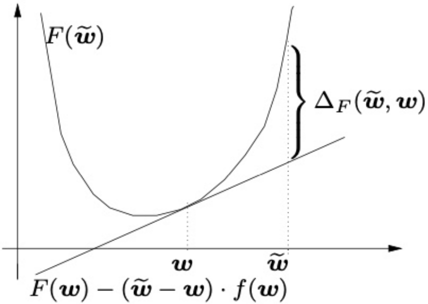
Bregman Divergences [Br,CL,Cs]

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

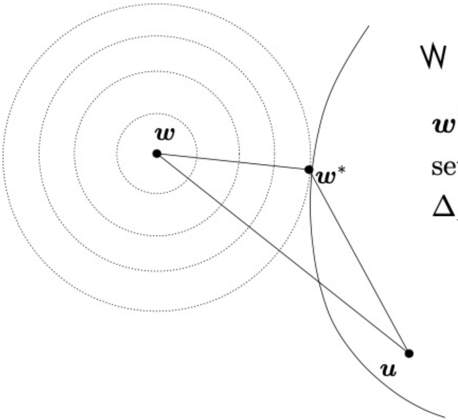
6. $\nabla_{\widetilde{\boldsymbol{w}}}\Delta_F(\widetilde{\boldsymbol{w}},\boldsymbol{w})$

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

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]



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

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

(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: $\frac{1}{p} + \frac{1}{q} = 1$)

$$
\begin{aligned}
F(\boldsymbol{w}) &= \frac{1}{2}||\boldsymbol{w}||_q^2 \\
f(\boldsymbol{w}) &= \nabla\frac{1}{2}||\boldsymbol{w}||_q^2 \\
\Delta_F(\widetilde{\boldsymbol{w}}, \boldsymbol{w}) &= \frac{1}{2}||\widetilde{\boldsymbol{w}}||_q^2 + \frac{1}{2}||\boldsymbol{w}||_q^2 - \widetilde{\boldsymbol{w}} \cdot f(\boldsymbol{w})
\end{aligned}
$$

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

General Motivation of Updates [KW]

Trade-off between two term:

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

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

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

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

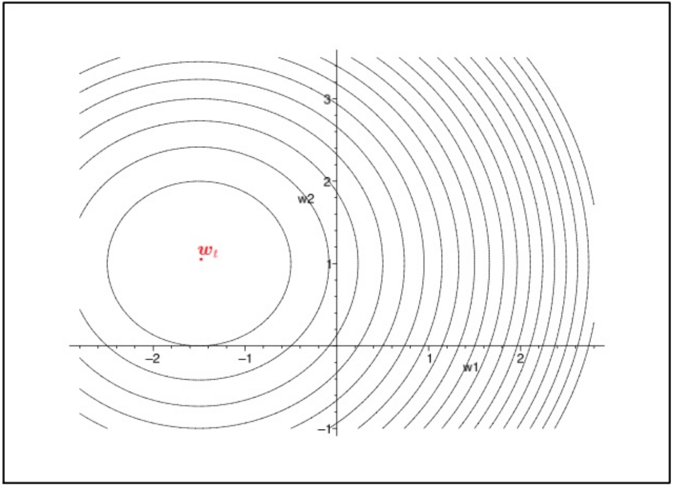
Divergence: Euclidean Distance Squared

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

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

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

$y_t = 1$



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

Loss & divergence are dependent

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

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]

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]

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

Any convex loss and any Bregman divergence!

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]

First step: Teleskoping

Summing over t [WJ,KW]

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

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]

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]

First step: Teleskoping

Summing over t [WJ,KW]

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

Any convex loss and any Bregman divergence!