot\boldsymbol{x}) - H(h^{-1}(y)) - (\boldsymbol{w}\cdot\boldsymbol{x} - h^{-1}(y))\,y$$

$$= \Delta_H(\boldsymbol{w}\cdot\boldsymbol{x}, h^{-1}(y))$$

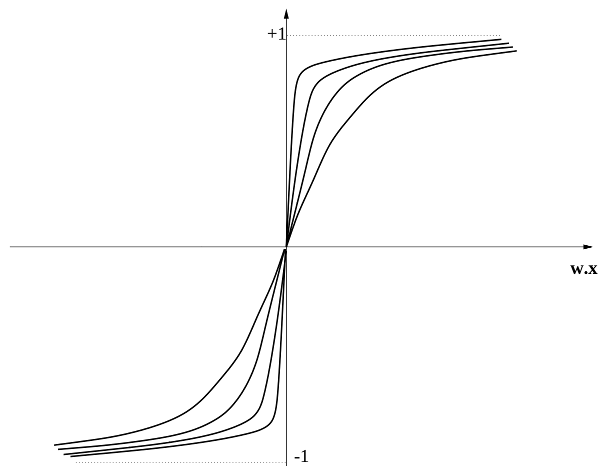
Use $\Delta_H(\boldsymbol{w} \cdot \boldsymbol{x}, h^{-1}(y))$ as loss of $\boldsymbol{w}$ on $(\boldsymbol{x}, y)$

Called matching loss for h [AHW,HKW]

Matching loss is convex in $\boldsymbol{w}$

transfer f. $h(z)$	$H(z)$	match. loss $d_H(\boldsymbol{w} \cdot \boldsymbol{x}, h^{-1}(y))$
z	$\frac{1}{2}z^2$	$\frac{1}{2}(\boldsymbol{w} \cdot \boldsymbol{x} - y)^2$ square loss
$\frac{e^z}{1+e^z}$	$\ln(1 + e^z)$	logistic loss
$\text{sign}(z)$	$\|z\|$	$\max\{0, -y\boldsymbol{w} \cdot \boldsymbol{x}\}$ hinge loss

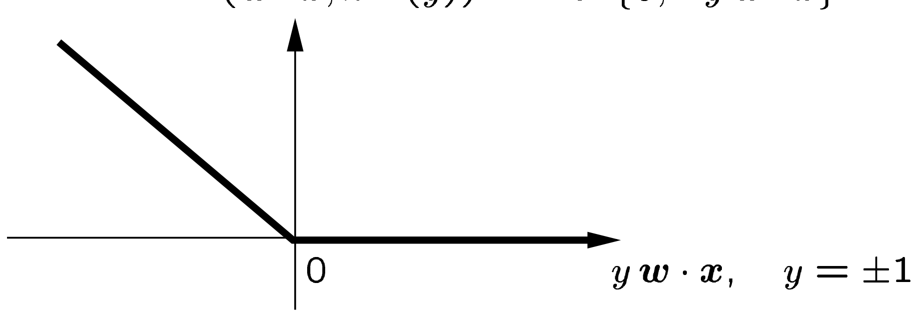
For transfer function $h(z) = \mathrm{sign}(z)$



$H(z) = |z|$

Matching loss is hinge loss [GW]

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



Convex in $\boldsymbol{w}$ but not differentiable

Motivation of linear threshold algs

Gradient descent with Hinge Loss	Perceptron
Expon. gradient with Hinge Loss	Normalized Winnow

Known linear threshold algs for ± 1-class are gradient-based algs with hinge loss

Trade-off between two divergences [KW]

$$\boldsymbol{w}_{t+1}$$
$$= \text{argmin}_{\boldsymbol{w}}(\underbrace{\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)}_{\substack{\text{parameter}\\ \text{divergence}}} + \eta \underbrace{\Delta_H(\boldsymbol{w} \cdot \boldsymbol{x}_t, h^{-1}(y_t))}_{\substack{\text{matching}\\ \text{loss}}})$$

Both divergences are convex in $\boldsymbol{w}$

$$\boldsymbol{w}_{t+1} = f^{-1}\left(f(\boldsymbol{w}_t) - \eta(h(\boldsymbol{w}_t \cdot \boldsymbol{x}_t) - y_t)\boldsymbol{x}_t\right)$$

Generalization of the "delta"-rule

Projections

$$\boldsymbol{w}_{t+1} = \text{argmin}_{\boldsymbol{w}}(\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t) + \eta(\boldsymbol{w} \cdot \boldsymbol{x}_t - y_t)^2)$$

