

Online learning using Bregman Divergences

Yoav Freund

January 29, 2020

Material follows Chapter 11 of “Prediction Learning and Games” and “Adaptive game playing using multiplicative weights” by Freund and Schapire.

Outline

Hedge(η)Algorithm

Bound on total loss

The hedging problem

- ▶ N possible actions
- ▶ At each time step $t = 1, 2, \dots, T$:
 - ▶ Algorithm chooses a distribution \mathbf{p}^t over actions.
 - ▶ Losses $0 \leq \ell_i^t \leq 1$ of all actions $i = 1, \dots, N$ are revealed.
 - ▶ Algorithm suffers **expected** loss $\mathbf{p}^t \cdot \ell_t$
- ▶ **Goal:** minimize total expected loss
- ▶ Here we have stochasticity - but only in **algorithm**, not in **outcome**

Hedge(η)

└ **Hedge**(η)Algorithm

The **Hedge**(η)Algorithm

Consider action i at time t

- ▶ Total loss:

$$L_i^t = \sum_{s=1}^{t-1} \ell_i^s$$

- ▶ Weight:

$$w_i^t = w_i^1 e^{-\eta L_i^t}$$

Note freedom to choose initial weight (w_i^1) $\sum_{i=1}^n w_i^1 = 1$.

- ▶ $\eta > 0$ is the learning rate parameter. Halving: $\eta \rightarrow \infty$

- ▶ Probability:

$$p_i^t = \frac{w_i^t}{\sum_{j=1}^N w_j^t}, \quad \mathbf{p}^t = \frac{\mathbf{w}^t}{\sum_{j=1}^N w_j^t}$$

Hedge(η)

└ Bound on total loss

Bound on the loss of **Hedge**(η)Algorithm

- Theorem (main theorem)

For any sequence of loss vectors ℓ_1, \dots, ℓ_T , and for any $i \in \{1, \dots, N\}$, we have

$$L_{\text{Hedge}(\eta)} \leq \frac{-\ln(w_i^1) + \eta L_i}{1 - e^{-\eta}}.$$

- **Proof:** by combining upper and lower bounds on $\sum_{i=1}^N w_i^{T+1}$

Comparing with the best distribution

- ▶ **Comparison class:** single experts. hindsight.
- ▶ Does not take advantage of multiple good experts.
- ▶ We will get tighter bounds by increasing the comparison class to include all **convex combinations** of the experts.

Hedge(η)

└ Bound on total loss

Recall Single step bound for **Hedge**(η)

The total weight has to decrease if the loss is large

$$\sum_{i=1}^N w_i^{t+1} \leq \left(\sum_{i=1}^N w_i^t \right) (1 - (1 - e^{-\eta}) \mathbf{p}^t \cdot \ell_t)$$

Enlarging the comparison set

- ▶ Bound compares cumulative loss to that of best expert in hindsight.
- ▶ Does not take advantage of multiple good experts.
- ▶ We will get tighter bounds by comparing to the best convex combination of experts.

Hedge(η)

└ Bound on total loss

Comparing with the best distribution

- ▶ Denote by \mathbf{q} an arbitrary distribution over N experts.
 $\mathbf{q} \in \Delta^N$. Distribution = convex combination.
- ▶ Compare loss of algorithm to loss of best convex combination of experts:

$$\sum_{t=1}^T L_A^t \leq +a \min_{\mathbf{q} \in \Delta^N} \sum_{t=1}^T \mathbf{q} \cdot \ell_t + cX$$

- ▶ When comparing to single best expert $X = \log N$
- ▶ **Intuition:** X should be small if best distribution \mathbf{q}^* is close to initial distribution \mathbf{p}^0

Hedge(η)

└ Bound on total loss

Relative Entropy Bound

- ▶ Relative Entropy or KL-Divergence:

$$\text{RE}(\mathbf{q} \parallel \mathbf{p}) \doteq \sum_{i=1}^N \mathbf{q}(i) \ln \frac{\mathbf{q}(i)}{\mathbf{p}(i)}$$

- ▶ For any distribution \mathbf{q} and any learning rate $\eta > 0$ The cumulative loss of **Hedge**(η) is bounded by:

$$\sum_{t=1}^T \mathbf{p}_t \cdot \ell_t \leq a_\eta \sum_{t=1}^T \mathbf{q} \cdot \ell_t + c_\eta \text{RE}(\mathbf{q} \parallel \mathbf{p}_0)$$

- ▶ Where $a_\eta = \frac{\eta}{1-e^{-\eta}}$, $c_\eta = \frac{1}{1-e^{-\eta}}$

Hedge(η)

└ Bound on total loss

Per Iteration bound

- ▶ For any distribution \mathbf{q} and any iteration t **Hedge**(η):

$$c_\eta \ (\text{RE}(\mathbf{q} \parallel \mathbf{p}_t) - \text{RE}(\mathbf{q} \parallel \mathbf{p}_{t+1})) \geq \mathbf{p}_t \cdot \ell_t - a_\eta \mathbf{q} \cdot \ell_t$$

Hedge(η)

└ Bound on total loss

Proof (from RE to ratio)

$$\begin{aligned} & \text{RE}(\mathbf{q} \parallel \mathbf{p}_t) - \text{RE}(\mathbf{q} \parallel \mathbf{p}_{t+1}) \\ &= \sum_{i=1}^N \mathbf{q}(i) \ln \frac{\mathbf{q}(i)}{\mathbf{p}_t(i)} - \sum_{i=1}^N \mathbf{q}(i) \ln \frac{\mathbf{q}(i)}{\mathbf{p}_{t+1}(i)} \\ &= \sum_{i=1}^N \mathbf{q}(i) \ln \frac{\mathbf{p}_{t+1}(i)}{\mathbf{p}_t(i)} \\ &= \sum_{i=1}^N \mathbf{q}(i) \ln \frac{\exp(-\eta \ell_t(i))}{Z_t} \end{aligned}$$

