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

Outline

The hedging problem

- N possible actions

The hedging problem

- N possible actions
- At each time step $t = 1, 2, \ldots, T$:

The hedging problem

- N possible actions
- At each time step $t = 1, 2, \ldots, T$:
  - Algorithm chooses a distribution $\mathbf{p}^t$ over actions.

The hedging problem

- N possible actions
- At each time step $t = 1, 2, \ldots, T$:
  - Algorithm chooses a distribution $\mathbf{p}^t$ over actions.
  - Losses $0 \leq \ell_i^t \leq 1$ of all actions $i = 1, \ldots, N$ are revealed.

The hedging problem

- N possible actions
- At each time step $t = 1, 2, \ldots, T$:
  - Algorithm chooses a distribution $\mathbf{p}^t$ over actions.
  - Losses $0 \leq \ell_i^t \leq 1$ of all actions $i = 1, \ldots, N$ are revealed.
  - Algorithm suffers **expected** loss $\mathbf{p}^t \cdot \ell_t$

The hedging problem

- N possible actions
- At each time step $t = 1, 2, \ldots, T$:
  - Algorithm chooses a distribution $\mathbf{p}^t$ over actions.
  - Losses $0 \leq \ell_i^t \leq 1$ of all actions $i = 1, \ldots, N$ are revealed.
  - Algorithm suffers **expected** loss $\mathbf{p}^t \cdot \ell_t$
- **Goal:** minimize total expected loss

The hedging problem

- N possible actions
- At each time step $t = 1, 2, \ldots, T$:
  - Algorithm chooses a distribution $\mathbf{p}^t$ over actions.
  - Losses $0 \leq \ell_i^t \leq 1$ of all actions $i = 1, \ldots, N$ 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**

The **Hedge**(η)Algorithm

Consider action i at time t

- Total loss:

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

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) $\sum_{i=1}^{n} w_i^1 = 1$.

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: $\eta \to \infty$

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: $\eta \to \infty$
- Probability:

$$p_i^t = \frac{w_i^t}{\sum_{j=1}^{N} w_i^t},$$

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: $\eta \to \infty$
- Probability:

$$p_i^t = \frac{w_i^t}{\sum_{j=1}^{N} w_i^t},$$

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: $\eta \to \infty$
- Probability:

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

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

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

- Theorem (main theorem)

  For any sequence of loss vectors $\ell_1, \ldots, \ell_T$*, and for any* $i \in \{1, \ldots, N\}$*, we have*

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

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

- Theorem (main theorem)

  For any sequence of loss vectors $\ell_1, \ldots, \ell_T$*, and for any* $i \in \{1, \ldots, N\}$*, we have*

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

  - 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. hindsite.

Comparing with the best distribution

- **Comparison class:** single experts. hindsite.
- Does not take advantage of multiple good experts.

Comparing with the best distribution

- **Comparison class:** single experts. hindsite.
- 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.

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

The total weight has to decrease if the loss is large

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

Enlarging the comparison set

- Bound compares cumulative loss to that of best expert in hindsite.

Enlarging the comparison set

- Bound compares cumulative loss to that of best expert in hindsite.
- Does not take advantage of multiple good experts.

Enlarging the comparison set

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

Comparing with the best distribution

- Denote by $\mathbf{q}$ an arbitrary distribution over N experts. $\mathbf{q} \in \Delta^N$. Distribution = convex combination.

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:

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

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 $X = \log N$

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 $\mathbf{p}^0$

Relative Entropy Bound

- Relative Entropy or KL-Divergence:

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

Relative Entropy Bound

- Relative Entropy or KL-Divergence:

$$\mathrm{RE}\,(\mathbf{q} \;\|\; \mathbf{p}) \doteq \sum_{i=1}^{N} \mathbf{q}(i) \ln \frac{\mathbf{q}(i)}{\mathbf{p}(i)}$$

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

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

Relative Entropy Bound

- Relative Entropy or KL-Divergence:

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

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

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

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

Per Iteration bound

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

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

Proof (from RE to ratio)

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

Because

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

Proof (from ratio to bound)

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

Divergence as Potential

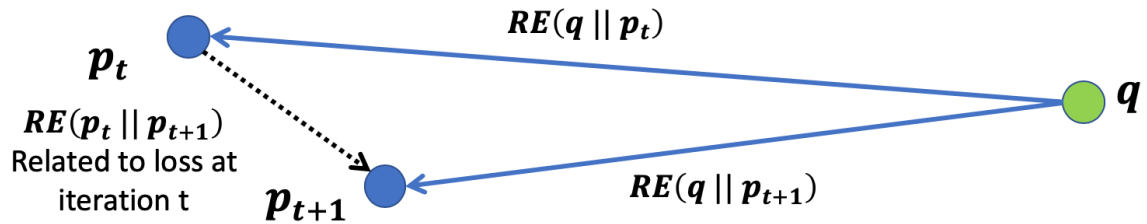
- For any distribution **q** and any iteration t **Hedge**(η):

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

Divergence as Potential

- For any distribution **q** and any iteration t **Hedge**(η):

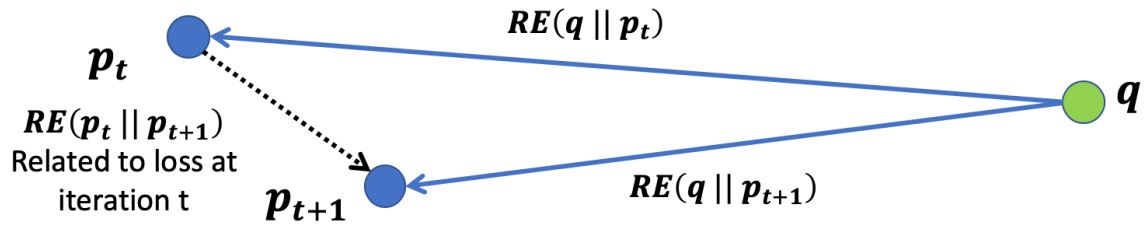
$$\mathrm{RE}(\mathbf{q} \parallel \mathbf{p}_t) - \mathrm{RE}(\mathbf{q} \parallel \mathbf{p}_{t+1}) \geq \frac{1}{c_\eta}\mathbf{p}_t \cdot \ell_t - \frac{a_\eta}{c_\eta}\,\mathbf{q} \cdot \ell_t$$



Divergence as Potential

- For any distribution **q** and any iteration t **Hedge**(η):

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

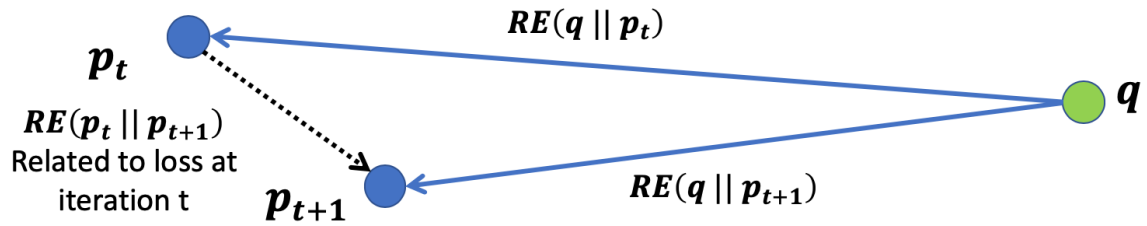


- If the loss of $\mathbf{p}_t$ is larger than that of **q** forces $\mathbf{p}_{t+1}$ do get closer to **q**.

