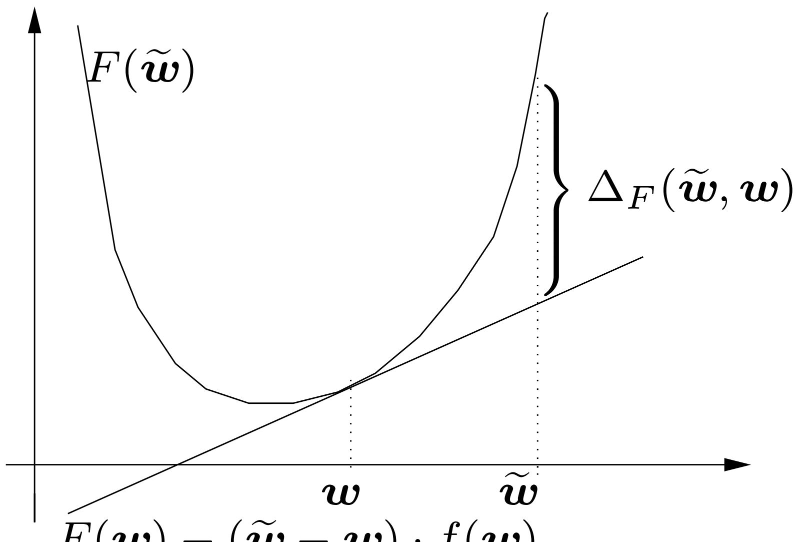
Bregman Divergences [Br,CL,Cs]

For any differentiable convex function F

$$\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) = F(\widetilde{\boldsymbol{w}}) - F(\boldsymbol{w}) - (\widetilde{\boldsymbol{w}} - \boldsymbol{w}) \cdot \underbrace{\nabla_{\boldsymbol{w}} F(\boldsymbol{w})}_{f(\boldsymbol{w})}$$

$$= F(\widetilde{\boldsymbol{w}}) - \begin{array}{c}\text{supporting hyperplane}\\ \text{through } (\boldsymbol{w}, F(\boldsymbol{w}))\end{array}$$



General Motivation of Updates [KW]

Trade-off between two term:

$$\boldsymbol{w}_{t+1} = \operatorname*{argmin}_{\boldsymbol{w}}(\underbrace{\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)}_{weight\ domain} + \eta_t \underbrace{L_t(\boldsymbol{w})}_{label\ domain})$$

$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)$ is "regularization term" and serves as measure of progress in the analysis.

When loss L is convex (in $\boldsymbol{w}$)

$$\nabla_{\boldsymbol{w}}(\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t) + \eta_t L_t(\boldsymbol{w})) = 0$$

iff

$$f(\boldsymbol{w}) - f(\boldsymbol{w}_t) + \eta_t \underbrace{\nabla L_t(\boldsymbol{w})}_{\approx \nabla L_t(\boldsymbol{w}_t)} = 0$$

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

Bregman Divergences: Simple Properties

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. Usually not symmetric: $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) \neq \Delta_F(\boldsymbol{w}, \widetilde{\boldsymbol{w}})$
4. Linearity (for $a \geq 0$):
   $\Delta_{F+a\,H}(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) = \Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) + a\,\Delta_H(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$
5. Unaffected by linear terms ($a \in \mathbf{R}$, $\boldsymbol{b} \in \mathbf{R}^n$):
   $\Delta_{H+a\widetilde{\boldsymbol{w}}+\boldsymbol{b}}(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) = \Delta_H(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$

Bregman Divergences: More Properties

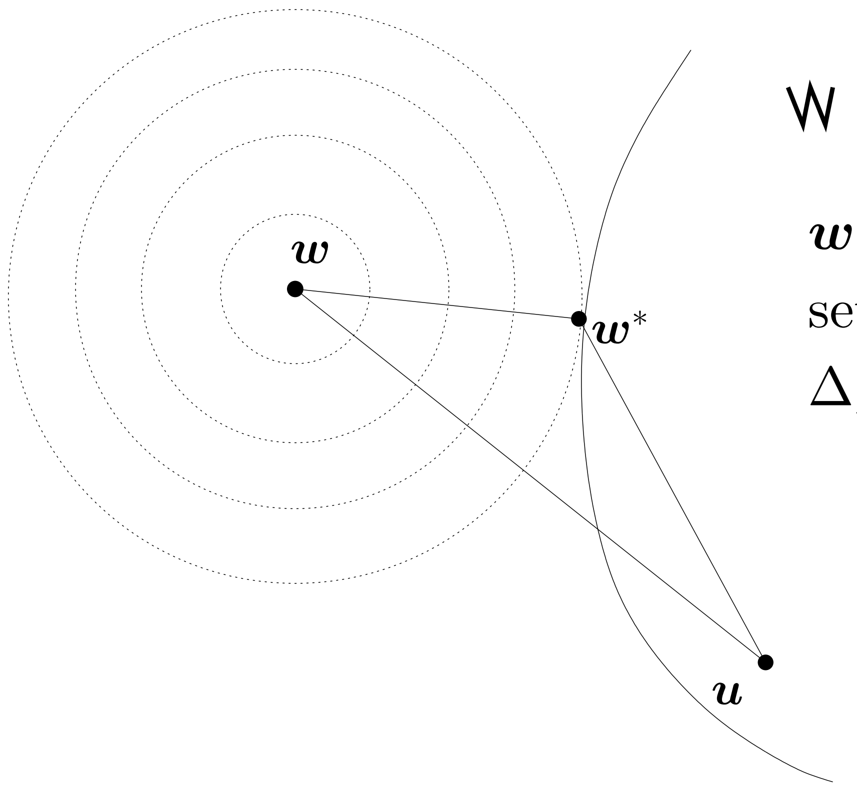
6. $\nabla_{\widetilde{\boldsymbol{w}}} \Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$

$$\begin{aligned} &= \nabla F(\widetilde{\boldsymbol{w}}) - \nabla_{\widetilde{\boldsymbol{w}}}(\widetilde{\boldsymbol{w}} \nabla_{\boldsymbol{w}} F(\boldsymbol{w})) \\ &= f(\widetilde{\boldsymbol{w}}) - f(\boldsymbol{w}) \end{aligned}$$

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

$$\begin{aligned} = \quad & F(\boldsymbol{w}_1) - F(\boldsymbol{w}_2) - (\boldsymbol{w}_1 - \boldsymbol{w}_2) f(\boldsymbol{w}_2) \\ & F(\boldsymbol{w}_2) - F(\boldsymbol{w}_3) - (\boldsymbol{w}_2 - \boldsymbol{w}_3) f(\boldsymbol{w}_3) \\ = \quad & \Delta_F(\boldsymbol{w}_1, \boldsymbol{w}_3) + (\boldsymbol{w}_1 - \boldsymbol{w}_2) \cdot (f(\boldsymbol{w}_3) - f(\boldsymbol{w}_2)) \end{aligned}$$

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



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

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

Theorem:

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

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

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

How to prove relative loss bounds?