Because

$$\mathbf{p}_{t+1}(i) = \mathbf{p}_t(i) \frac{e^{-\eta \ell_t(i)}}{Z_t}; \quad Z_t = \sum_{i=1}^N \mathbf{p}_t(i) \exp(-\eta \ell_t(i))$$

Hedge(η)

└ Bound on total loss

Proof (from ratio to bound)

$$\begin{aligned} & \sum_{i=1}^N \mathbf{q}(i) \ln \frac{\exp(-\eta \ell_t(i))}{Z_t} \\ &= -\eta \mathbf{q} \cdot \ell_t - \ln Z_t \\ &\geq -\eta \mathbf{q} \cdot \ell_t - \ln \left[\sum_{i=1}^N \mathbf{p}_t(i) (1 - (1 - e^{-\eta}) \ell_t(i)) \right] \\ &\quad \text{because } e^{-\eta x} \leq (1 - (1 - e^{-\eta})x) \text{ for } x \in [0, 1] \\ &\geq (1 - e^{-\eta}) \mathbf{p}_t \cdot \ell_t - \eta \mathbf{q} \cdot \ell_q \\ &\quad \text{because } \ln(1 - x) \leq -x \text{ for } x < 1 \end{aligned}$$

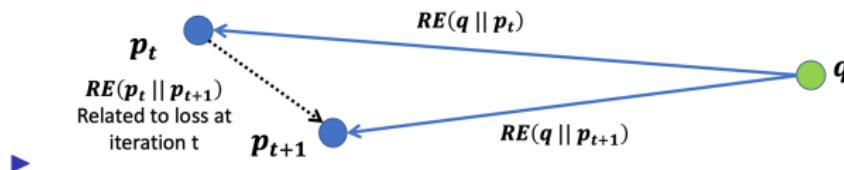
Hedge(η)

└ Bound on total loss

Divergence as Potential

- For any distribution \mathbf{q} and any iteration t **Hedge(η)**:

$$RE(\mathbf{q} \parallel \mathbf{p}_t) - RE(\mathbf{q} \parallel \mathbf{p}_{t+1}) \geq \frac{1}{c_\eta} \mathbf{p}_t \cdot \ell_t - \frac{a_\eta}{c_\eta} \mathbf{q} \cdot \ell_t$$



- If the loss of \mathbf{p}_t is larger than that of \mathbf{q} forces \mathbf{p}_{t+1} do get closer to \mathbf{q} .
- The distance from \mathbf{p}_0 to \mathbf{q}^* bounds the total regret.

Hedge(η)

└ Bound on total loss

online Gradient Descent

- ▶ ω_t - prediction vector
- ▶ x_t - outcome vector
- ▶ $L(\omega, x)$ - loss function (maps prediction,outcome to non-negative loss, differentiable)
- ▶ algorithm suffers $L(\omega_t, x_t)$
- ▶ Algorithm updates $\omega_{t+1} = \omega_t - \eta \nabla_\omega L_t(x_t \omega_t)$
- ▶ Small η = small correction = strong regularization.

Hedge(η)

└ Bound on total loss

Divergence as Regularization

- ▶ Use a divergence term $D(\omega_t, \omega_{t+1})$ to measure the size of the update.
- ▶ Find ω_{t+1} to balance loss reduction and divergence:

$$\operatorname{argmin}_{\omega} = L(\omega, x_t) + \eta D(\omega, \omega_t)$$

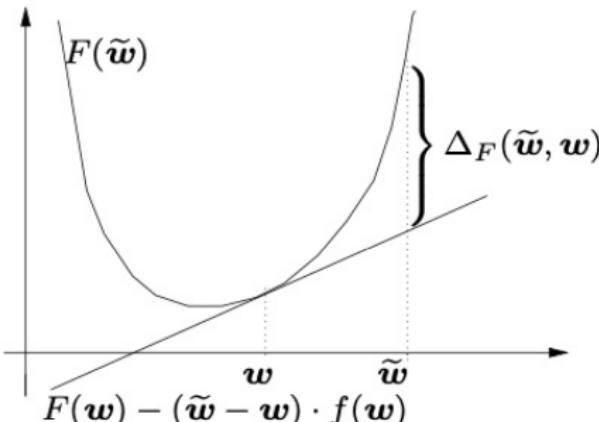
- ▶ If update is small, we can think of it in terms of gradients:

$$\nabla_{\omega}(L(\omega, x_t) + \eta D(\omega, \omega_t))$$

Bregman Divergences [Br, CL, Cs]

For **any** differentiable convex function F

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



Bregman Divergences: Simple Properties

1. $\Delta_F(\tilde{\mathbf{w}}, \mathbf{w})$ is convex in $\tilde{\mathbf{w}}$
2. $\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) \geq 0$
If F convex equality holds iff $\tilde{\mathbf{w}} = \mathbf{w}$
3. Usually not symmetric: $\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) \neq \Delta_F(\mathbf{w}, \tilde{\mathbf{w}})$

4. Linearity (for $a \geq 0$):

$$\Delta_{F+aH}(\tilde{\mathbf{w}}, \mathbf{w}) = \Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) + a \Delta_H(\tilde{\mathbf{w}}, \mathbf{w})$$

5. Unaffected by linear terms ($a \in \mathbf{R}$, $\mathbf{b} \in \mathbf{R}^n$):

$$\Delta_{H+a\tilde{\mathbf{w}}+\mathbf{b}}(\tilde{\mathbf{w}}, \mathbf{w}) = \Delta_H(\tilde{\mathbf{w}}, \mathbf{w})$$