Divergence as Potential

- For any distribution **q** and any iteration t **Hedge**(η):

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



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

online Gradient Descent

- ω_t - prediction vector

online Gradient Descent

- ω_t - prediction vector
- x_t - outcome vector

online Gradient Descent

- ω_t - prediction vector
- x_t - outcome vector
- $L(\omega, x)$ - loss function (maps prediction,outcome to non-negative loss, differentiable)

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 $L(\omega_t, x_t)$

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 $\omega_{t+1} = \omega_t - \eta \bigtriangledown_\omega L_t(x_t \omega_t)$

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 $\omega_{t+1} = \omega_t - \eta \bigtriangledown_\omega L_t(x_t \omega_t)$
- Small η = small correction = strong regularization.

Divergence as Regularization

- Use a divergence term $D(\omega_t, \omega_{t+1})$ to measure the size of the update.

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:

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

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:

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

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

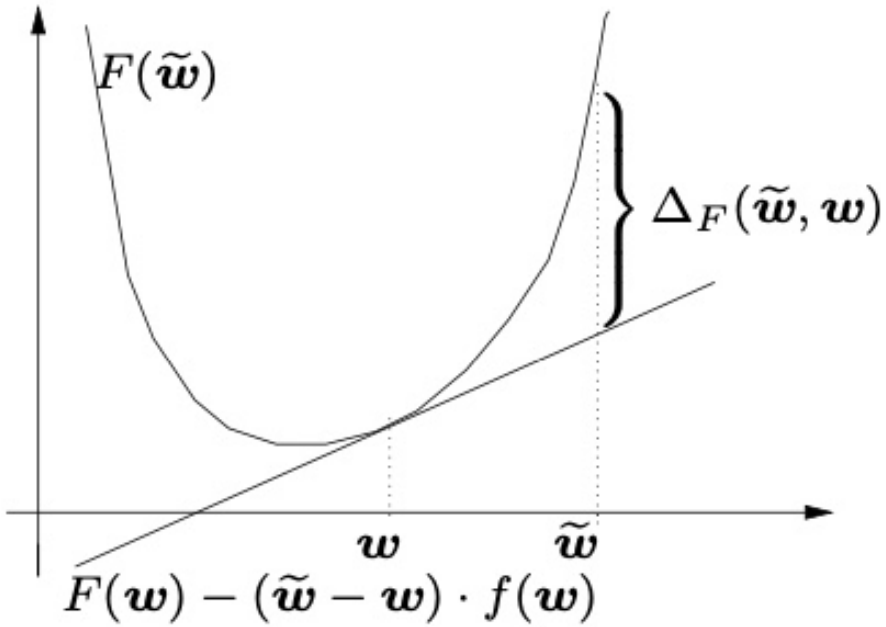
$$\bigtriangledown_\omega(L(\omega, x_t) + \eta D(\omega, \omega_t))$$

Bregman Divergences [Br,CL,Cs]

For any differentiable convex function F

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

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



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

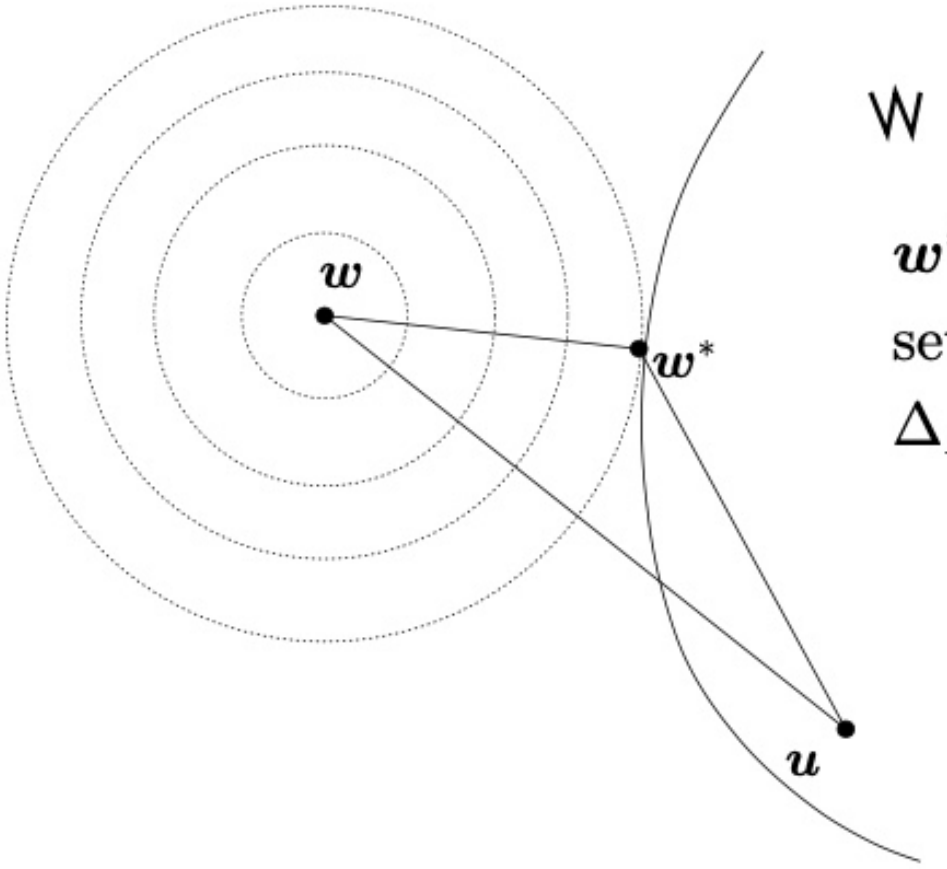
6. $\nabla_{\widetilde{\boldsymbol{w}}} \Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$