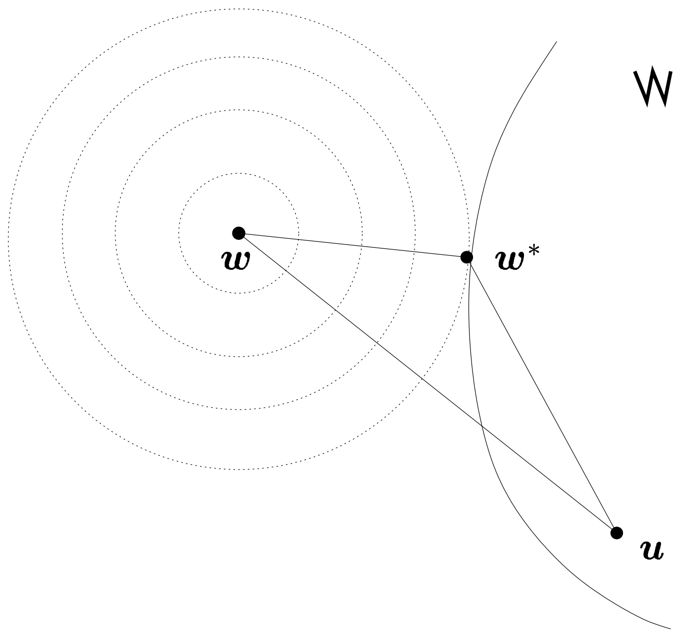
When η is large then $\boldsymbol{w}_{t+1}$ is projection of $\boldsymbol{w}_t$ onto plane $\boldsymbol{w} \cdot \boldsymbol{x}_t = y_t$

$$\boldsymbol{w}_{t+1} = \text{argmin}_{\{\boldsymbol{w} \,:\, \boldsymbol{w}_t \cdot \boldsymbol{x}_t = y_t\}} \ \Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)$$

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

[La,KW]

A Pythagorean Theorem [Br,Cs,A,HW]



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

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

Th:

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

Kernel trick [BGV]

- Prediction determined by $\boldsymbol{w} \cdot \boldsymbol{x}$
- $\boldsymbol{w}$ is linear combination of past examples

$$\boldsymbol{x} \;\rightarrow\; \Phi(\boldsymbol{x})$$

$$\underbrace{\left(\sum_t \alpha_t \Phi(\boldsymbol{x}_t)\right)}_{\boldsymbol{w}} \cdot \Phi(\boldsymbol{x}) \;=\; \sum_t \alpha_t \underbrace{\Phi(\boldsymbol{x}_t) \cdot \Phi(\boldsymbol{x})}_{K(\boldsymbol{x}_t, \boldsymbol{x})}$$

When trick applicable?

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

Rotation invariance $\Rightarrow$ $\boldsymbol{w}$ lin. comb. of ex.
[KWA]

When applicable (cont.)?

Bregman update:

$$\begin{aligned} f(\boldsymbol{w}_T) &= f(\boldsymbol{w}_1) + \eta \sum_t \nabla_{\boldsymbol{w}} L(y_t, h(\boldsymbol{w}_t \cdot \boldsymbol{x}_t)) \\ &= \underbrace{f(\boldsymbol{w}_1)}_{0} + \text{lin. comb. of ex.} \end{aligned}$$

geometric	information theoretic
rotation invariant	no
$f = id$	$f = \log$
kernels	???

How do we prove relative loss bounds?

Loss: $L_t(\boldsymbol{w}) = L((\boldsymbol{x}_t, y_t), \boldsymbol{w})$ convex in $\boldsymbol{w}$

Divergence: $\Delta_F(\boldsymbol{u}, \boldsymbol{w})$

Update: $f(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}_t) = -\eta \nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_t)$

$$L_t(\boldsymbol{u}) \overbrace{\geq}^{\text{convexity}} L_t(\boldsymbol{w}_t) + (\boldsymbol{u} - \boldsymbol{w}_t) \cdot \underbrace{\nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_t)}_{\text{update}}$$

$$= L_t(\boldsymbol{w}_t) - \frac{1}{\eta} \underbrace{(\boldsymbol{u} - \boldsymbol{w}_t) \cdot (f(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}_t))}_{\text{prop. 7 of } \Delta_F}$$

$$= L_t(\boldsymbol{w}_t) + \frac{1}{\eta} \Big(\Delta_F(\boldsymbol{u}, \boldsymbol{w}_{t+1}) - \Delta_F(\boldsymbol{u}, \boldsymbol{w}_t) - \Delta_F(\boldsymbol{w}_t, \boldsymbol{w}_{t+1}) \Big)$$