Bregman Divergences: more properties

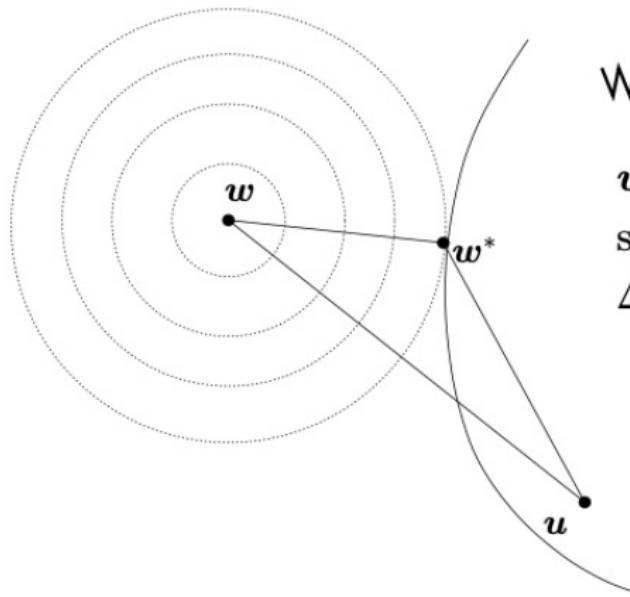
$$6. \nabla_{\tilde{\boldsymbol{w}}} \Delta_F(\tilde{\boldsymbol{w}}, \boldsymbol{w})$$

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

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

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

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



\mathcal{W}

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

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

Theorem:

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

Examples

Squared Euclidean Distance

$$F(\mathbf{w}) = \|\mathbf{w}\|_2^2 / 2$$

$$f(\mathbf{w}) = \mathbf{w}$$

$$\begin{aligned}\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) &= \|\tilde{\mathbf{w}}\|_2^2 / 2 - \|\mathbf{w}\|_2^2 / 2 - (\tilde{\mathbf{w}} - \mathbf{w}) \cdot \mathbf{w} \\ &= \|\tilde{\mathbf{w}} - \mathbf{w}\|_2^2 / 2\end{aligned}$$

(Unnormalized) Relative Entropy

$$F(\mathbf{w}) = \sum_i (w_i \ln w_i - w_i)$$

$$f(\mathbf{w}) = \ln \mathbf{w}$$

$$\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) = \sum_i \left(\widetilde{w_i} \ln \frac{\widetilde{w_i}}{w_i} + w_i - \widetilde{w_i} \right)$$

Examples-2 [GLS,GL]

p-norm Algs (q is dual to p : $\frac{1}{p} + \frac{1}{q} = 1$)

$$F(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|_q^2$$

$$f(\mathbf{w}) = \nabla \frac{1}{2} \|\mathbf{w}\|_q^2$$

$$\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) = \frac{1}{2} \|\tilde{\mathbf{w}}\|_q^2 + \frac{1}{2} \|\mathbf{w}\|_q^2 - \tilde{\mathbf{w}} \cdot f(\mathbf{w})$$

When $p = q = 2$ this reduces to squared Euclidean distance (Widrow-Hoff).

General Motivation of Updates [KW]

Trade-off between two term:

$$\mathbf{w}_{t+1} = \operatorname{argmin}_{\mathbf{w}} \left(\underbrace{\Delta_F(\mathbf{w}, \mathbf{w}_t)}_{\text{weight domain}} + \eta_t \underbrace{L_t(\mathbf{w})}_{\text{label domain}} \right)$$

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

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

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

iff

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

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

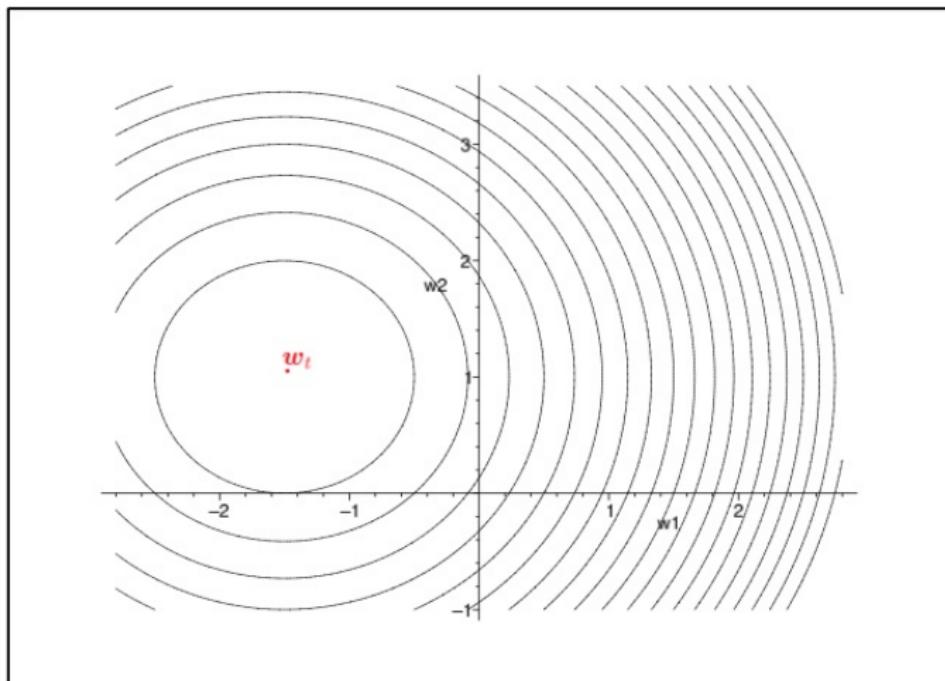
Divergence: Euclidean Distance Squared

$$\Delta_F(\mathbf{w}, \mathbf{w}_t) = \|\mathbf{w} - \mathbf{w}_t\|_2^2 / 2$$

$$\mathbf{w}_t = (-3/2, 1)$$

$$\mathbf{x}_t = (1, -0.5)$$

$$y_t = 1$$



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

Loss & divergence are dependent

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

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

Yield bounds of the form

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

a, b constants, $a > 1$.