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

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

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

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



$\mathcal{W}$

$\boldsymbol{w}^*$ 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

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

Examples-2 [GLS,GL]

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

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

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

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

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

General Motivation of Updates [KW]

Trade-off between two term:

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

$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)$ 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

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

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

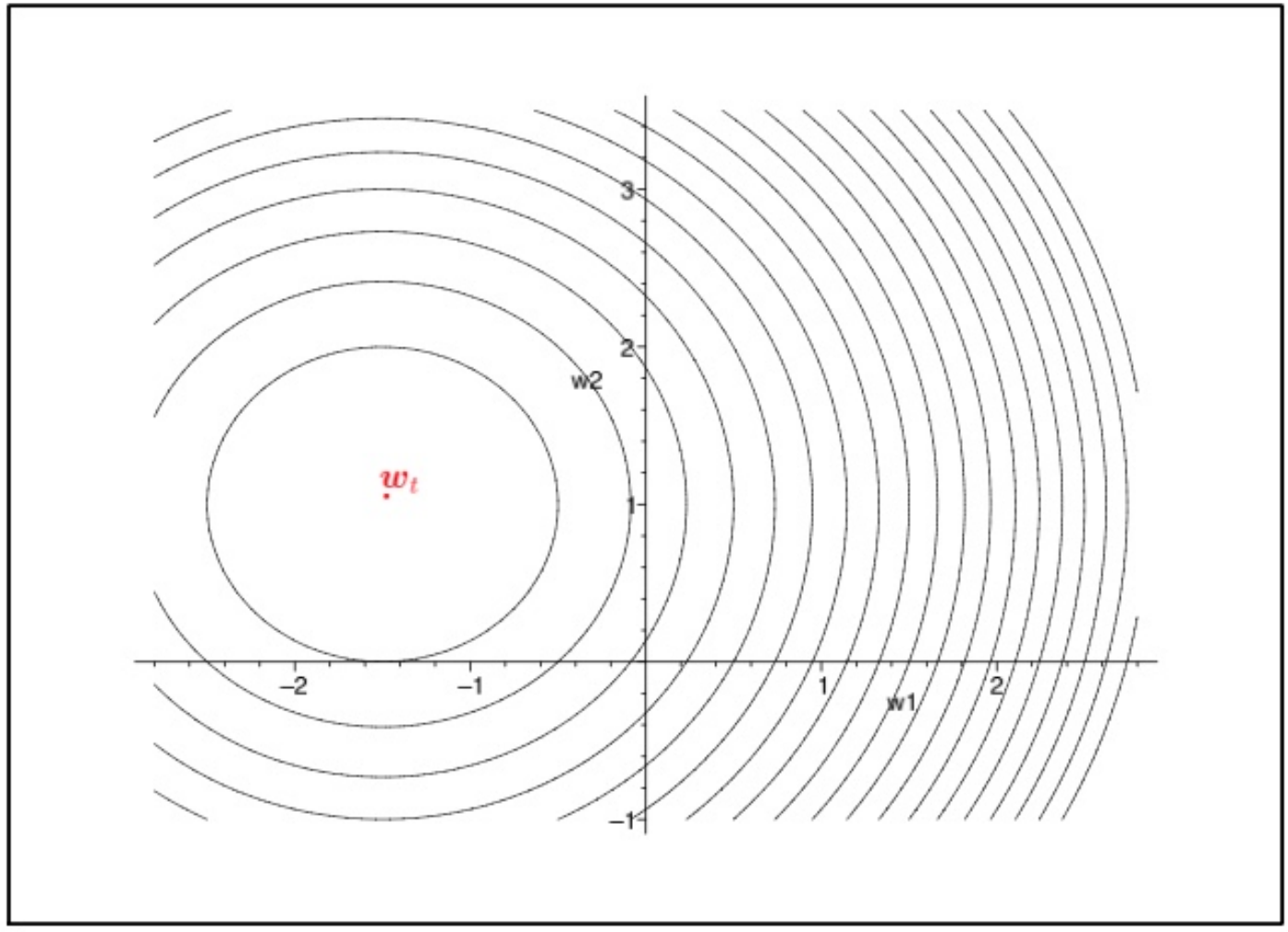
Divergence: Euclidean Distance Squared

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

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

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

$y_t = 1$



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

Loss & divergence are dependent

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

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

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

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

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

First step: Teleskoping

Summing over t [WJ,KW]

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

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

$$\leq \sum_t L_t(\boldsymbol{u}) + \frac{1}{\eta}\Delta_F(\boldsymbol{u}, \boldsymbol{w}_1) + \frac{1}{\eta}\sum_t \Delta_F(\boldsymbol{w}_t, \boldsymbol{w}_{t+1})$$

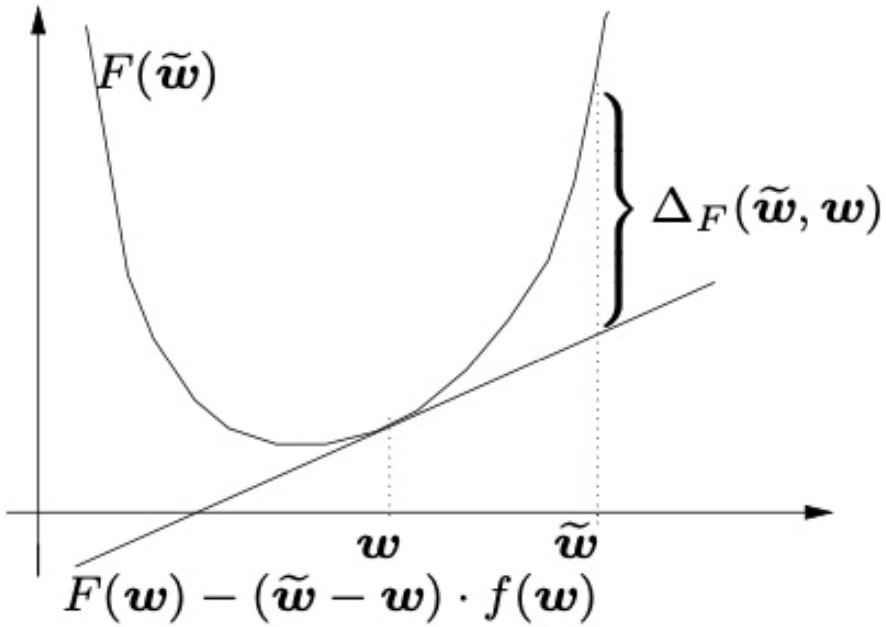
Any convex loss and any Bregman divergence!