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

Divergence: $\Delta_F(\boldsymbol{u}, \boldsymbol{w}) = F(\boldsymbol{u}) - F(\boldsymbol{w}) - (\boldsymbol{u} - \boldsymbol{w}) \cdot f(\boldsymbol{w})$

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

$$
\begin{aligned}
L_t(\boldsymbol{u}) \quad & \overbrace{\geq}^{\text{convexity}} \quad L_t(\boldsymbol{w}_t) + (\boldsymbol{u} - \boldsymbol{w}_t) \cdot \underbrace{\nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_t)}_{\text{update}} \\
& = \quad L_t(\boldsymbol{w}_t) - \frac{1}{\eta} \underbrace{(\boldsymbol{u} - \boldsymbol{w}_t) \cdot (f(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}_t))}_{\text{prop. 7 of } \Delta_F} \\
& = \quad L_t(\boldsymbol{w}_t) + \frac{1}{\eta} (\Delta_F(\boldsymbol{u}, \boldsymbol{w}_{t+1}) - \Delta_F(\boldsymbol{u}, \boldsymbol{w}_t) - \Delta_F(\boldsymbol{w}_t, \boldsymbol{w}_{t+1}))
\end{aligned}
$$

First step: Teleskoping

Summing over t [WJ,KW]

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

Any convex loss and any Bregman divergence!

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

Loss & divergence are 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 [AG]

Bounds for Linear Regression with Square Loss

Gradient Descent

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

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

Scaled Exponentiated Gradient

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

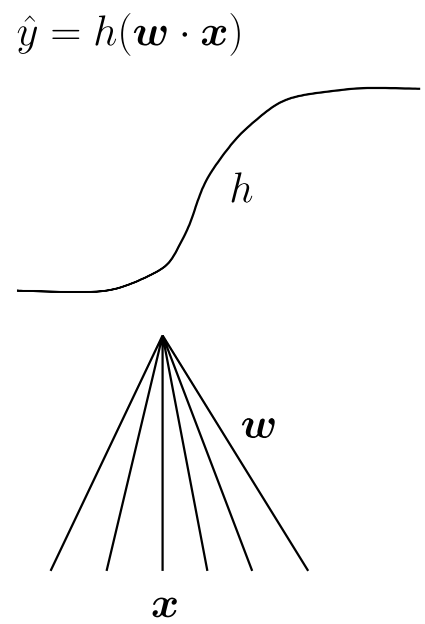
$||\boldsymbol{x}_t||_\infty \leq X_\infty,\ ||\boldsymbol{u}||_1 \leq U_1,\ c > 0$

p-norm Algorithm

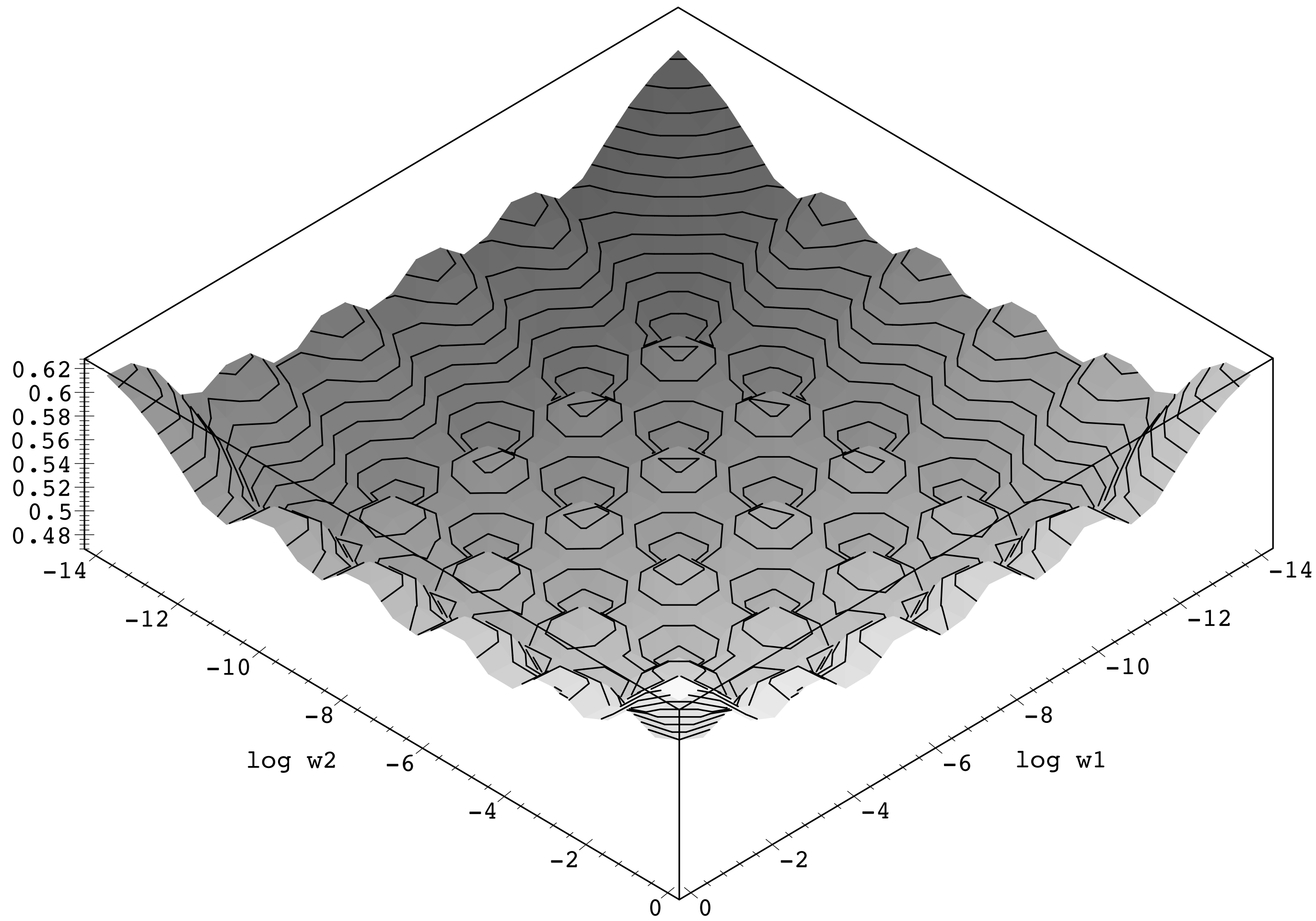
$$\sum_t L_t(\boldsymbol{w}_t) \leq (1+c) \sum_t L_t(\boldsymbol{u}) + \frac{1+c}{c}\, (p-1)\, X_p^2\, U_q^2$$