Regret bounds ($a = 1$):

time changing η , subtler analysis

[AG]

How to prove relative loss bounds?

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

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

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

convexity

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

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

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

First step: Teleskopung

Summing over t

[WJ,KW]

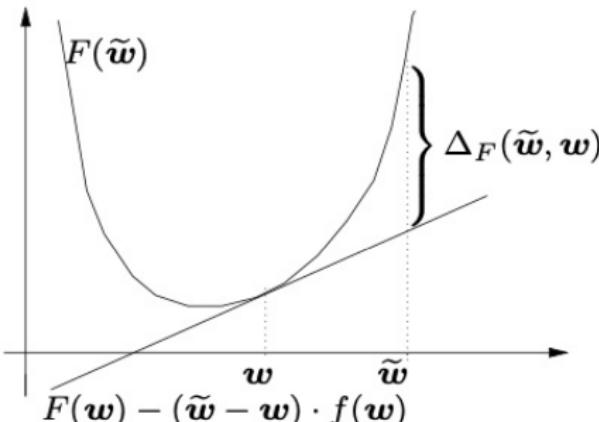
$$\begin{aligned} \sum_t L_t(\mathbf{w}_t) &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \sum_t \left(\Delta_F(\mathbf{u}, \mathbf{w}_t) - \underbrace{\Delta_F(\mathbf{u}, \mathbf{w}_{t+1})}_{\geq 0} \right. \\ &\quad \left. + \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \right) \\ &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \left(\Delta_F(\mathbf{u}, \mathbf{w}_1) - \underbrace{\Delta_F(\mathbf{u}, \mathbf{w}_{T+1})}_{\geq 0} \right) \\ &\quad + \frac{1}{\eta} \sum_t \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \\ &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \Delta_F(\mathbf{u}, \mathbf{w}_1) + \frac{1}{\eta} \sum_t \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \end{aligned}$$

Any convex loss and any Bregman divergence!

Bregman Divergences [Br, CL, Cs]

For **any** differentiable convex function F

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



First step: Teleskopung

Summing over t

[WJ,KW]

$$\begin{aligned} \sum_t L_t(\mathbf{w}_t) &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \sum_t \left(\Delta_F(\mathbf{u}, \mathbf{w}_t) - \underbrace{\Delta_F(\mathbf{u}, \mathbf{w}_{t+1})}_{\geq 0} \right. \\ &\quad \left. + \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \right) \\ &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \left(\Delta_F(\mathbf{u}, \mathbf{w}_1) - \underbrace{\Delta_F(\mathbf{u}, \mathbf{w}_{T+1})}_{\geq 0} \right) \\ &\quad + \frac{1}{\eta} \sum_t \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \\ &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \Delta_F(\mathbf{u}, \mathbf{w}_1) + \frac{1}{\eta} \sum_t \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \end{aligned}$$

Any convex loss and any Bregman divergence!

How to prove relative loss bounds?

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

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

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

convexity

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

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

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

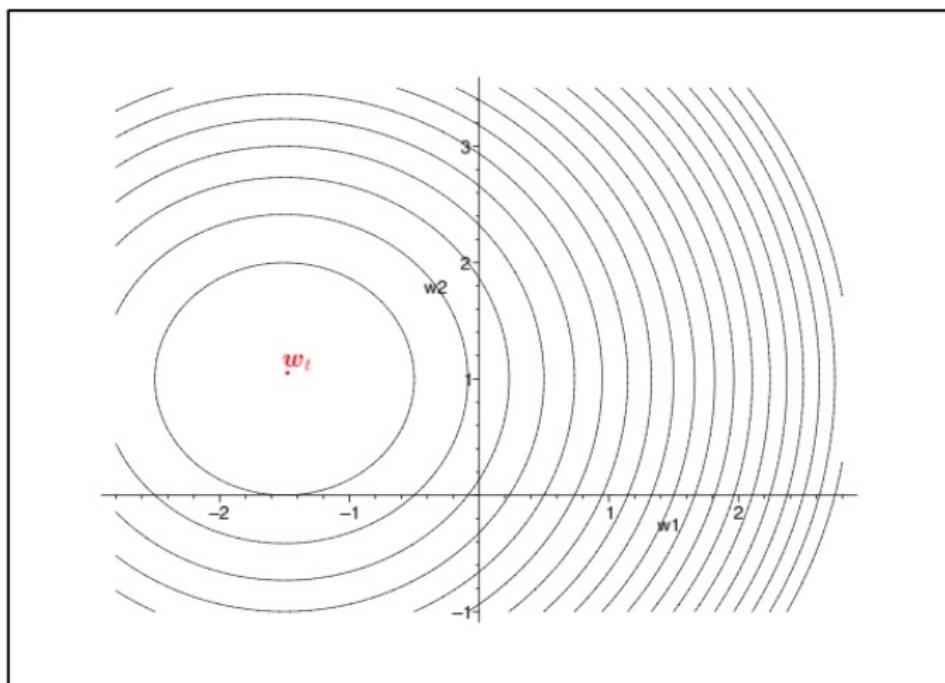
Divergence: Euclidean Distance Squared

$$\Delta_F(\mathbf{w}, \mathbf{w}_t) = \|\mathbf{w} - \mathbf{w}_t\|_2^2 / 2$$

$$\mathbf{w}_t = (-3/2, 1)$$

$$\mathbf{x}_t = (1, -0.5)$$

$$y_t = 1$$



General Motivation of Updates [KW]

