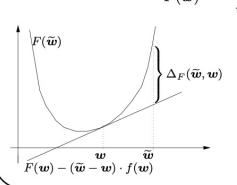
Bregman Divergences [Br,CL,Cs]

For any differentiable convex function F

$$egin{array}{lcl} \Delta_F(\widetilde{m{w}},m{w}) &=& F(\widetilde{m{w}}) - F(m{w}) - (\widetilde{m{w}} - m{w}) \cdot \underbrace{
abla_{m{w}} F(m{w})}_{f(m{w})} \end{array}$$
 $=& F(\widetilde{m{w}}) - rac{ ext{supporting hyperplane}}{ ext{through } (m{w},F(m{w}))}$



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 $a \geq 0$):

$$\Delta_{F+a\,H}(\widetilde{oldsymbol{w}},oldsymbol{w}) = \Delta_{F}(\widetilde{oldsymbol{w}},oldsymbol{w}) + a\,\Delta_{H}(\widetilde{oldsymbol{w}},oldsymbol{w})$$

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

$$\Delta_{H+a\widetilde{oldsymbol{w}}+oldsymbol{b}}(\widetilde{oldsymbol{w}},oldsymbol{w})=\Delta_{H}(\widetilde{oldsymbol{w}},oldsymbol{w})$$

Bregman Divergences: more properties

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

$$= \nabla F(\widetilde{\boldsymbol{w}}) - \nabla_{\widetilde{\boldsymbol{w}}}(\widetilde{\boldsymbol{w}}\nabla_{\boldsymbol{w}}F(\boldsymbol{w}))$$
$$= f(\widetilde{\boldsymbol{w}}) - f(\boldsymbol{w})$$

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

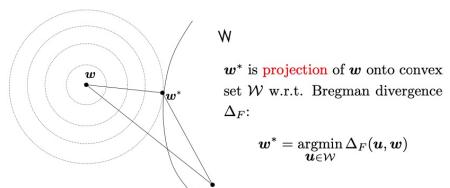
$$= F(\boldsymbol{w}_1) - F(\boldsymbol{w}_2) - (\boldsymbol{w}_1 - \boldsymbol{w}_2) f(\boldsymbol{w}_2)$$

$$F(\boldsymbol{w}_2) - F(\boldsymbol{w}_2) - (\boldsymbol{w}_1 - \boldsymbol{w}_2) f(\boldsymbol{w}_2)$$

$$F(\mathbf{w}_2) - F(\mathbf{w}_3) - (\mathbf{w}_2 - \mathbf{w}_3) f(\mathbf{w}_3)$$

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

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



Theorem:

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

Examples

Squared Euclidean Distance

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

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

(Unnormalized) Relative Entropy

$$F(\boldsymbol{w}) = \sum_{i} (w_i \ln w_i - w_i)$$

$$\frac{\widetilde{i}}{i} + w_i - \widetilde{w_i}$$

$$egin{array}{lcl} f(oldsymbol{w}) &=& \ln oldsymbol{w} \ \Delta_F(\widetilde{oldsymbol{w}},oldsymbol{w}) &=& \sum_i \left(\widetilde{w_i} \, \ln rac{\widetilde{w_i}}{w_i} + w_i - \widetilde{w_i}
ight) \end{array}$$

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:

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

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

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

$$abla_{oldsymbol{w}}(\Delta_F(oldsymbol{w},oldsymbol{w}_t)+oldsymbol{\eta_t}L_t(oldsymbol{w}))=0$$

i

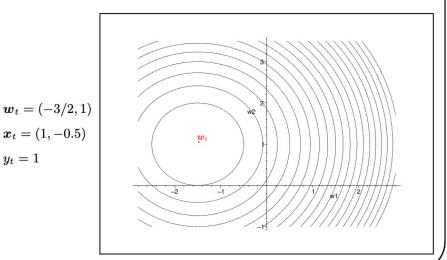
$$f(\boldsymbol{w}) - f(\boldsymbol{w}_t) + \frac{\eta_t}{\approx \nabla L_t(\boldsymbol{w}_t)} = 0$$

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

Divergence: Euclidean Distance Squared

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

 $y_t = 1$



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

s [AG

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

How to prove relative loss bounds?