Summing over t [WJ,KW]

$$\begin{aligned}\sum_t L_t(\boldsymbol{w}_t) \le \sum_t & L_t(\boldsymbol{u}) \\ & + \frac{1}{\eta}\sum_t\Big(\Delta_F(\boldsymbol{u},\boldsymbol{w}_t) - \Delta_F(\boldsymbol{u},\boldsymbol{w}_{t+1}) \\ & \qquad + \Delta_F(\boldsymbol{w}_t,\boldsymbol{w}_{t+1})\Big) \\ \le & \sum_t L_t(\boldsymbol{u}) \\ & + \frac{1}{\eta}\Big(\Delta_F(\boldsymbol{u},\boldsymbol{w}_1) - \underbrace{\Delta_F(\boldsymbol{u},\boldsymbol{w}_{T+1})}_{\ge 0}\Big) \\ & \qquad + \frac{1}{\eta}\sum_t \Delta_F(\boldsymbol{w}_t,\boldsymbol{w}_{t+1})\end{aligned}$$

$$\begin{aligned}\sum_t L_t(\boldsymbol{w}_t) \le \sum_t L_t(\boldsymbol{u}) & + \frac{1}{\eta}\Delta_F(\boldsymbol{u},\boldsymbol{w}_1) \\ & + \frac{1}{\eta}\sum_t \Delta_F(\boldsymbol{w}_t,\boldsymbol{w}_{t+1})\end{aligned}$$

Any convex loss and any Bregman divergence!

Key step:

Relate $\Delta_F(\boldsymbol{w}_t, \boldsymbol{w}_{t+1})$ to loss $L_t(\boldsymbol{w}_t)$

Loss & divergence dependent

Get $\Delta_F(\boldsymbol{w}_t, \boldsymbol{w}_{t+1}) \leq \text{const}\, L_t(\boldsymbol{w}_t)$

Then solve for $\sum_t L_t(\boldsymbol{w}_t)$

Yield bounds of the form

$$\sum_t L_t(\boldsymbol{w}_t) \leq a \sum_t L_t(\boldsymbol{u}) + b\, \Delta_F(\boldsymbol{u}, \boldsymbol{w}_1)$$

a, b constants, $a > 1$.

Regret bounds ($a = 1$):

time changing η, subtler analysis [AW]

Some Bounds [KW,GLS,GL]
(Linear Regression with Square Loss)

Gradient Descent

$$\sum_t L_t(\boldsymbol{w}_t) \le (1+c) \sum_t L_t(\boldsymbol{u}) + \frac{1+c}{c} X_2^2 U_2^2$$

$||\boldsymbol{x}_t||_2 \le X_2$, $||\boldsymbol{u}||_2 \le U_2$, $c > 0$

Scaled Exponentiated Gradient

$$\sum_t L_t(\boldsymbol{w}_t) \le (1+c) \sum_t L_t(\boldsymbol{u}) + \frac{1+c}{c} \ln n \, X_\infty^2 U_1^2$$

$||\boldsymbol{x}_t||_\infty \le X_\infty$, $||\boldsymbol{u}||_1 \le U_1$, $c > 0$

p-norm alg

$$\sum_t L_t(\boldsymbol{w}_t) \le (1+c) \sum_t L_t(\boldsymbol{u}) + \frac{1+c}{c} (p-1) \, X_p^2 U_q^2$$

$||\boldsymbol{x}_t||_p \le X_p$, $||\boldsymbol{u}||_q \le U_q$, $c > 0$

Hadamard example [KW]

$$\underbrace{\begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}}_{\substack{H \\ \text{instances}}} \underbrace{\begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix}}_{\substack{\text{unit} \\ \text{target}}} = \underbrace{\begin{pmatrix} 1 \\ 1 \\ -1 \\ -1 \end{pmatrix}}_{\text{labels}}$$

Rotation invariant alg.	$O(n)$ total loss
Information theoretic alg.	$O(\log n)$ 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

GD	$\underbrace{\vert\vert\boldsymbol{x}\vert\vert_2}_{\sqrt{n}}{}^2\underbrace{\vert\vert\boldsymbol{w}\vert\vert_2}_{1}{}^2 = n$
EG	$\underbrace{\vert\vert\boldsymbol{x}\vert\vert_\infty}_{1}{}^2\underbrace{\vert\vert\boldsymbol{w}\vert\vert_1}_{1}{}^2 \log n = \log n$

$F(w) = \frac{||\boldsymbol{w}||_2^2}{2}$ only rotation invariant norm

Where do Bregman divergences come from?