Bregman Divergences [Br,CL,Cs]

For any differentiable convex function F

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

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



First step: Teleskoping

Summing over t [WJ,KW]

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

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

$$\leq \sum_t L_t(\boldsymbol{u}) + \frac{1}{\eta}\Delta_F(\boldsymbol{u}, \boldsymbol{w}_1) + \frac{1}{\eta}\sum_t \Delta_F(\boldsymbol{w}_t, \boldsymbol{w}_{t+1})$$

Any convex loss and any Bregman divergence!

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

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

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

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

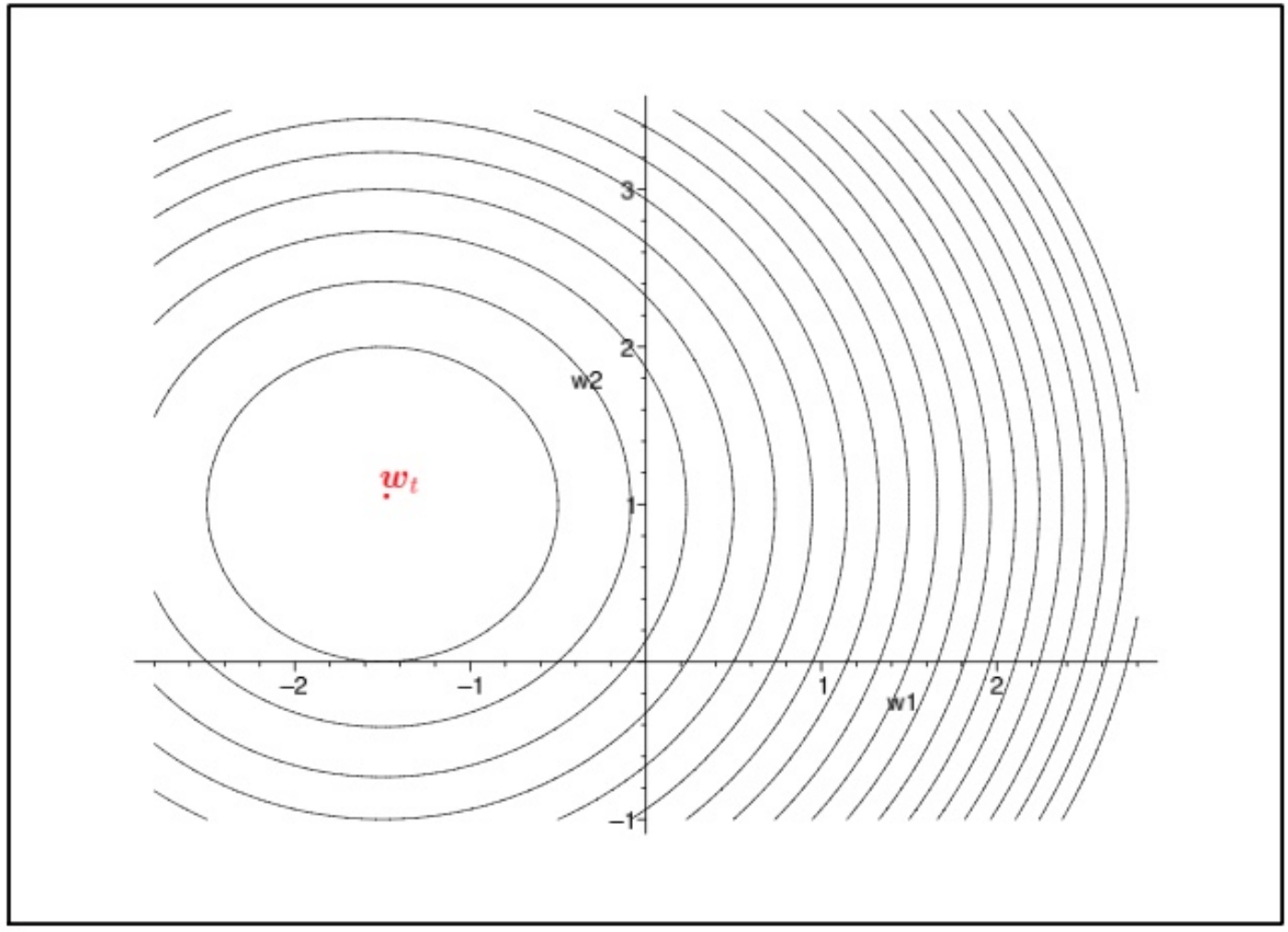
Divergence: Euclidean Distance Squared

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

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

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

$y_t = 1$



General Motivation of Updates [KW]

Trade-off between two term:

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

$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)$ 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

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

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

Examples-2 [GLS,GL]

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

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

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

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

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

Examples

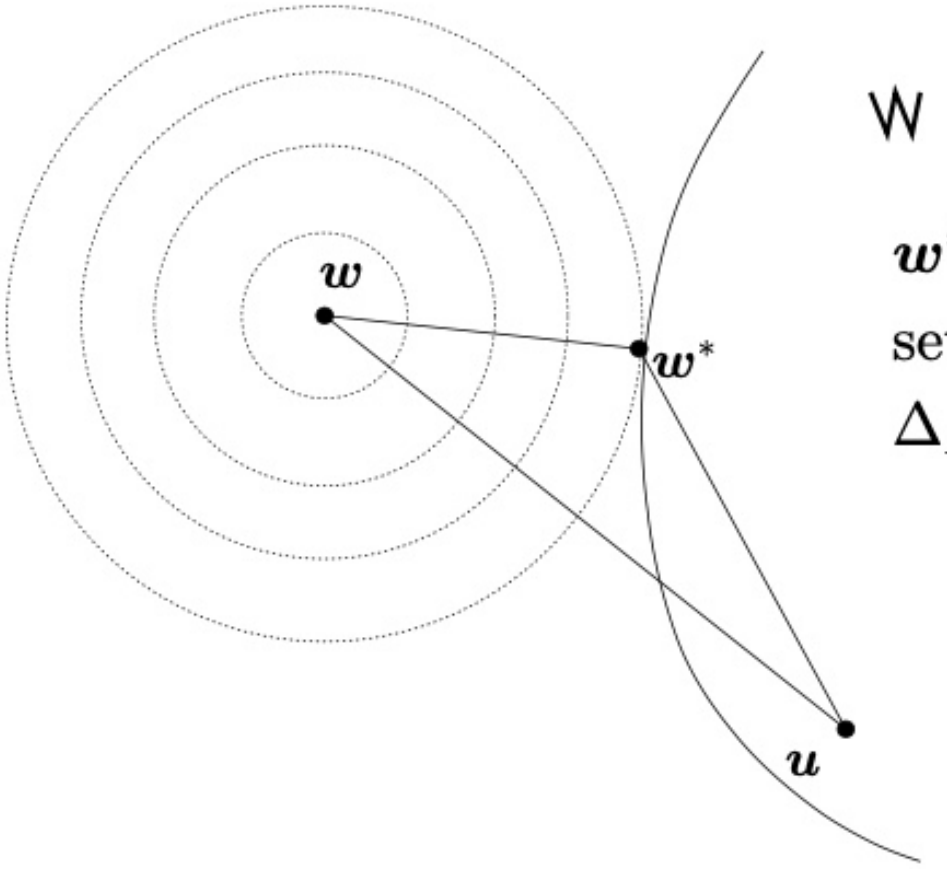
Squared Euclidean Distance

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

(Unnormalized) Relative Entropy

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

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



$\mathcal{W}$

$\boldsymbol{w}^*$ 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:

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

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

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

Bregman Divergences: Simple Properties

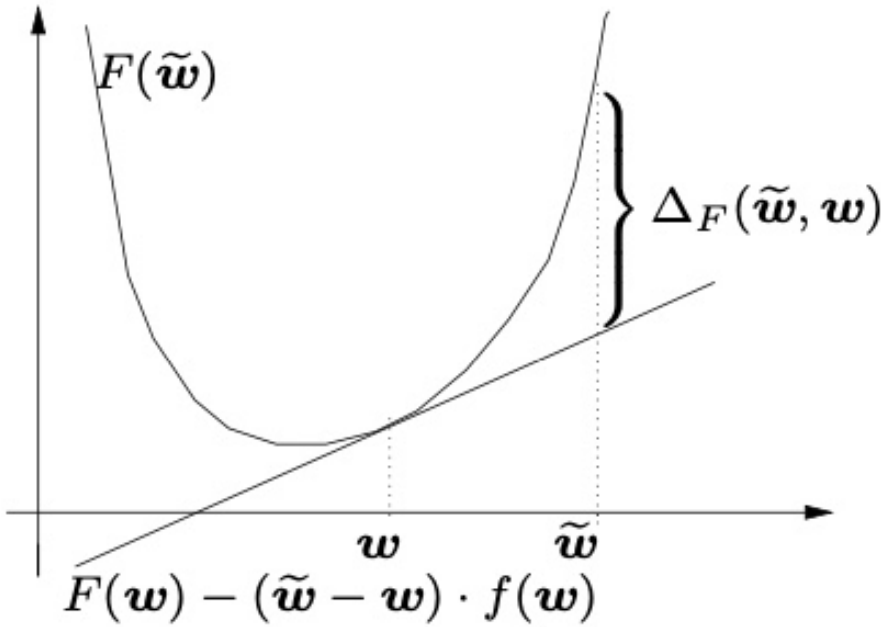
1. $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$ is convex in $\widetilde{\boldsymbol{w}}$
2. $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) \geq 0$
   If F convex equality holds iff $\widetilde{\boldsymbol{w}} = \boldsymbol{w}$
3. Usually not symmetric: $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) \neq \Delta_F(\boldsymbol{w}, \widetilde{\boldsymbol{w}})$
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 [Br,CL,Cs]

For any differentiable convex function F

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

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



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:

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

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

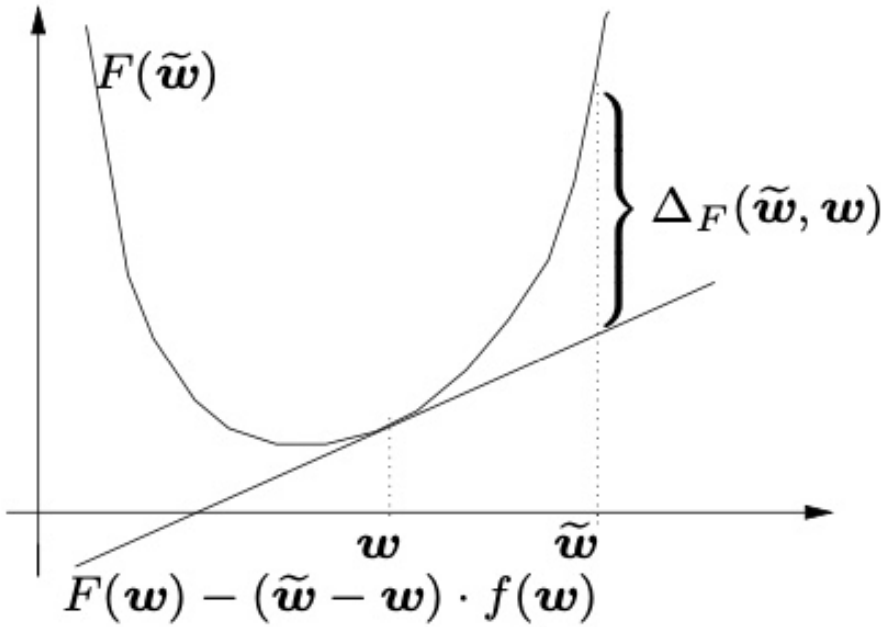
$$\bigtriangledown_\omega (L(\omega, x_t) + \eta D(\omega, \omega_t))$$

Bregman Divergences [Br,CL,Cs]

For any differentiable convex function F

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

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



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

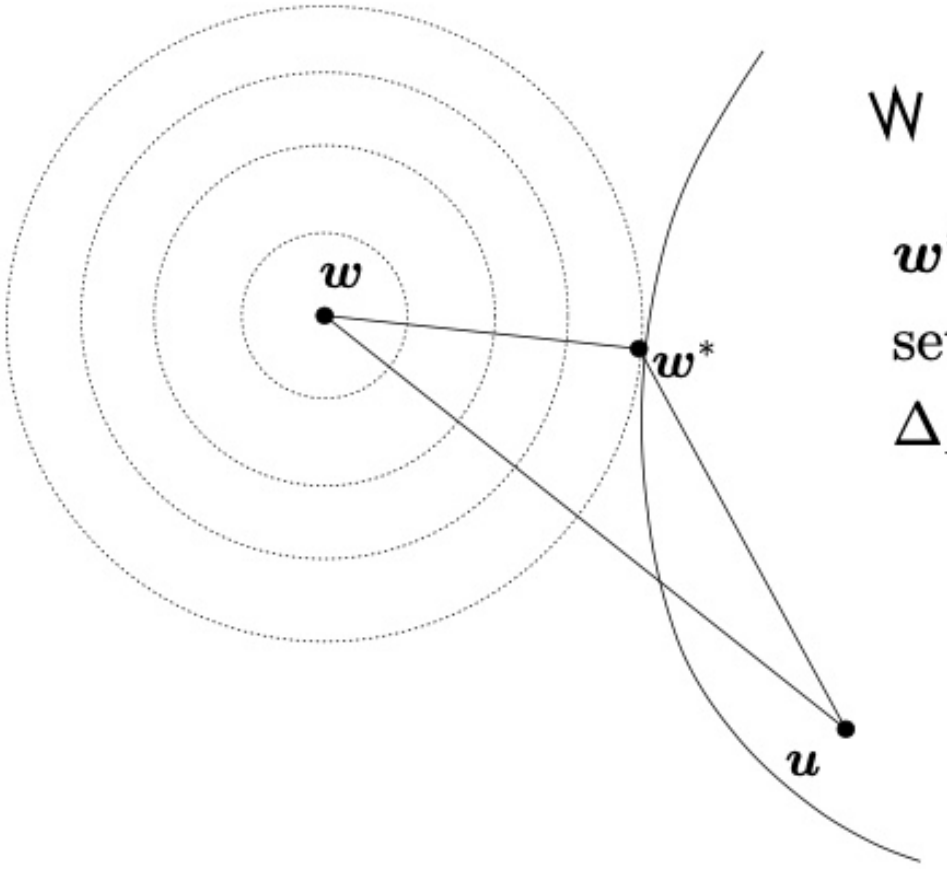
6. $\nabla_{\widetilde{\boldsymbol{w}}} \Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$