Trade-off between two term:

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

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

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

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

iff

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

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

Examples-2 [GLS,GL]

p-norm Algs (q is dual to p : $\frac{1}{p} + \frac{1}{q} = 1$)

$$F(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|_q^2$$

$$f(\mathbf{w}) = \nabla \frac{1}{2} \|\mathbf{w}\|_q^2$$

$$\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) = \frac{1}{2} \|\tilde{\mathbf{w}}\|_q^2 + \frac{1}{2} \|\mathbf{w}\|_q^2 - \tilde{\mathbf{w}} \cdot f(\mathbf{w})$$

When $p = q = 2$ this reduces to squared Euclidean distance (Widrow-Hoff).

Examples

Squared Euclidean Distance

$$F(\mathbf{w}) = \|\mathbf{w}\|_2^2 / 2$$

$$f(\mathbf{w}) = \mathbf{w}$$

$$\begin{aligned}\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) &= \|\tilde{\mathbf{w}}\|_2^2 / 2 - \|\mathbf{w}\|_2^2 / 2 - (\tilde{\mathbf{w}} - \mathbf{w}) \cdot \mathbf{w} \\ &= \|\tilde{\mathbf{w}} - \mathbf{w}\|_2^2 / 2\end{aligned}$$

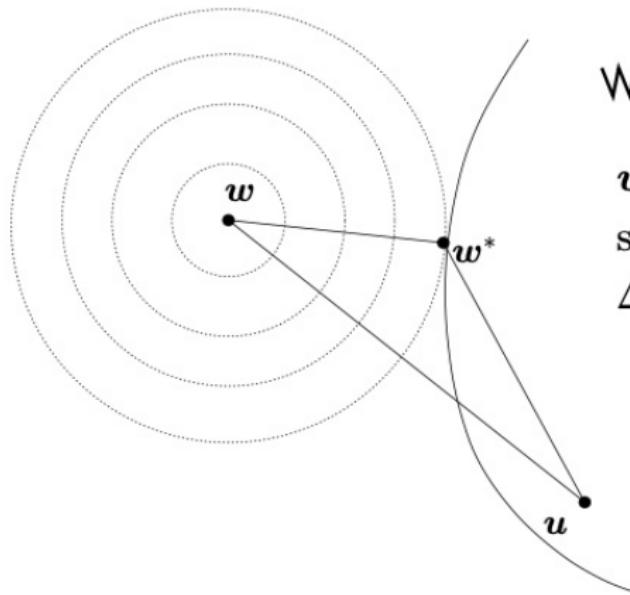
(Unnormalized) Relative Entropy

$$F(\mathbf{w}) = \sum_i (w_i \ln w_i - w_i)$$

$$f(\mathbf{w}) = \ln \mathbf{w}$$

$$\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) = \sum_i \left(\widetilde{w_i} \ln \frac{\widetilde{w_i}}{w_i} + w_i - \widetilde{w_i} \right)$$

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



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

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

Theorem:

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

Bregman Divergences: more properties

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

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

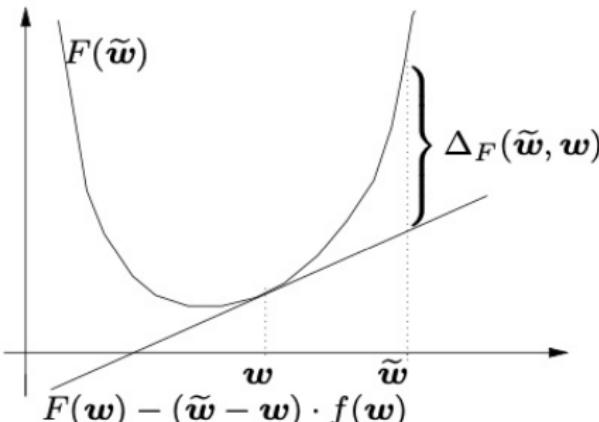
$$7. \Delta_F(\boldsymbol{w}_1, \boldsymbol{w}_2) + \Delta_F(\boldsymbol{w}_2, \boldsymbol{w}_3)$$

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

Bregman Divergences [Br, CL, Cs]

For **any** differentiable convex function F

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



Bregman Divergences: Simple Properties

1. $\Delta_F(\tilde{\mathbf{w}}, \mathbf{w})$ is convex in $\tilde{\mathbf{w}}$
2. $\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) \geq 0$
If F convex equality holds iff $\tilde{\mathbf{w}} = \mathbf{w}$
3. Usually not symmetric: $\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) \neq \Delta_F(\mathbf{w}, \tilde{\mathbf{w}})$

4. Linearity (for $a \geq 0$):

$$\Delta_{F+aH}(\tilde{\mathbf{w}}, \mathbf{w}) = \Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) + a \Delta_H(\tilde{\mathbf{w}}, \mathbf{w})$$

5. Unaffected by linear terms ($a \in \mathbf{R}$, $\mathbf{b} \in \mathbf{R}^n$):

$$\Delta_{H+a\tilde{\mathbf{w}}+\mathbf{b}}(\tilde{\mathbf{w}}, \mathbf{w}) = \Delta_H(\tilde{\mathbf{w}}, \mathbf{w})$$