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

Nonlinear Regression

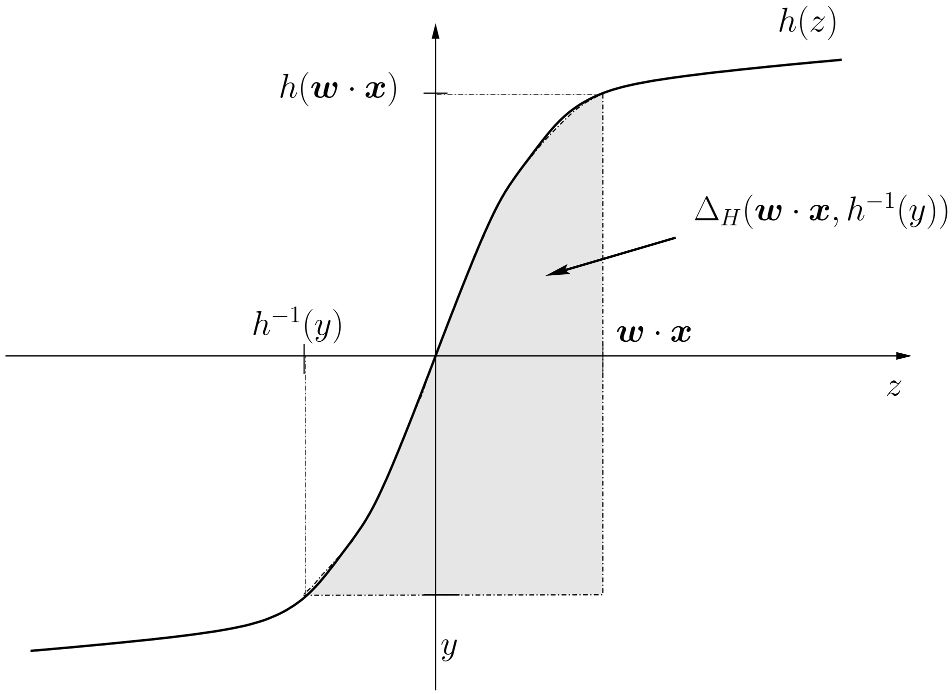


- Sigmoid function $h(z) = \frac{1}{1+e^{-z}}$
- For a set of examples $(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_T, y_T)$
  total loss $\sum_{t=1}^{T} (h(\boldsymbol{w} \cdot \boldsymbol{x}) - y_t)^2/2$
  can have exponentially many minima
  in weight space [Bu,AHW]



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

Bregman Div. Lead to Good Loss Function



$$(h = \nabla H)$$

$$\begin{aligned}\int_{h^{-1}(y)}^{\boldsymbol{w}\cdot\boldsymbol{x}} (h(z) - y)\, dz &= H(\boldsymbol{w}\cdot\boldsymbol{x}) - H(h^{-1}(y)) - (\boldsymbol{w}\cdot\boldsymbol{x} - h^{-1}(y))\, y \\ &= \Delta_H(\boldsymbol{w}\cdot\boldsymbol{x}, h^{-1}(y))\end{aligned}$$

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
$\frac{e^z}{1+e^z}$	$\ln(1+e^z)$	$\ln(1+e^{\boldsymbol{w}\cdot\boldsymbol{x}}) - y\boldsymbol{w} \cdot \boldsymbol{x}$ $+y \ln y + (1-y)\ln(1-y)$ logistic loss
$\mathrm{sign}(z)$	$\lvert z \rvert$	$\max\{0, -y\boldsymbol{w} \cdot \boldsymbol{x}\}$ hinge loss

Idea behind the matching loss

If transfer function and loss match, then

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

Then update has simple form:

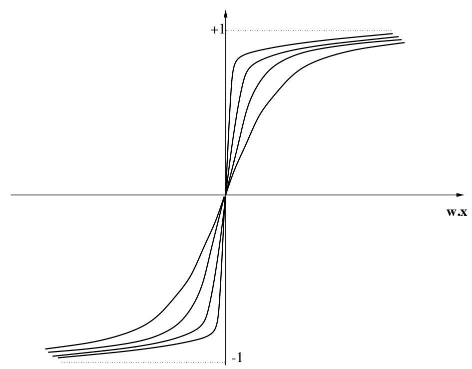
$$f(\boldsymbol{w}_{t+1}) = f(\boldsymbol{w}_t) - \eta_t (h(\boldsymbol{w}_t \cdot \boldsymbol{x}) - y_t) \boldsymbol{x}_t$$

This can be exploited in proofs

But not absolutely necessary
One only needs convexity of $L(h(\boldsymbol{w} \cdot \boldsymbol{x}), y)$ in $\boldsymbol{w}$ [Ce]

Sigmoid in the Limit

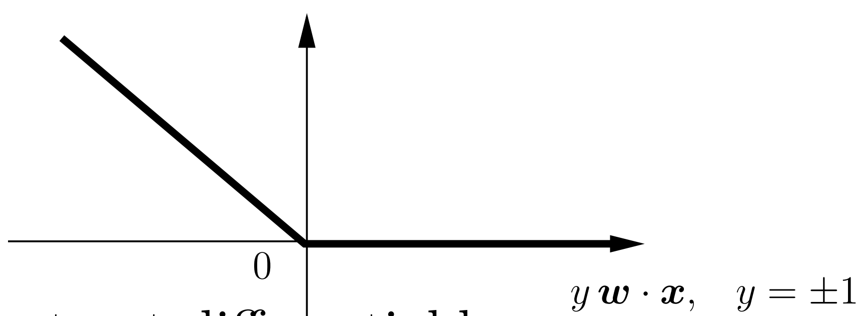
For transfer function $h(z) = \text{sign}(z)$



$H(z) = |z|$

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 algorithms for ± 1-classification case are gradient-based algorithms with hinge loss

Perceptron

$$\boldsymbol{w}_{t+1}$$

$$= \operatorname*{argmin}_{\boldsymbol{w}} \left(||\boldsymbol{w} - \boldsymbol{w}_t||^2/2 + \eta \, HL(\boldsymbol{w} \cdot \boldsymbol{x}_t, g^{-1}(y_t)) \right)$$

$$= \boldsymbol{w}_t - \eta \, (\mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x}_t)) - y_t) \, \boldsymbol{x}_t$$

$$\approx \boldsymbol{w}_t - \eta \, (\underbrace{\mathrm{sign}(\boldsymbol{w}_t \cdot \boldsymbol{x}_t)}_{\hat{y}_t} - y_t) \, \boldsymbol{x}_t$$

Normalized Winnow

$$\boldsymbol{w}_{t+1}$$

$$= \operatorname*{argmin}_{\boldsymbol{w}} \left(\sum_{i=1}^{n} w_i \ln \frac{w_i}{w_{t,i}} + \eta \, HL(\boldsymbol{w} \cdot \boldsymbol{x}_t, g^{-1}(y_t)) \right)$$

$$= w_{t,i} \; e^{-\eta \, (\mathrm{sign}(\boldsymbol{w} \cdot \boldsymbol{x}_t) - y_t) \, x_{t.i}} / \text{ normalization}$$

$$\approx w_{t,i} \; e^{-\eta \, (\underbrace{\mathrm{sign}(\boldsymbol{w}_t \cdot \boldsymbol{x}_t)}_{\hat{y}_t} - y_t) \, x_{t.i}} / \text{ normalization}$$

Trade-off between two divergences [KW]

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

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

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

Generalization of the “delta”-rule

Duality

Special case:

$$\min_{\boldsymbol{w}} \Delta_F(\boldsymbol{w}, \boldsymbol{w}_t) + \Delta_H(\boldsymbol{x}_t \cdot \boldsymbol{w}, h^{-1}(y_t))$$

$$= -\min_{\alpha} \Delta_{\mathcal{G}}(\alpha + y_t, h(0)) + \Delta_{\mathcal{F}}(f(\boldsymbol{w}_t) - \alpha \boldsymbol{x}_t, f(\mathbf{0})) + \text{const.}$$