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

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

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

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



$\mathcal{W}$

$\boldsymbol{w}^*$ 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

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

Examples-2 [GLS,GL]

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

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

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

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

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

General Motivation of Updates [KW]

Trade-off between two term:

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

$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)$ 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

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

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

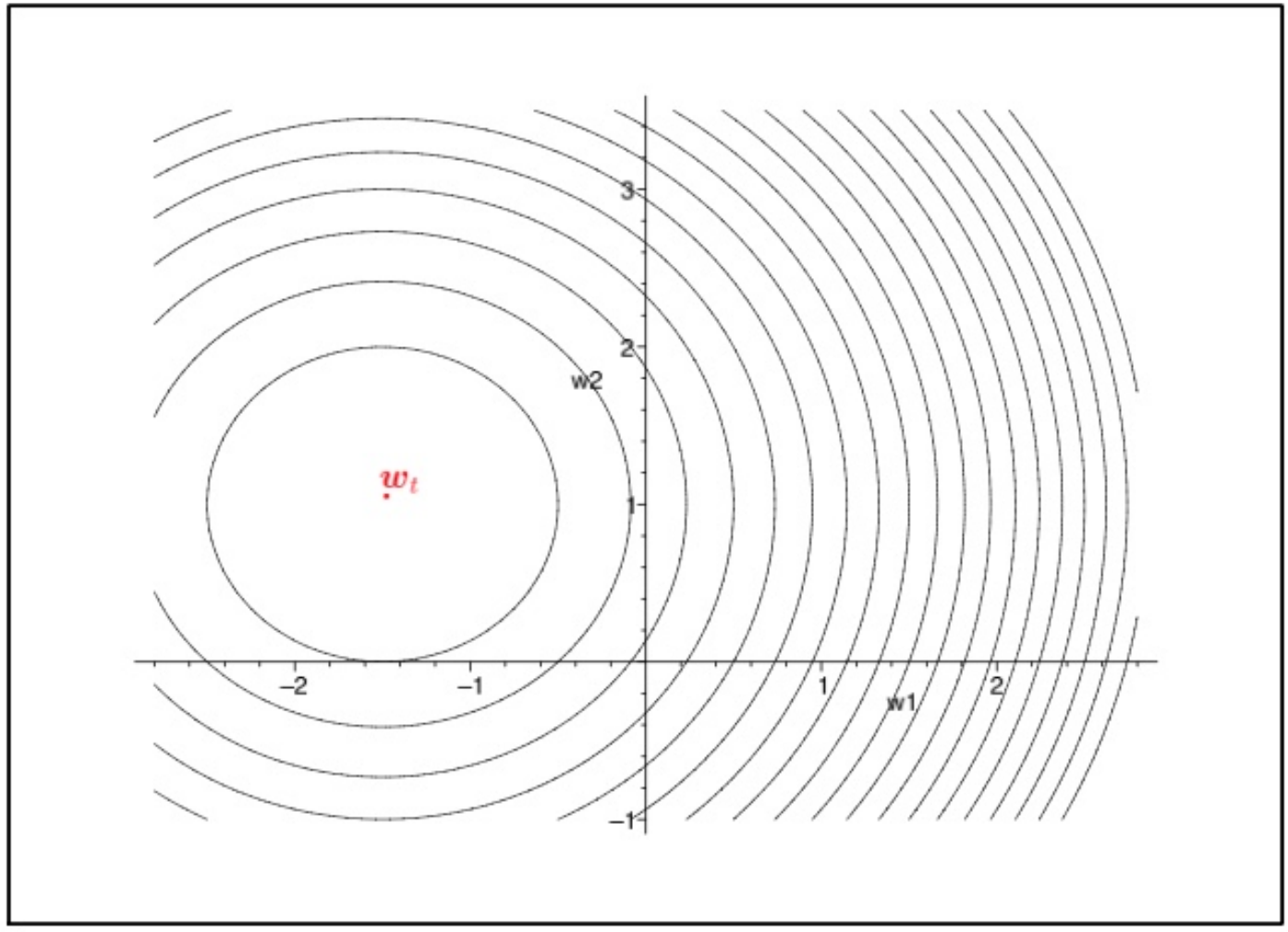
Divergence: Euclidean Distance Squared

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

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

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

$y_t = 1$



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

Loss & divergence are dependent

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

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

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

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

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

First step: Teleskoping

Summing over t [WJ,KW]

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

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

$$\leq \sum_t L_t(\boldsymbol{u}) + \frac{1}{\eta}\Delta_F(\boldsymbol{u}, \boldsymbol{w}_1) + \frac{1}{\eta}\sum_t \Delta_F(\boldsymbol{w}_t, \boldsymbol{w}_{t+1})$$