Bregman Divergences: more properties

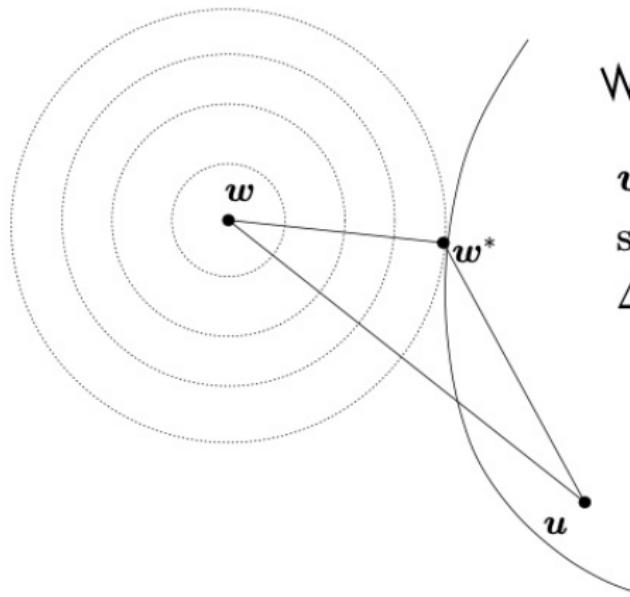
$$6. \nabla_{\tilde{\boldsymbol{w}}} \Delta_F(\tilde{\boldsymbol{w}}, \boldsymbol{w})$$

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

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

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

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



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

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

Theorem:

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

Examples

Squared Euclidean Distance

$$F(\mathbf{w}) = \|\mathbf{w}\|_2^2 / 2$$

$$f(\mathbf{w}) = \mathbf{w}$$

$$\begin{aligned}\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) &= \|\tilde{\mathbf{w}}\|_2^2 / 2 - \|\mathbf{w}\|_2^2 / 2 - (\tilde{\mathbf{w}} - \mathbf{w}) \cdot \mathbf{w} \\ &= \|\tilde{\mathbf{w}} - \mathbf{w}\|_2^2 / 2\end{aligned}$$

(Unnormalized) Relative Entropy

$$F(\mathbf{w}) = \sum_i (w_i \ln w_i - w_i)$$

$$f(\mathbf{w}) = \ln \mathbf{w}$$

$$\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) = \sum_i \left(\widetilde{w_i} \ln \frac{\widetilde{w_i}}{w_i} + w_i - \widetilde{w_i} \right)$$

Examples-2 [GLS,GL]

p-norm Algs (q is dual to p : $\frac{1}{p} + \frac{1}{q} = 1$)

$$F(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|_q^2$$

$$f(\mathbf{w}) = \nabla \frac{1}{2} \|\mathbf{w}\|_q^2$$

$$\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) = \frac{1}{2} \|\tilde{\mathbf{w}}\|_q^2 + \frac{1}{2} \|\mathbf{w}\|_q^2 - \tilde{\mathbf{w}} \cdot f(\mathbf{w})$$

When $p = q = 2$ this reduces to squared Euclidean distance (Widrow-Hoff).

General Motivation of Updates [KW]

Trade-off between two term:

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

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

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

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

iff

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

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

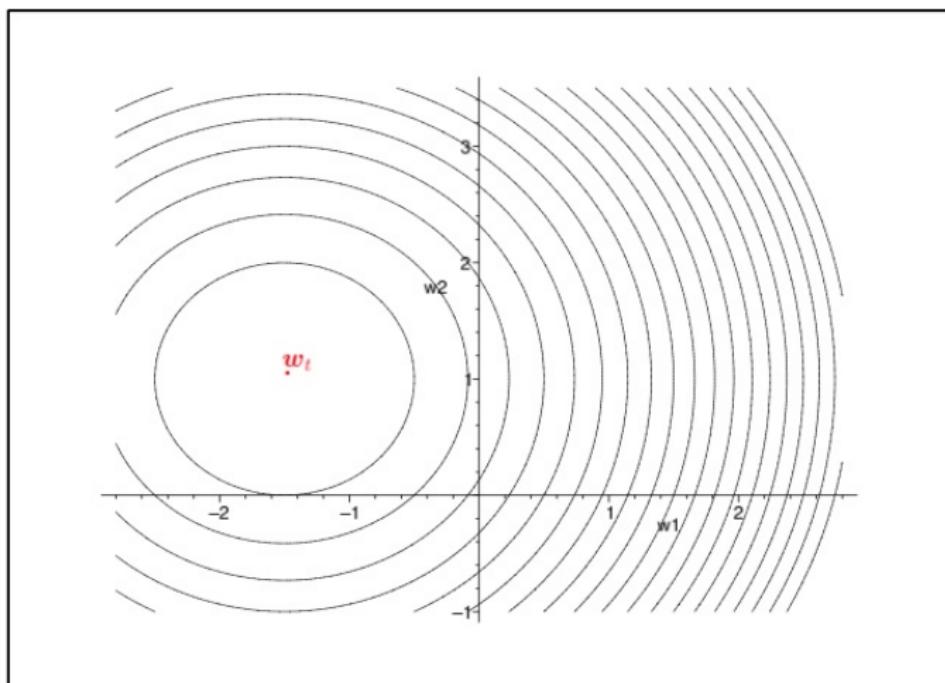
Divergence: Euclidean Distance Squared

$$\Delta_F(\mathbf{w}, \mathbf{w}_t) = \|\mathbf{w} - \mathbf{w}_t\|_2^2 / 2$$

$$\mathbf{w}_t = (-3/2, 1)$$

$$\mathbf{x}_t = (1, -0.5)$$

$$y_t = 1$$



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

Loss & divergence are dependent

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

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

Yield bounds of the form

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

a, b constants, $a > 1$.