General:

$$\min_{\boldsymbol{w}} \Delta_F(\boldsymbol{w} + \boldsymbol{\mu}, \underbrace{f^{-1}(\boldsymbol{\phi})}_{\boldsymbol{w}_t}) + \Delta_H(X\boldsymbol{w} + \boldsymbol{\nu}, h^{-1}(\boldsymbol{y}))$$

$$= -\min_{\boldsymbol{\alpha}} \Delta_{\mathcal{G}}(\boldsymbol{\alpha} + \boldsymbol{y}, h(\boldsymbol{\nu})) + \Delta_{\mathcal{F}}(\boldsymbol{\phi} - X^\top \boldsymbol{\alpha}, f(\boldsymbol{\mu})) + \text{const.}$$

where F and $\mathcal{F}$ are convex conjugate functions:

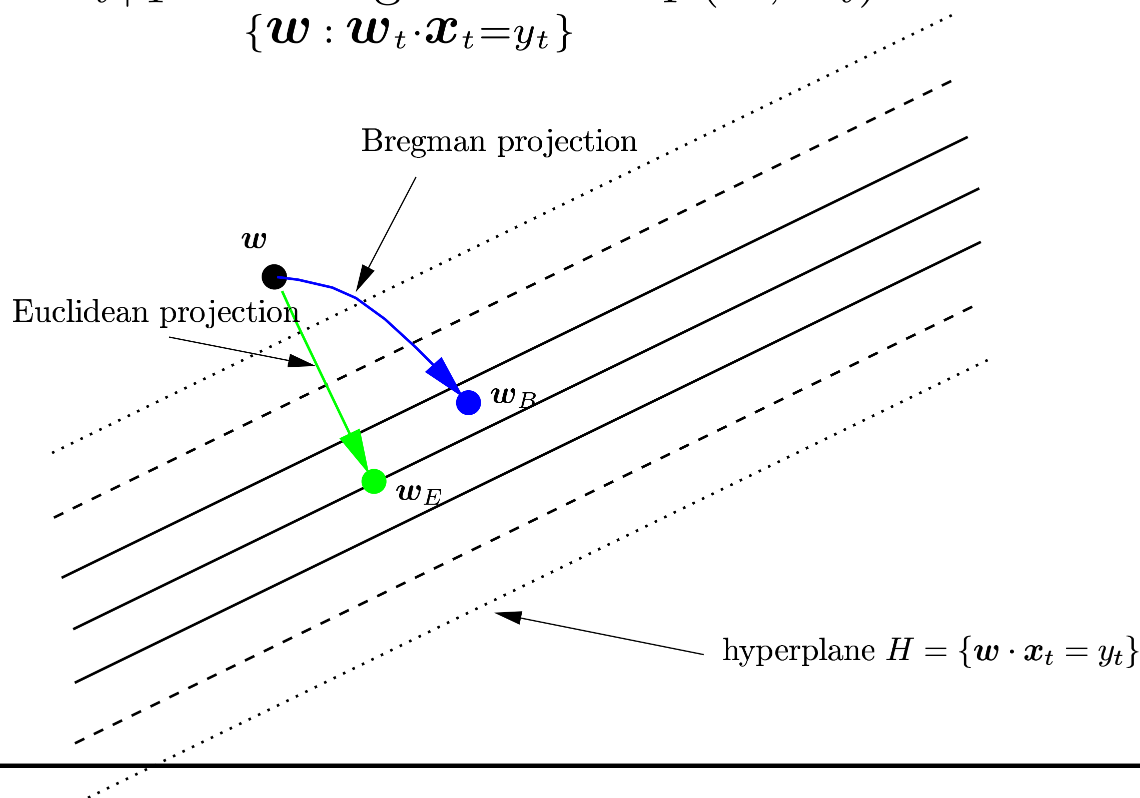
$$\mathcal{F}(\boldsymbol{x}) = \sup_{\boldsymbol{y}} \boldsymbol{x} \cdot \boldsymbol{y} - F(\boldsymbol{y}) = \boldsymbol{x} \cdot (\nabla F)^{-1}(\boldsymbol{x}) - F((\nabla F)^{-1}(\boldsymbol{x}))$$

Projections onto Hyperplanes

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

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

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



Relation to Boosting

The AdaBoost update of the probability vector $\boldsymbol{w}_t$:

$$w_i^{(t+1)} = w_i^{(t)} \exp(-\alpha_t y_i h_t(\boldsymbol{x}_i))$$

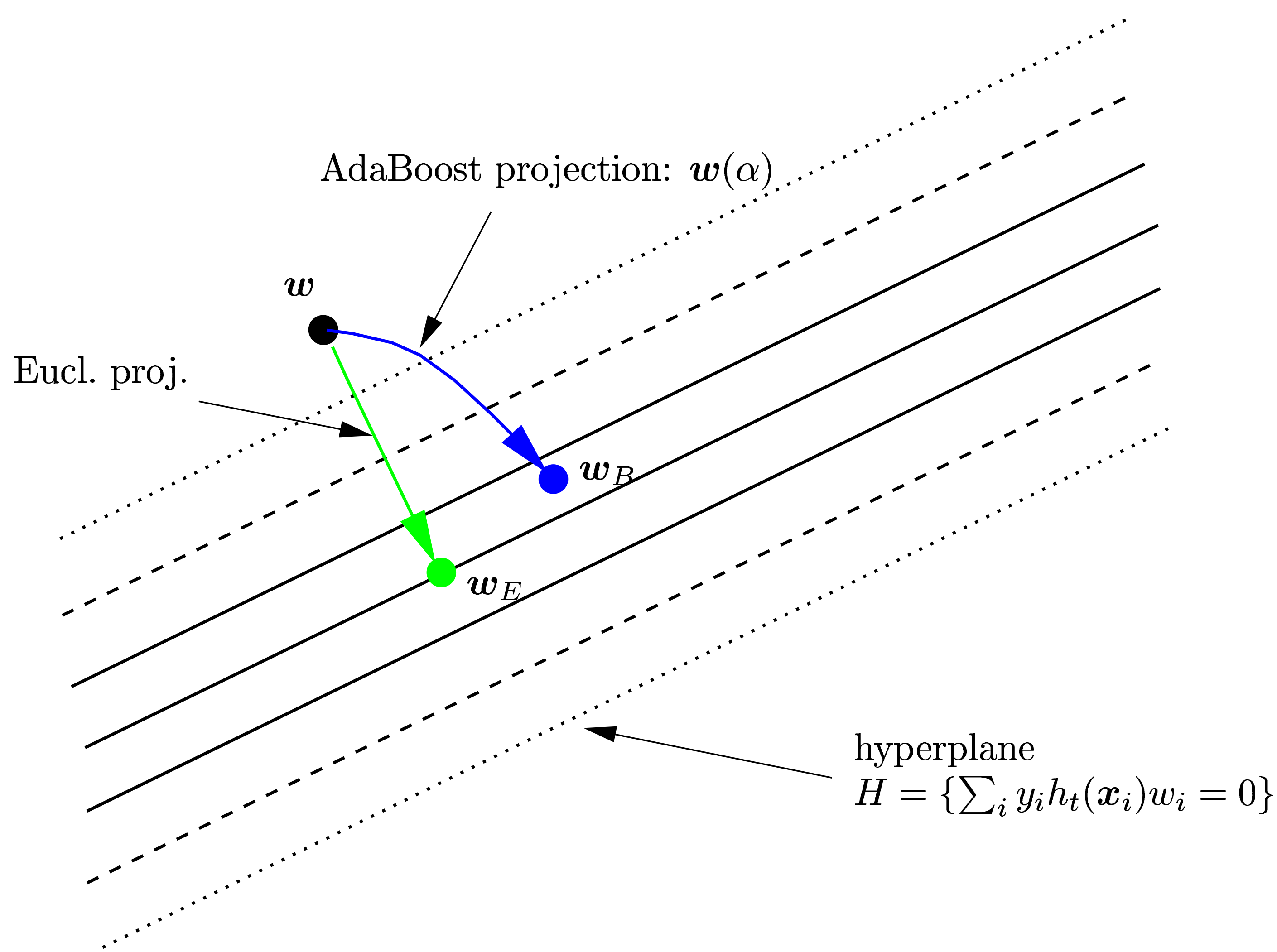
Is a projection w.r.t. divergence [CKW,La,KW,CSS]