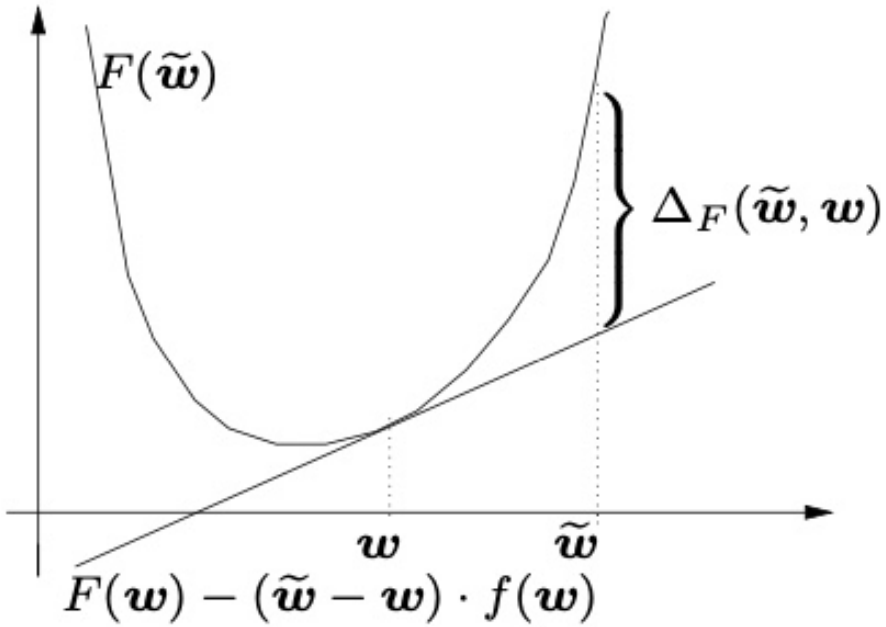
Any convex loss and any Bregman divergence!

Bregman Divergences [Br,CL,Cs]

For any differentiable convex function F

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

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



First step: Teleskoping

Summing over t [WJ,KW]

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

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

$$\leq \sum_t L_t(\boldsymbol{u}) + \frac{1}{\eta}\Delta_F(\boldsymbol{u}, \boldsymbol{w}_1) + \frac{1}{\eta}\sum_t \Delta_F(\boldsymbol{w}_t, \boldsymbol{w}_{t+1})$$

Any convex loss and any Bregman divergence!

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

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

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

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

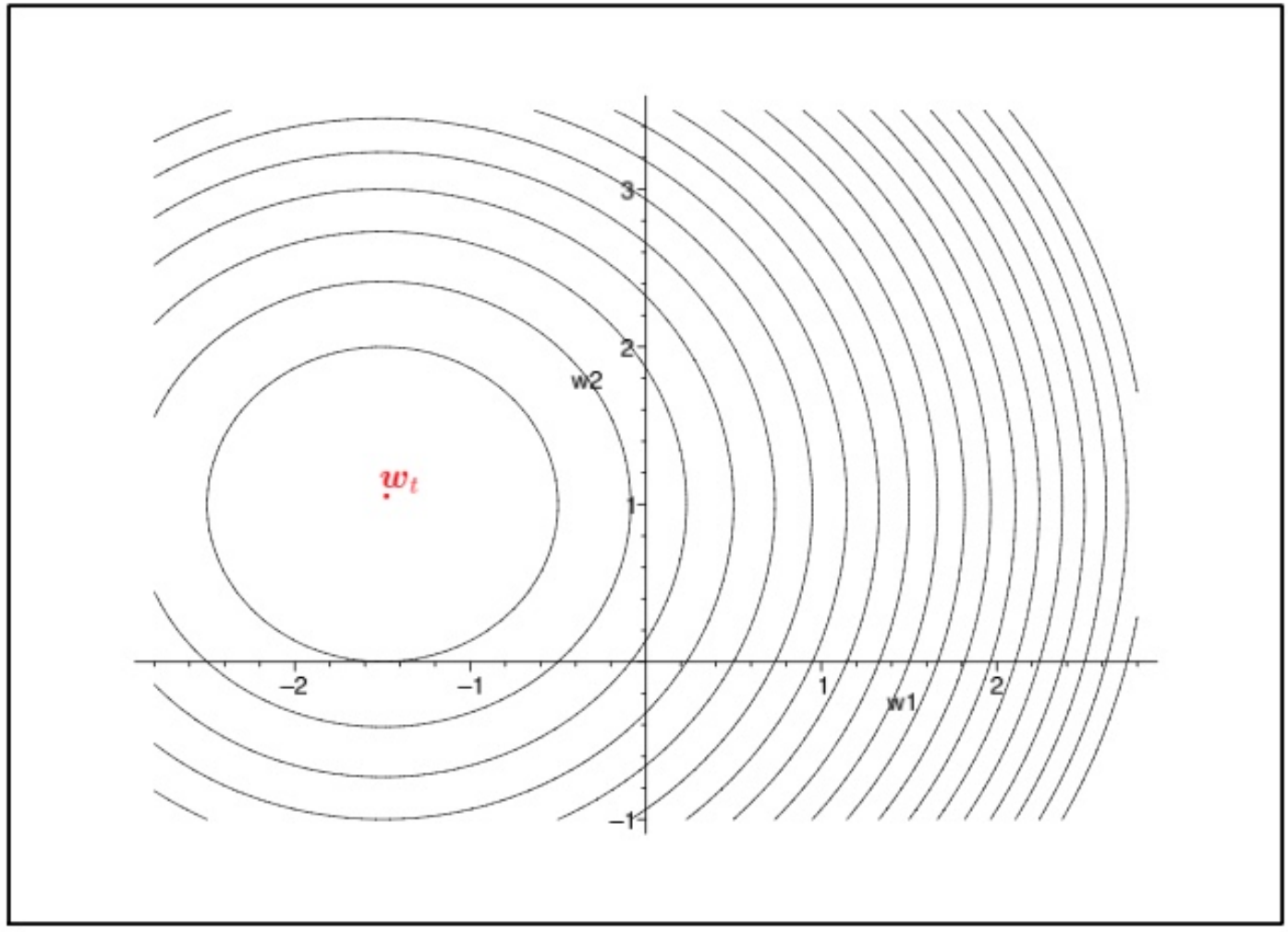
Divergence: Euclidean Distance Squared

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

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

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

$y_t = 1$



General Motivation of Updates [KW]

Trade-off between two term:

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

$\Delta_F(\boldsymbol{w}, \boldsymbol{w}_t)$ 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

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

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

Examples-2 [GLS,GL]