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

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

$$L_t(\boldsymbol{u}) \quad \widehat{\geq} \quad L_t(\boldsymbol{w}_t) + (\boldsymbol{u} - \boldsymbol{w}_t) \cdot \underline{\nabla_{\boldsymbol{w}} L_t(\boldsymbol{w}_t)}$$

$$egin{array}{lll} L_t(oldsymbol{u}) & \geq & L_t(oldsymbol{w}_t) + (oldsymbol{u} - oldsymbol{w}_t) \cdot oldsymbol{ar{v}} oldsymbol{w}_t L_t(oldsymbol{w}_t) \ & & ext{update} \ & = & L_t(oldsymbol{w}_t) - rac{1}{2}(oldsymbol{u} - oldsymbol{w}_t) \cdot (oldsymbol{f}(oldsymbol{w}_t)) \ \end{array}$$

Loss:

$$= L_t(\boldsymbol{w}_t) - \frac{1}{2} (\boldsymbol{u} - \boldsymbol{w}_t) \cdot (f(\boldsymbol{w}_{t+1}) - f(\boldsymbol{w}_t))$$

$$= L_t(\boldsymbol{w}_t) - \frac{1}{n} (\boldsymbol{u} - \boldsymbol{w}_t) \cdot (\boldsymbol{f}(\boldsymbol{w}_{t+1}) - \boldsymbol{f}(\boldsymbol{w}_t))$$

$$= L_t(oldsymbol{w}_t) - rac{1}{\eta} \underbrace{(oldsymbol{u} - oldsymbol{w}_t) \cdot (oldsymbol{f}(oldsymbol{w}_{t+1}) - oldsymbol{f}(oldsymbol{w}_t))}_{}$$

$$(v_t) - rac{1}{\eta} \underbrace{(oldsymbol{u} - oldsymbol{w}_t) \cdot (f(oldsymbol{w}_{t+1}) - f(oldsymbol{w}_t))}_{oldsymbol{v}}$$

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

$$\frac{1}{\eta} \underbrace{\left(\frac{u - w_t}{w_t} \right) \cdot \left(\frac{w_{t+1}}{w_{t+1}} \right) - \frac{1}{\eta} \left(\frac{w_t}{w_t} \right) }_{\text{prop. } 7 \text{ of } \Delta}$$

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

$$\eta \stackrel{\text{(at)}}{=} \eta \stackrel{\text{(b)}}{=} \eta \stackrel$$

First step: Teleskoping

Summing over t

$$\sum_t L_t(oldsymbol{w}_t) \leq \sum_t L_t(oldsymbol{u}) + rac{1}{oldsymbol{\eta}} \sum_t \Bigl(\Delta_F(oldsymbol{u}, oldsymbol{w}_t) - \Delta_F(oldsymbol{u}, oldsymbol{w}_{t+1}) + \Delta_F(oldsymbol{w}_t, oldsymbol{w}_{t+1}) \Bigr)$$

$$egin{aligned} &+\Delta_F(oldsymbol{w}_t,oldsymbol{w}_{t+1}) \ &\leq \sum_{oldsymbol{t}} L_{oldsymbol{t}}(oldsymbol{u}) + rac{1}{oldsymbol{\eta}} \Big(\Delta_F(oldsymbol{u},oldsymbol{w}_1) - \underbrace{\Delta_F(oldsymbol{u},oldsymbol{w}_{T+1})} \Big) \end{aligned}$$