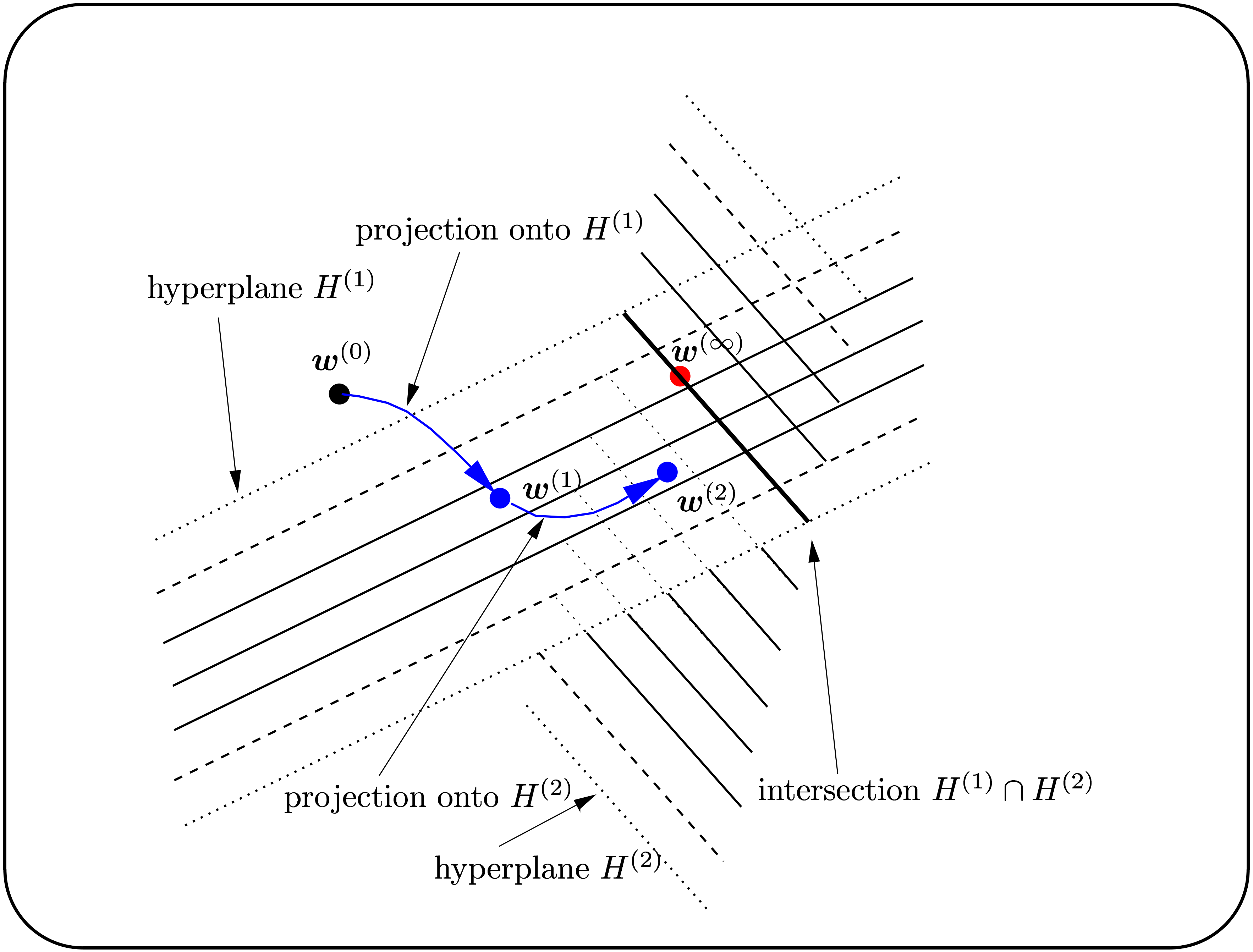
$$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t) = \sum_i w_i \ln \frac{w_i}{w_{t,i}}$$

Such that the weighted training error of h_t w.r.t. $\boldsymbol{w}^{(t+1)}$ is $\frac{1}{2}$

("diversification" of Boosting mentioned in Ron Meir's talk)

$$w_i(\alpha) = w_i^{(t)} \exp(-\alpha y_i h_t(\boldsymbol{x}_i)) \qquad \alpha = \operatorname*{argmin}_{\alpha'} \sum_i w_i(\alpha')$$





projection onto $H^{(1)}$
hyperplane $H^{(1)}$
$\boldsymbol{w}^{(0)}$
$\boldsymbol{w}^{(\infty)}$
$\boldsymbol{w}^{(1)}$
$\boldsymbol{w}^{(2)}$
projection onto $H^{(2)}$
intersection $H^{(1)} \cap H^{(2)}$
hyperplane $H^{(2)}$

Where do Bregman divergences come from?

- Exponential family of distributions
- Inherent duality

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

primal param. dual param.

$$\begin{array}{ccc} \boldsymbol{w}_t & \xrightarrow{f} & f(\boldsymbol{w}_t) \\ \boldsymbol{w}_{t+1} & \xleftarrow{f^{-1}} & -\eta \nabla L_t(\boldsymbol{w}_t) \end{array}$$

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

- Expectation parameter

$$\boldsymbol{\mu} = \int_{\boldsymbol{x}} \boldsymbol{x} P_G(\boldsymbol{x}|\boldsymbol{\theta}) d\boldsymbol{x} = E_{\boldsymbol{\theta}}(\boldsymbol{x}) = g(\boldsymbol{\theta})$$

  where $g(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} G(\boldsymbol{\theta})$

- Second convex function $F(\boldsymbol{\mu})$ on space $g(\boldsymbol{\Theta})$

$$F(\boldsymbol{\mu}) \;\; = \;\; \boldsymbol{\theta} \cdot \boldsymbol{\mu} - G(\boldsymbol{\theta})$$

- $G(\boldsymbol{\theta})$ and $F(\boldsymbol{\mu})$ are *convex conjugate* functions
- Let $f(\boldsymbol{\mu}) = \nabla_{\boldsymbol{\mu}} F(\boldsymbol{\mu})$
- $f(\boldsymbol{\mu}) = g^{-1}(\boldsymbol{\mu})$

Primal & Dual Parameters

natural paramater		expectation parameter
$\boldsymbol{\theta}$	$\overset{g}{\longrightarrow}$ $\underset{f}{\longleftarrow}$	$\boldsymbol{\mu}$
$G(\boldsymbol{\theta})$		$F(\boldsymbol{\mu})$