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

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

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

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

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

Examples

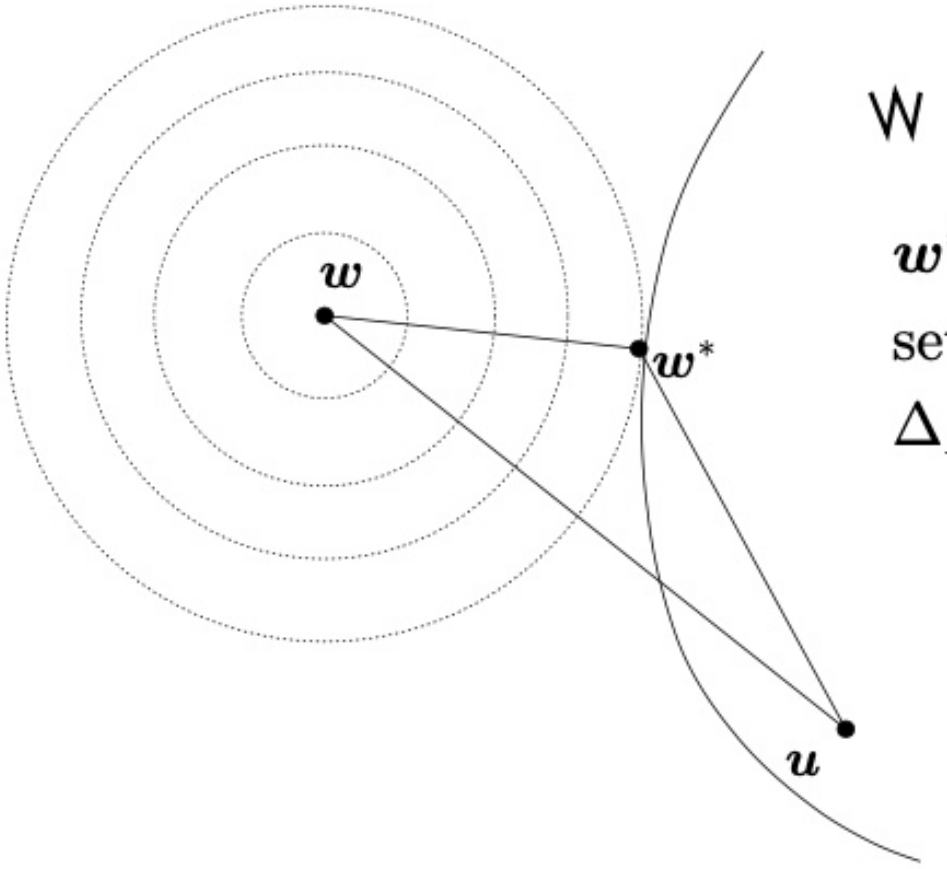
Squared Euclidean Distance

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

(Unnormalized) Relative Entropy

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

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



$\mathcal{W}$

$\boldsymbol{w}^*$ 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:

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

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

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

Bregman Divergences: Simple Properties

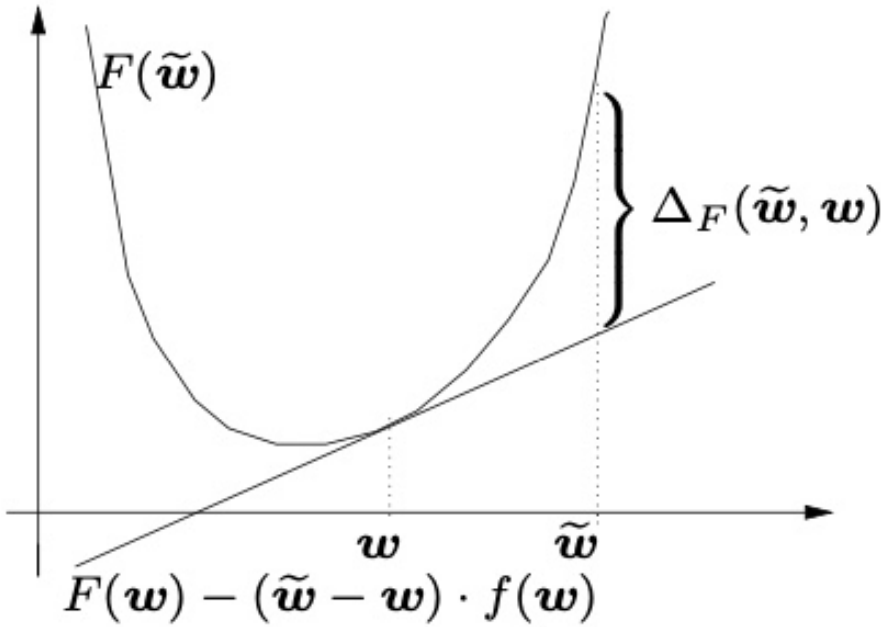
1. $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w})$ is convex in $\widetilde{\boldsymbol{w}}$
2. $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) \geq 0$
   If F convex equality holds iff $\widetilde{\boldsymbol{w}} = \boldsymbol{w}$
3. Usually not symmetric: $\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) \neq \Delta_F(\boldsymbol{w}, \widetilde{\boldsymbol{w}})$
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 [Br,CL,Cs]

For any differentiable convex function F

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

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

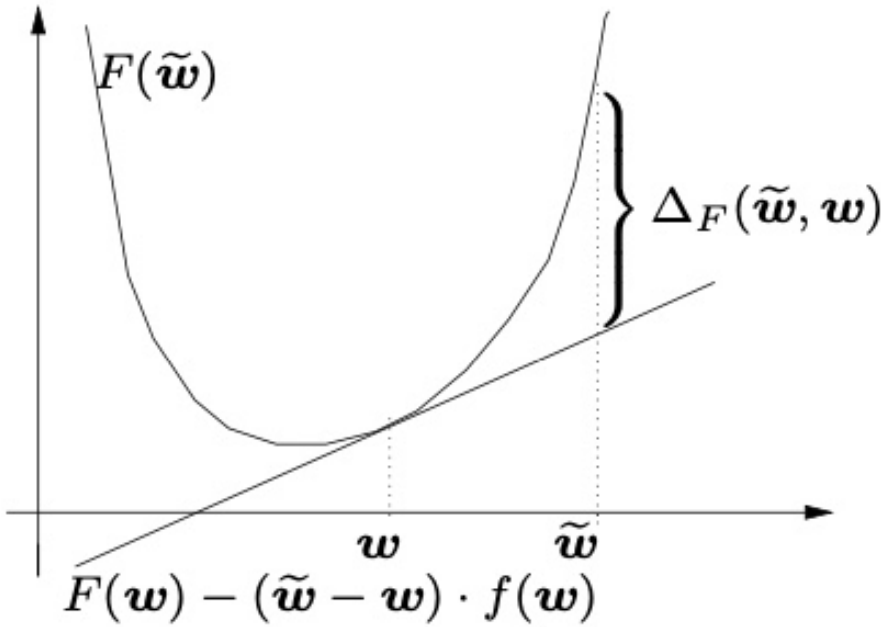


Bregman Divergences [Br,CL,Cs]

For any differentiable convex function F

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

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



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:

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

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

$$\bigtriangledown_\omega (L(\omega, x_t) + \eta D(\omega, \omega_t))$$