Regret bounds ($a = 1$):

time changing η , subtler analysis

[AG]

How to prove relative loss bounds?

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

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

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

convexity

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

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

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

First step: Teleskopung

Summing over t

[WJ,KW]

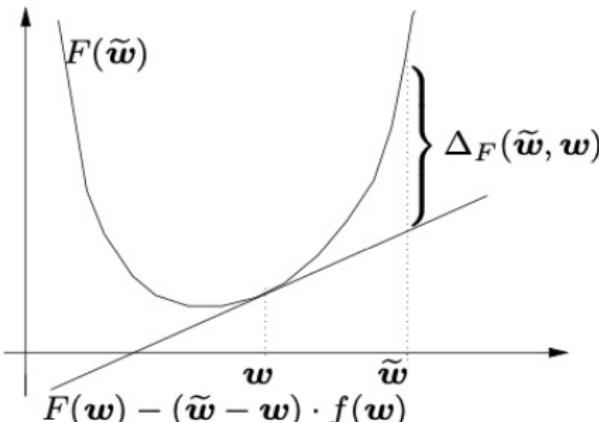
$$\begin{aligned} \sum_t L_t(\mathbf{w}_t) &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \sum_t \left(\Delta_F(\mathbf{u}, \mathbf{w}_t) - \underbrace{\Delta_F(\mathbf{u}, \mathbf{w}_{t+1})}_{\geq 0} \right. \\ &\quad \left. + \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \right) \\ &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \left(\Delta_F(\mathbf{u}, \mathbf{w}_1) - \underbrace{\Delta_F(\mathbf{u}, \mathbf{w}_{T+1})}_{\geq 0} \right) \\ &\quad + \frac{1}{\eta} \sum_t \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \\ &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \Delta_F(\mathbf{u}, \mathbf{w}_1) + \frac{1}{\eta} \sum_t \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \end{aligned}$$

Any convex loss and any Bregman divergence!

Bregman Divergences [Br, CL, Cs]

For **any** differentiable convex function F

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



First step: Teleskopung

Summing over t

[WJ,KW]

$$\begin{aligned} \sum_t L_t(\mathbf{w}_t) &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \sum_t \left(\Delta_F(\mathbf{u}, \mathbf{w}_t) - \underbrace{\Delta_F(\mathbf{u}, \mathbf{w}_{t+1})}_{\geq 0} \right. \\ &\quad \left. + \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \right) \\ &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \left(\Delta_F(\mathbf{u}, \mathbf{w}_1) - \underbrace{\Delta_F(\mathbf{u}, \mathbf{w}_{T+1})}_{\geq 0} \right) \\ &\quad + \frac{1}{\eta} \sum_t \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \\ &\leq \sum_t L_t(\mathbf{u}) + \frac{1}{\eta} \Delta_F(\mathbf{u}, \mathbf{w}_1) + \frac{1}{\eta} \sum_t \Delta_F(\mathbf{w}_t, \mathbf{w}_{t+1}) \end{aligned}$$

Any convex loss and any Bregman divergence!

How to prove relative loss bounds?

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

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

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

convexity

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

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

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

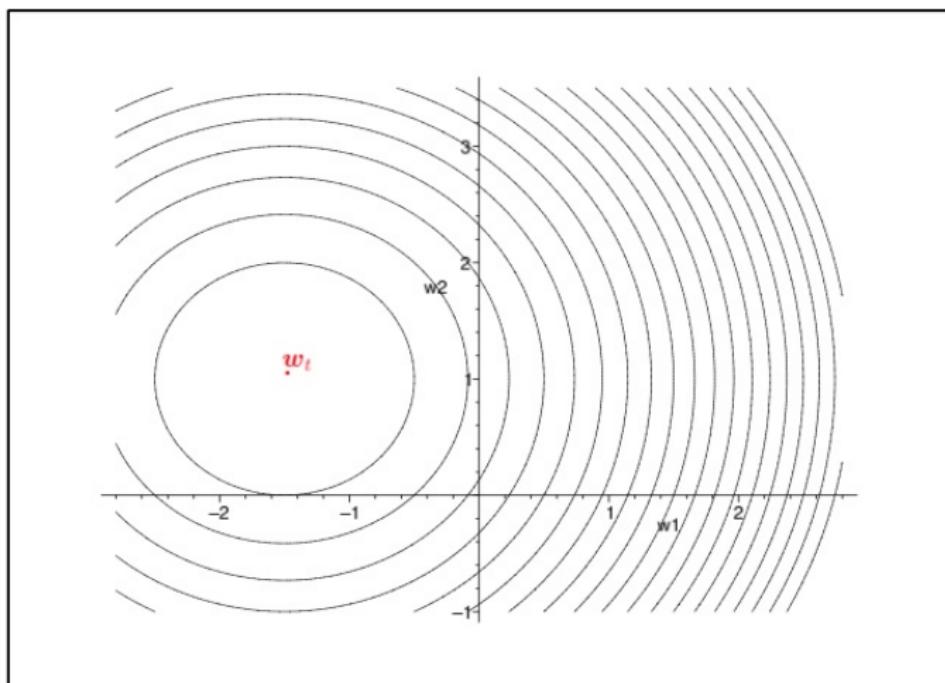
Divergence: Euclidean Distance Squared

$$\Delta_F(\mathbf{w}, \mathbf{w}_t) = \|\mathbf{w} - \mathbf{w}_t\|_2^2 / 2$$

$$\mathbf{w}_t = (-3/2, 1)$$

$$\mathbf{x}_t = (1, -0.5)$$

$$y_t = 1$$



General Motivation of Updates [KW]

Trade-off between two term:

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

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

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

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

iff

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

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

Examples-2 [GLS,GL]

p-norm Algs (q is dual to p : $\frac{1}{p} + \frac{1}{q} = 1$)

$$F(\mathbf{w}) = \frac{1}{2} \|\mathbf{w}\|_q^2$$

$$f(\mathbf{w}) = \nabla \frac{1}{2} \|\mathbf{w}\|_q^2$$

$$\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) = \frac{1}{2} \|\tilde{\mathbf{w}}\|_q^2 + \frac{1}{2} \|\mathbf{w}\|_q^2 - \tilde{\mathbf{w}} \cdot f(\mathbf{w})$$

When $p = q = 2$ this reduces to squared Euclidean distance (Widrow-Hoff).