- Exponential family of distributions

- Inherent duality

Exponential Family of Distributions

- Parametric density functions

$$P_G(\boldsymbol{x}|\boldsymbol{\theta}) = e^{\boldsymbol{\theta}\cdot\boldsymbol{x} - G(\boldsymbol{\theta})}\, P_0(\boldsymbol{x})$$

- $\boldsymbol{\theta}$ and $\boldsymbol{x}$ vectors in R^d

- Cumulant function $G(\boldsymbol{\theta})$
  assures normalization

$$G(\boldsymbol{\theta}) = \ln \int e^{\boldsymbol{\theta}\cdot\boldsymbol{x}}\, P_0(\boldsymbol{x})\, d\boldsymbol{x}$$

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

- G characterizes members of the family

- $\boldsymbol{\theta}$ is *natural* parameter

Bregman Divergences as Relative Entropies between Exponential Distributions

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

$$\Delta_G(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta})$$

$$= \int_{\boldsymbol{x}} P_G(\boldsymbol{x}|\boldsymbol{\theta}) \ln \frac{P_G(\boldsymbol{x}|\boldsymbol{\theta})}{P_G(\boldsymbol{x}|\widetilde{\boldsymbol{\theta}})} d\boldsymbol{x}$$

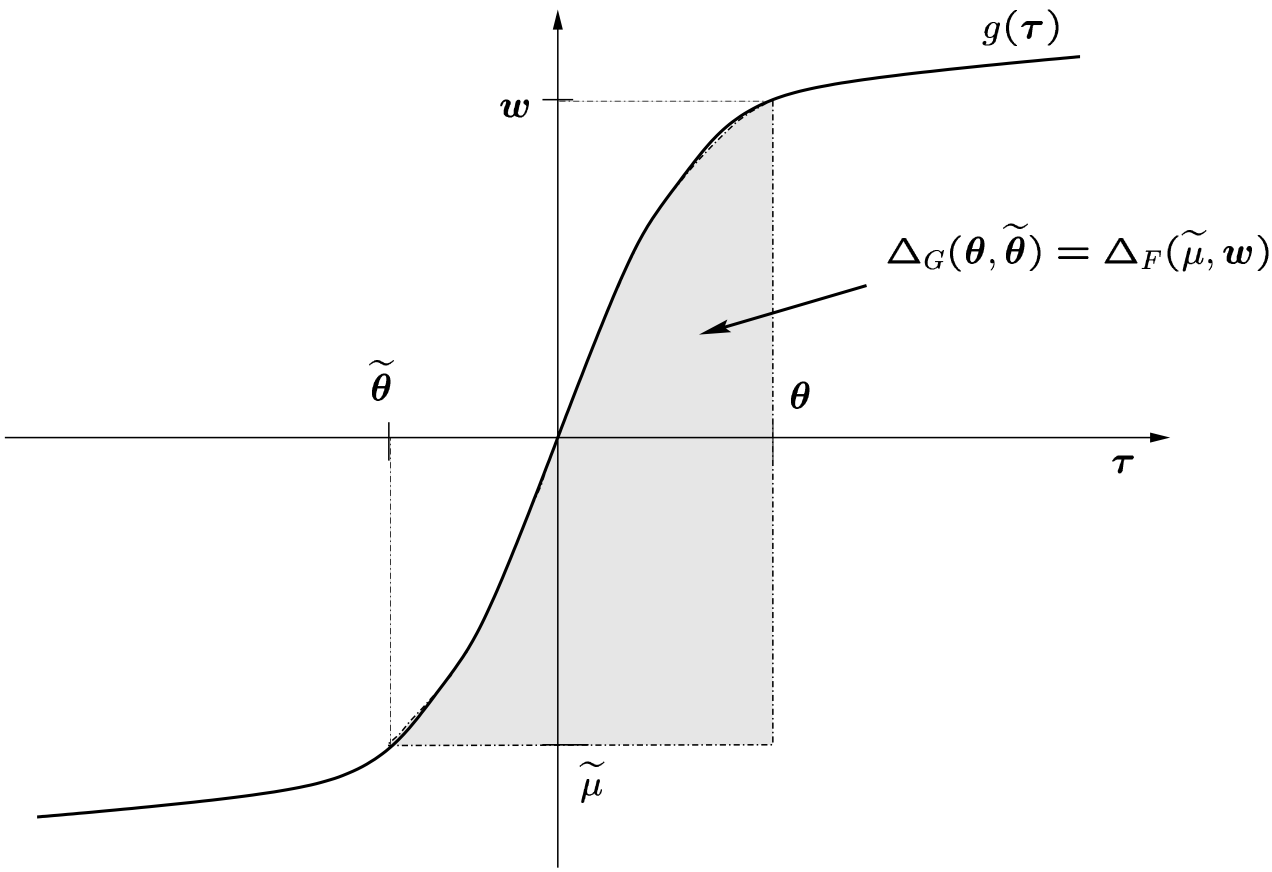
$$= G(\widetilde{\boldsymbol{\theta}}) - G(\boldsymbol{\theta}) - (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \cdot \underbrace{g(\boldsymbol{\theta})}_{\boldsymbol{w}}$$