- $\boldsymbol{\theta}$ and $\boldsymbol{\mu}$ are dual parameters
- Parameter transformations
  $g(\boldsymbol{\theta}) = \boldsymbol{\mu}$ and $f(\boldsymbol{\mu}) = \boldsymbol{\theta}$ [A,BN]

Gaussian (unit variance)

$$
\begin{aligned}
P(\boldsymbol{x}|\boldsymbol{\theta}) &\sim e^{-\frac{1}{2}(\boldsymbol{\theta}-\boldsymbol{x})^2} \\
&= e^{\boldsymbol{\theta}\cdot\boldsymbol{x}-\frac{1}{2}\boldsymbol{\theta}^2} e^{\frac{1}{2}\boldsymbol{x}^2}
\end{aligned}
$$

Cumulant function: $G(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^2$

Parameter transformations:

$g(\boldsymbol{\theta}) = \boldsymbol{\theta} = \boldsymbol{\mu}$ and $f(\boldsymbol{\mu}) = \boldsymbol{\mu} = \boldsymbol{\theta}$

Dual convex function:
$$
\begin{aligned}
F(\boldsymbol{\mu}) &= \boldsymbol{\theta}\cdot\boldsymbol{\mu} - G(\boldsymbol{\theta}) \\
&= \frac{1}{2}\boldsymbol{\mu}^2
\end{aligned}
$$

Square loss: $L_t(\boldsymbol{\theta}) = \frac{1}{2}(\boldsymbol{\theta}_t - \boldsymbol{x}_t)^2$

Bernoulli

Examples x_t are coin flips in $\{0,1\}$

$$P(x|\mu) = \mu^x (1-\mu)^{1-x}$$

μ is the probability (expectation) of 1
Natural parameter: $\theta = \ln \frac{\mu}{1-\mu}$

$$P(x|\theta) = \exp\left(\theta x - \ln(1+e^\theta)\right)$$

Cumulant function: $G(\theta) = \ln(1+e^\theta)$

Parameter transformations:

$$\mu = g(\theta) = \frac{e^\theta}{1+e^\theta} \text{ and } \theta = f(\mu) = \ln \frac{\mu}{1-\mu}$$

Dual function: $F(\mu) = \mu \ln \mu + (1-\mu)\ln(1-\mu)$

Log loss: $$\begin{aligned} L_t(\theta) &= -x_t\theta + ln(1+e^\theta) \\ &= -x_t \ln \mu - (1-x_t)\ln(1-\mu) \end{aligned}$$

Poisson

Examples x_t are natural numbers in $\{0, 1, \dots\}$

$$P(x|\mu) = \frac{e^{-\mu}\mu^x}{x!}$$

μ is expectation of x

Natural parameter: $\theta = \ln \mu$

$$P(x|\theta) = \exp\left(\theta x - e^\theta\right) \frac{1}{x!}$$

Cumulant function: $G(\theta) = e^\theta$

Parameter transformations:

$$\mu = g(\theta) = e^\theta \text{ and } \theta = f(\mu) = \ln \mu$$

Dual function: $F(\mu) = \mu \ln \mu - \mu$

$$\begin{aligned} \text{Loss:} \quad L_t(\theta) &= -x_t\theta + e^\theta + \ln x_t! \\ &= -x_t \ln \mu + \mu + \ln x_t! \end{aligned}$$

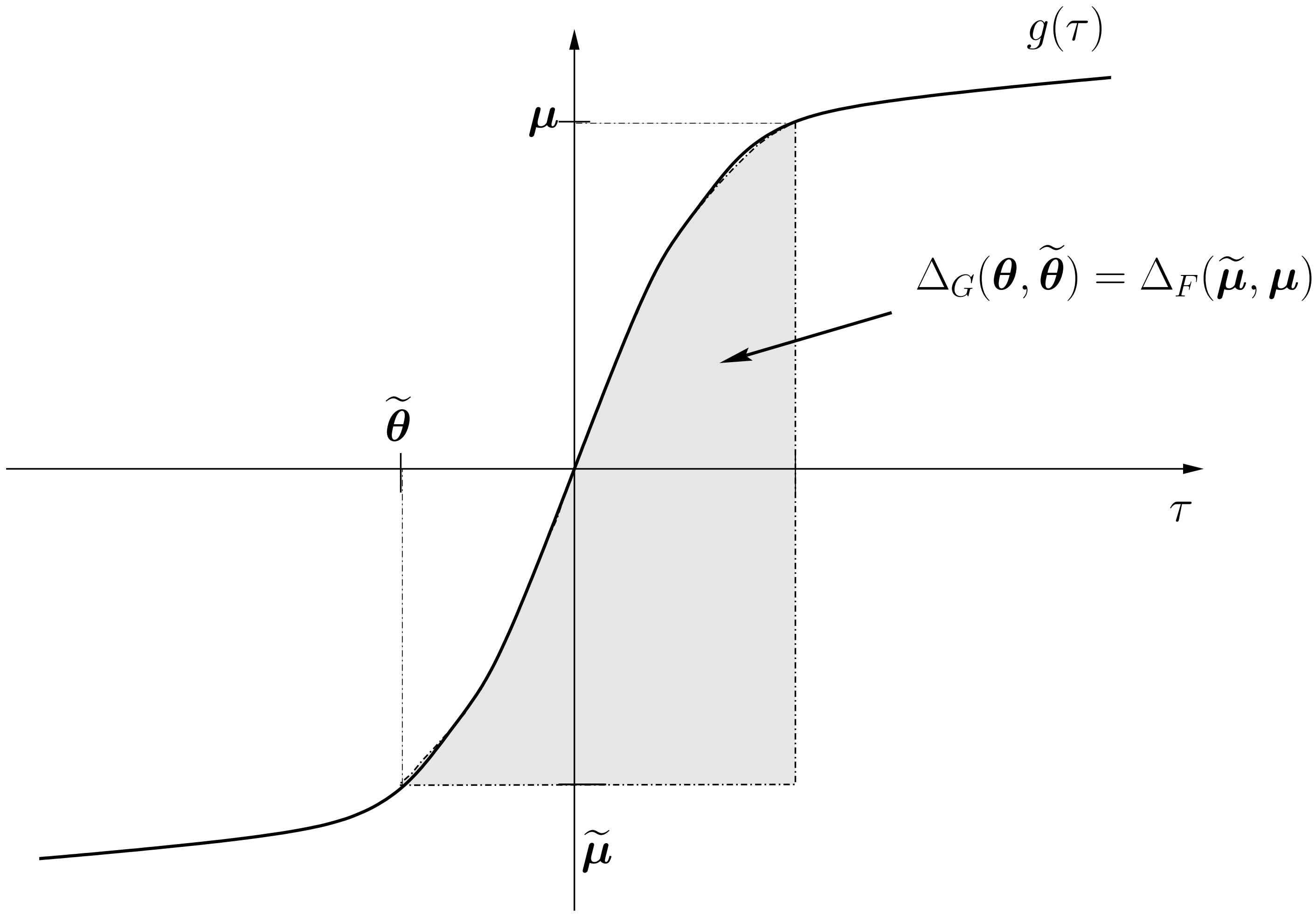
Bregman Div. as Rel. Ent. between Distributions