Examples

Squared Euclidean Distance

$$F(\mathbf{w}) = \|\mathbf{w}\|_2^2 / 2$$

$$f(\mathbf{w}) = \mathbf{w}$$

$$\begin{aligned}\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) &= \|\tilde{\mathbf{w}}\|_2^2 / 2 - \|\mathbf{w}\|_2^2 / 2 - (\tilde{\mathbf{w}} - \mathbf{w}) \cdot \mathbf{w} \\ &= \|\tilde{\mathbf{w}} - \mathbf{w}\|_2^2 / 2\end{aligned}$$

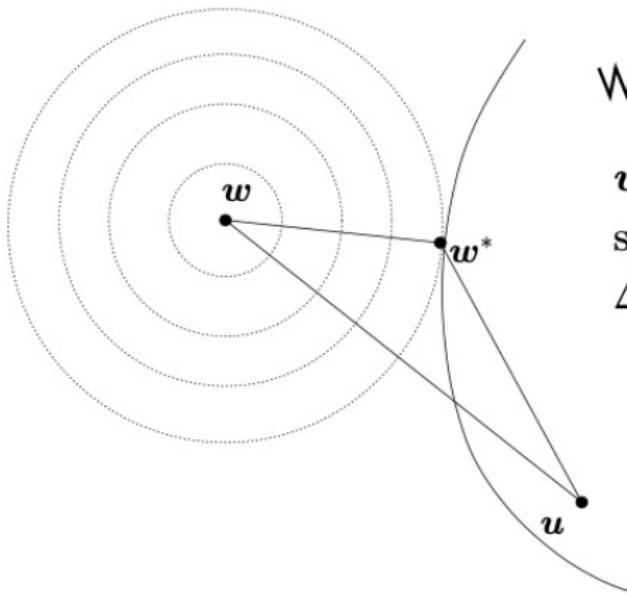
(Unnormalized) Relative Entropy

$$F(\mathbf{w}) = \sum_i (w_i \ln w_i - w_i)$$

$$f(\mathbf{w}) = \ln \mathbf{w}$$

$$\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) = \sum_i \left(\widetilde{w_i} \ln \frac{\widetilde{w_i}}{w_i} + w_i - \widetilde{w_i} \right)$$

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



\mathcal{W}

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

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

Theorem:

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

Bregman Divergences: more properties

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

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

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

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

Bregman Divergences: Simple Properties

1. $\Delta_F(\tilde{\mathbf{w}}, \mathbf{w})$ is convex in $\tilde{\mathbf{w}}$
2. $\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) \geq 0$
If F convex equality holds iff $\tilde{\mathbf{w}} = \mathbf{w}$
3. Usually not symmetric: $\Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) \neq \Delta_F(\mathbf{w}, \tilde{\mathbf{w}})$

4. Linearity (for $a \geq 0$):

$$\Delta_{F+aH}(\tilde{\mathbf{w}}, \mathbf{w}) = \Delta_F(\tilde{\mathbf{w}}, \mathbf{w}) + a \Delta_H(\tilde{\mathbf{w}}, \mathbf{w})$$

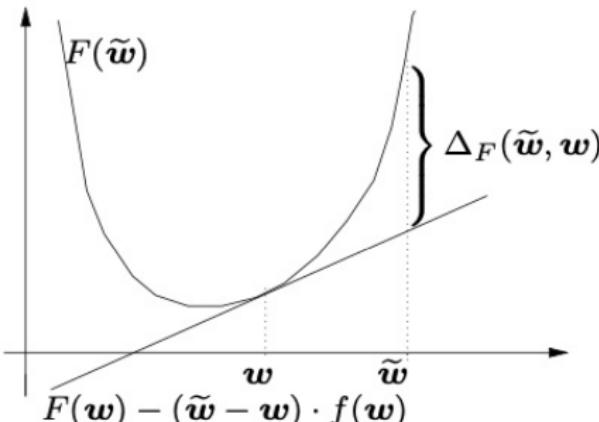
5. Unaffected by linear terms ($a \in \mathbf{R}$, $\mathbf{b} \in \mathbf{R}^n$):

$$\Delta_{H+a\tilde{\mathbf{w}}+\mathbf{b}}(\tilde{\mathbf{w}}, \mathbf{w}) = \Delta_H(\tilde{\mathbf{w}}, \mathbf{w})$$

Bregman Divergences [Br, CL, Cs]

For **any** differentiable convex function F

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



Hedge(η)

└ Bound on total loss

Divergence as Regularization

- ▶ Use a divergence term $D(\omega_t, \omega_{t+1})$ to measure the size of the update.
- ▶ Find ω_{t+1} to balance loss reduction and divergence:

$$\operatorname{argmin}_{\omega} = L(\omega, x_t) + \eta D(\omega, \omega_t)$$

- ▶ If update is small, we can think of it in terms of gradients:

$$\nabla_{\omega}(L(\omega, x_t) + \eta D(\omega, \omega_t))$$