[WJ,KW]

$$+rac{1}{m}{\sum \Delta_F(oldsymbol{w}_t,oldsymbol{w}_{t+1})}$$

$$+rac{1}{n}{\sum\Delta_F(oldsymbol{w}_t,oldsymbol{w}_{t+1})}$$

$$+rac{1}{oldsymbol{\eta}}{\sum_t}\Delta_F(oldsymbol{w}_t,oldsymbol{w}_{t+1})$$

$$+rac{1}{oldsymbol{\eta}}{\sum_t}\Delta_F(oldsymbol{w}_t,oldsymbol{w}_{t+1})$$

$$\frac{\eta}{\eta} \sum_{t} \Delta_{F}(\omega_{t}, \omega_{t+1})$$

$$\eta \frac{1}{t}$$

$$\sum I_{n}(u) + \frac{1}{2} \Delta_{n}(u, u) + \frac{1}{2} \sum \Delta_{n}(u, u)$$

$$< \sum L_t(oldsymbol{u}) + rac{1}{\epsilon} \Delta_E(oldsymbol{u}, oldsymbol{w}_1) + rac{1}{\epsilon} \sum \Delta_E(oldsymbol{w}_t, oldsymbol{w}_{t+1})$$

$$\leq \sum L_t(oldsymbol{u}) + rac{1}{n} \Delta_F(oldsymbol{u}, oldsymbol{w}_1) + rac{1}{n} \sum \Delta_F(oldsymbol{w}_t, oldsymbol{w}_{t+1})$$

Any convex loss and any Bregman divergence!

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

s [AG

First step: Teleskoping

Summing over
$$t$$

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

$$+\Delta_F(oldsymbol{w}_t,oldsymbol{w}_{t+1})\Big)$$

$$\leq \qquad \sum_{t} L_{t}(\boldsymbol{u}) + \frac{1}{\eta} \Big(\Delta_{F}(\boldsymbol{u}, \boldsymbol{w}_{1}) - \underbrace{\Delta_{F}(\boldsymbol{u}, \boldsymbol{w}_{T+1})}_{\geq 0} \Big)$$

$$\leq \sum_t L_t(oldsymbol{u}) + rac{1}{\eta} \left(\Delta_F(oldsymbol{u},oldsymbol{w}_1) - \underbrace{\Delta_F(oldsymbol{u},oldsymbol{w}_{T+1})}_{\geq 0}
ight)$$

$$\frac{1}{t}$$

$$\geq 0$$

$$+\frac{1}{2}\sum \Delta_E(w_t,w_{t+1})$$

$$+rac{1}{\eta}{\sum_t}\Delta_F(oldsymbol{w}_t,oldsymbol{w}_{t+1})$$

$$+\frac{1}{n}\sum \Delta_F(\boldsymbol{w}_t, \boldsymbol{w}_{t+1})$$

$$+rac{1}{oldsymbol{\eta}}\sum_t \Delta_F(oldsymbol{w}_t,oldsymbol{w}_{t+1})$$

[WJ,KW]

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

Any convex loss and any Bregman divergence!

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

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

a, b constants, a > 1.

Regret bounds (a = 1):

time changing η , subtler analysis

AG]

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

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

a, b constants, a > 1.

Regret bounds (a = 1):

time changing η , subtler analysis

AG]

First step: Teleskoping

 $\sum L_t(\boldsymbol{u}) + \frac{1}{n} \left(\Delta_F(\boldsymbol{u}, \boldsymbol{w}_1) - \underbrace{\Delta_F(\boldsymbol{u}, \boldsymbol{w}_{T+1})} \right)$

 $\leq \sum L_t(oldsymbol{u}) + rac{1}{n} \Delta_F(oldsymbol{u}, oldsymbol{w}_1) + rac{1}{n} \sum \Delta_F(oldsymbol{w}_t, oldsymbol{w}_{t+1})$

 $+\frac{1}{\eta}\sum_{t}\Delta_F(\boldsymbol{w}_t,\boldsymbol{w}_{t+1})$

[WJ,KW]

Summing over t

Summing over
$$t$$

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

Any convex loss and any Bregman divergence!

 $+\Delta_F(oldsymbol{w}_t,oldsymbol{w}_{t+1})\Big)$