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

$$
\begin{aligned}
\Delta_G(\widetilde{\boldsymbol{\theta}}, \boldsymbol{\theta}) &= \int_{\boldsymbol{x}} P_G(\boldsymbol{x}|\boldsymbol{\theta}) \ln \frac{P_G(\boldsymbol{x}|\boldsymbol{\theta})}{P_G(\boldsymbol{x}|\widetilde{\boldsymbol{\theta}})} d\boldsymbol{x} \\
&= G(\widetilde{\boldsymbol{\theta}}) - G(\boldsymbol{\theta}) - (\widetilde{\boldsymbol{\theta}} - \boldsymbol{\theta}) \cdot \boldsymbol{\mu} \\
&\overset{F(\boldsymbol{\mu}) = \boldsymbol{\theta}\cdot\boldsymbol{\mu} - G(\boldsymbol{\theta})}{=} F(\boldsymbol{\mu}) - F(\widetilde{\boldsymbol{\mu}}) - (\boldsymbol{\mu} - \widetilde{\boldsymbol{\mu}}) \cdot \widetilde{\boldsymbol{\theta}} \\
&= \Delta_F(\boldsymbol{\mu}, \widetilde{\boldsymbol{\mu}})
\end{aligned}
$$

[A,BN,AW]

Area unchanged When Slide Flipped



$$
\begin{aligned}
\Delta_G(\boldsymbol{\theta}, \widetilde{\boldsymbol{\theta}}) &= G(\boldsymbol{\theta}) - G(\widetilde{\boldsymbol{\theta}}) - (\boldsymbol{\theta} - \widetilde{\boldsymbol{\theta}}) \cdot g(\widetilde{\boldsymbol{\theta}}) \\
&= \int_{\widetilde{\boldsymbol{\theta}}}^{\boldsymbol{\theta}} (g(\boldsymbol{\tau}) - g(\widetilde{\boldsymbol{\theta}})) \cdot d\boldsymbol{\tau} \\
&\overset{\text{flip}}{=} \int_{\boldsymbol{\mu}}^{\widetilde{\boldsymbol{\mu}}} (f(\boldsymbol{\sigma}) - f(\boldsymbol{\mu})) \cdot d\boldsymbol{\sigma} \\
&= F(\widetilde{\boldsymbol{\mu}}) - F(\boldsymbol{\mu}) - (\widetilde{\boldsymbol{\mu}} - \boldsymbol{\mu}) \cdot f(\boldsymbol{\mu}) \\
&= \Delta_F(\widetilde{\boldsymbol{\mu}}, \boldsymbol{\mu})
\end{aligned}
$$

Dual divergence for Bernoulli

$$G(\boldsymbol{\theta}) = \ln(1 + e^{\theta}) \qquad F(\mu) = \mu \ln \mu + (1-\mu)\ln(1-\mu)$$

$$g(\theta) = \tfrac{e^{\theta}}{1+e^{\theta}} = \mu \qquad f(\mu) = \ln \tfrac{\mu}{1-\mu} = \theta$$

$$\Delta_G(\widetilde{\theta}, \theta) = \ln(1 + e^{\widetilde{\theta}}) - \ln(1 + e^{\theta}) - (\widetilde{\theta} - \theta)\frac{e^{\theta}}{1 + e^{\theta}}$$

$$\Delta_F(\mu, \widetilde{\mu}) = \mu \ln \frac{\mu}{\widetilde{\mu}} + (1-\mu)\ln\frac{1-\mu}{1-\widetilde{\mu}}$$

Binary relative entropy

Dual divergence for Poisson

$$G(\boldsymbol{\theta}) = e^{\theta} \qquad F(\mu) = \mu \ln \mu - \mu$$

$$g(\theta) = e^{\theta} = \mu \quad f(\mu) = \ln \mu = \theta$$

$$\Delta_G(\widetilde{\theta}, \theta) = e^{\widetilde{\theta}} - e^{\theta} - (\widetilde{\theta} - \theta)e^{\theta}$$

$$\Delta_F(\mu, \widetilde{\mu}) = \mu \ln \frac{\mu}{\widetilde{\mu}} + \widetilde{\mu} - \mu$$

Unnormalized relative entropy

Dual matching loss for sigmoid transfer func.

$$H(z) = \ln(1+e^z) \qquad K(r) = r\ln r + (1-r)\ln(1-r)$$

$$h(z) = \tfrac{e^z}{1+e^z} = r \qquad k(r) = \ln\tfrac{r}{1-r} = z$$

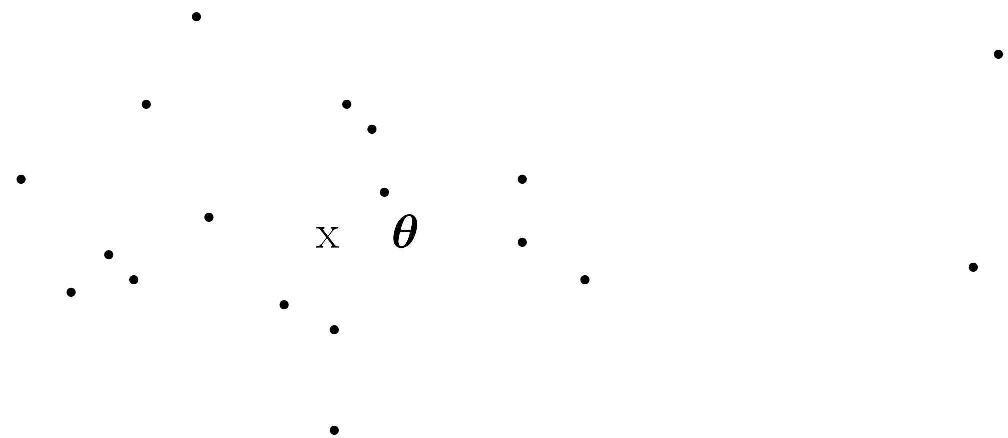
$$K \text{ dual to } H \text{ and } k = h^{-1}$$

$$\begin{aligned}&\Delta_H(\boldsymbol{w}\cdot\boldsymbol{x}, h^{-1}(y))\\ &\quad = \ln(1+e^{\boldsymbol{w}\cdot\boldsymbol{x}}) - y\boldsymbol{w}\cdot\boldsymbol{x} + y\ln y + (1-y)\ln(1-y)\end{aligned}$$

By duality logistic loss is same as entropic loss

$$\begin{aligned}&\Delta_K(y, h(\boldsymbol{w}\cdot\boldsymbol{x}))\\ &\quad = y\ln\frac{y}{h(\boldsymbol{w}\cdot\boldsymbol{x})} + (1-y)\ln\frac{1-y}{1-h(\boldsymbol{w}\cdot\boldsymbol{x})}\end{aligned}$$

Example: Gaussian density estimation



Off-line versus on-line

- Loss on example $\boldsymbol{x}_t$

$$L_t(\boldsymbol{\theta}) = -\ln P(\boldsymbol{x}_t|\boldsymbol{\theta}) \quad = \quad \frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{\theta})^2$$

Derivation of Updates

- Want to bound

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \inf_{\boldsymbol{\theta}} L_{1..T}(\boldsymbol{\theta})$$