$$\overset{F(\boldsymbol{w}) = \boldsymbol{\theta} \cdot \boldsymbol{w} - G(\boldsymbol{\theta})}{=} F(\boldsymbol{w}) - F(\widetilde{\boldsymbol{w}}) - (\boldsymbol{w} - \widetilde{\boldsymbol{w}}) \cdot \underbrace{f(\widetilde{\boldsymbol{w}})}_{\widetilde{\boldsymbol{\theta}}}$$

$$= \Delta_F(\boldsymbol{w}, \widetilde{\boldsymbol{w}})$$

[A,BN,AW]

Area unchanged When Slide Flipped



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 [TW]

Learner against adversary

$$\inf_{\boldsymbol{\theta}_1} \sup_{\boldsymbol{x}_1} \inf_{\boldsymbol{\theta}_2} \sup_{\boldsymbol{x}_2} \inf_{\boldsymbol{\theta}_3} \sup_{\boldsymbol{x}_3} \ldots \inf_{\boldsymbol{\theta}_T} \sup_{\boldsymbol{x}_T}$$

$$\underbrace{\sum_{t=1}^{T} \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{x}_t)^2}_{\text{total loss of on-line algorithm}} \quad - \quad \underbrace{\inf_{\boldsymbol{\theta}} \left(\sum_{t=1}^{T} \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{x}_t)^2 \right)}_{\text{total loss of off-line algorithm}}$$

Instances must be bounded: $||\boldsymbol{x}_t||_2 \leq X$

Minimax algorithm usually intractable

Bernoulli is another exception [Sh]

Gaussian

Max likelihood	$\boldsymbol{\theta}_t = \frac{\sum_{q=1}^{t-1} \boldsymbol{x}_q}{t-1}$
Forward Alg.	$\boldsymbol{\theta}_t = \frac{\sum_{q=1}^{t-1} \boldsymbol{x}_q}{t-1+1}$
Bound	$\frac{1}{2}X^2(1+\ln T)$
Minimax Alg.	$\boldsymbol{\theta}_t = \frac{\sum_{q=1}^{t-1} \boldsymbol{x}_q}{t+\ln T - \ln(t+O(\ln T))}$
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 [Fo,TW]

$$\begin{aligned} \boldsymbol{\theta}_t &= \operatorname*{arginf}_{\boldsymbol{\theta}} \sup_{\boldsymbol{x}_t} \sum_{q=1}^{t} L_q(\boldsymbol{\theta}_q) - \inf_{\boldsymbol{\theta}} L_{1..t}(\boldsymbol{\theta}) \\ &= \operatorname*{arginf}_{\boldsymbol{\theta}} \sup_{\boldsymbol{x}_t} L_t(\boldsymbol{\theta}_t) - \inf_{\boldsymbol{\theta}} L_{1..t}(\boldsymbol{\theta}) \end{aligned}$$

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 [P]

Relative entropy is regularizer and barrier

Conjecture:

$$\begin{array}{c}\text{argmin}_{\boldsymbol{w}} \\ |\boldsymbol{w}|_1 = 1 \\ \text{constraints}\end{array} \sum_i w_i \log \frac{w_i}{1/n}$$

and

$$\begin{array}{c}\text{argmin}_{\boldsymbol{w}} \\ |\boldsymbol{w}|_1 = 1 \\ \text{constraints} \\ w_i \geq 0\end{array} \sum_i w_i^2$$

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 (ϵ, δ)-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