- Off-line algorithm has all T examples

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_T\}$$

- Setup for choosing best parameter setting

$$\boldsymbol{\theta}_B = \operatorname*{argmin}_{\boldsymbol{\theta}}(\eta_B^{-1} \underbrace{\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_1)}_{\substack{\text{divergence} \\ \text{to initial}}} + \underbrace{L_{1..T}(\boldsymbol{\theta})}_{\substack{\text{total} \\ \text{loss}}})$$

Here $\eta_B^{-1} > 0$ is a tradeoff parameter

On-line Algorithm [AW]

- In trial t, the first t examples

$$\{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_t\}$$

  have been presented

- Motivation for on-line parameter update:
  do as well as best off-line algorithm up to trial t

- At end of trial t algorithm minimizes

$$\boldsymbol{\theta}_{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \; (\eta_1^{-1} \underbrace{\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_1)}_{\substack{\text{divergence}\\ \text{to initial}}} \; + \; \underbrace{L_{1..t}(\boldsymbol{\theta})}_{\substack{\text{loss}\\ \text{so far}}} \;)$$

  Tradeoff parameter $\eta_1^{-1} \geq 0$

Alternate Motivation of Same On-Line Update

$$\boldsymbol{\theta}_{t+1} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}}\Big(\eta_t^{-1} \underbrace{\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_t)}_{\text{divergence to last}} + \underbrace{L_t(\boldsymbol{\theta})}_{\text{current loss}}\Big)$$

where $\eta_t = \dfrac{1}{\eta_1^{-1} + t - 1}$

Parameter Updates

Off-line: $$\boldsymbol{\mu}_B = \frac{\eta_B^{-1}\boldsymbol{\mu}_1 + \sum_{t=1}^{T} \boldsymbol{x}_t}{\eta_B^{-1} + T}$$

On-Line in trial t: $$\boldsymbol{\mu}_{t+1} = \frac{\eta_1^{-1}\boldsymbol{\mu}_1 + \sum_{q=1}^{t} \boldsymbol{x}_q}{\eta_1^{-1} + t} = \boldsymbol{\mu}_t - \eta_{t+1}(\boldsymbol{\mu}_t - \boldsymbol{x}_t)$$

$$\boldsymbol{\theta}_{t+1} = g^{-1}\left(g(\boldsymbol{\theta}_t) - \eta_{t+1}(\boldsymbol{\mu}_t - \boldsymbol{x}_t)\right)$$

- On-line algorithm has freedom to use a tradeoff parameter η_1^{-1} that could be different from the off-line parameter η_B^{-1}
- Two choices for η_1^{-1}

  Case $\eta_1^{-1} = \eta_B^{-1}$:

  Incremental Off-Line Algorithm

  Case $\eta_1^{-1} = \eta_B^{-1} + 1$:

  Forward Algorithm [V]

Shrinkage Towards Initial

$$\boldsymbol{\mu}_B = \overline{\boldsymbol{x}_T} - \eta_B^{-1}(\eta_B^{-1} + T)^{-1}(\overline{\boldsymbol{x}_T} - \boldsymbol{\mu}_1)$$

where $\overline{\boldsymbol{x}_T} = \frac{\sum_{t=1}^{T} \boldsymbol{x}_t}{T}$

Shrinkage factor $\eta_B^{-1}(\eta_B^{-1} + T)^{-1}$

	Off-line	Forward on-line
Gauss $\boldsymbol{\mu}_1 = 0, \eta_B^{-1} = 0$	$\boldsymbol{\mu}_B = \frac{\sum_{t=1}^{T} \boldsymbol{x}_t}{T}$	$\boldsymbol{\mu}_t = \frac{\sum_{q=1}^{t-1} \boldsymbol{x}_q}{t}$
Bernoulli $\boldsymbol{\mu}_1 = \frac{1}{2}, \eta_B^{-1} = 0$	$\boldsymbol{\mu}_B = \frac{\sum_{t=1}^{T} \boldsymbol{x}_t}{T}$	$\boldsymbol{\mu}_t = \frac{\frac{1}{2}+\sum_{q=1}^{t-1} \boldsymbol{x}_q}{t}$

Key Lemma [AW]

For any example $\boldsymbol{x}_t$ and any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\underbrace{L_t(\boldsymbol{\theta}_t)}_{\substack{\text{loss of}\\ \text{algorithm}}} - \underbrace{L_t(\boldsymbol{\theta})}_{\substack{\text{loss of}\\ \text{comparator } \boldsymbol{\theta}}}$$

$$= \eta_t^{-1} \underbrace{\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_t)}_{\substack{\text{divergence}\\ \text{to last par.}}} - \eta_{t+1}^{-1} \underbrace{\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_{t+1})}_{\substack{\text{divergence}\\ \text{to updated par.}}}$$

$$+ \underbrace{\eta_{t+1}^{-1} \Delta_G(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})}_{\substack{\text{cost of}\\ \text{update}}}$$

Main Theorem

For any sequence of examples and any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$

$$\underbrace{\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t)}_{\text{total loss of algorithm}} - \underbrace{L_{1..T}(\boldsymbol{\theta})}_{\text{total loss of comparator } \boldsymbol{\theta}}$$

$$= \eta_1^{-1} \underbrace{\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_1)}_{\text{divergence to initial par.}} - \eta_{T+1}^{-1} \underbrace{\Delta_G(\boldsymbol{\theta}, \boldsymbol{\theta}_{T+1})}_{\text{divergence to last par.}}$$

$$+ \underbrace{\sum_{t=1}^{T} \eta_{t+1}^{-1} \Delta_G(\boldsymbol{\theta}_t, \boldsymbol{\theta}_{t+1})}_{\text{cost of all updates}}$$

Proven by simply summing the Key Lemma

Bounds for the Forward Algorithm

$$\sum_{t=1}^{T} L_t(\boldsymbol{\theta}_t) - \inf_{\boldsymbol{\theta}} L_{1..T}(\boldsymbol{\theta}) \overset{\text{Gauss}}{=} \sum_{t=1}^{T} \eta_t \, \boldsymbol{x}_t^2/2 - \sum_{t=1}^{T-1} \eta_t \, \boldsymbol{\mu}_{t+1}^2/2 \qquad [AW]$$

$$\leq \frac{X^2}{2} \ln(1 + \frac{T}{\eta_1^{-1} - 1})$$

$$\overset{\text{Bernoulli}}{\leq} \frac{1}{2} \ln(T+1) + 1 \qquad [Fr, XB, AW]$$

$$\overset{\text{lin. regr.}}{\leq} \frac{1}{2} Y^2 n \ln\left(1 + \frac{TX^2}{a}\right) \qquad [V, Fo, AW]$$

$$X^2 = \max_{t=1}^{T} \boldsymbol{x}_t^2, \; Y = \max_{t=1}^{T} y_t, \; \boldsymbol{w}_t = \left(a\boldsymbol{I} + \sum_{q=1}^{t} \boldsymbol{x}_q \boldsymbol{x}_q'\right)^{-1} \sum_{q=1}^{t-1} \boldsymbol{x}_q y_q$$

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

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

Gaussian and Bernoulli are exceptions [TW,Sh]

Gaussian

Forward Alg. $\boldsymbol{\theta}_t = \frac{\sum_{q=1}^{t-1} \boldsymbol{x}_q}{t}$

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]

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

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

For Bernoulli Last-step Minimax slightly better than Forward Alg
(Laplace Estimator)