## 5.8 Regret bounds with Bergman divergences

We generalize the notion of distance by Bregman div.

> **[Definition 28] Bregman divergence.**
> Let $f$ be a continuously-differentiable convex function.
> The Bregman divergence is defined as:
> $$D_f(w \| u) := f(w) - f(u) - \nabla f(u) \cdot (w - u)$$

Intuitively, $D_f(w\|u)$ captures the error of the linear approximation of $f$ based on $\nabla f(u)$.

Property: $D_f(w\|u) \geq 0$ (since $f$ is convex)

Note: $D_f$ is not symmetric so no distance metrix.

**[Example 34] Quadratic regularizer**

Let $f(w) = \frac{1}{2}\|w\|_2^2$

Then $D_f(w\|u) = \frac{1}{2}\|w\|_2^2 - \frac{1}{2}\|u\|_2^2 - u\cdot(w-u) = \frac{1}{2}\|w-u\|_2^2$

**[Example 35] Entropic**

Let $f(w) = \sum_{j=1}^{d} w_j \log w_j$ for $w \in \Delta_d$

The $D_f(w\|u) = \sum_{j=1}^{d} w_j \log w_j - \sum_{j=1}^{d} u_j \log u_j - \sum_{j=1}^{d} (\log u_j + 1)(w_j - u_j)$

$$= \sum_{j=1}^{d} w_j \log \frac{w_j}{u_j} - \sum_{j=1}^{d} w_j + \sum_{j=1}^{d} u_j$$

$$= \sum_{j=1}^{d} w_j \log \frac{w_j}{u_j} \qquad \left( \sum_{j=1}^{d} w_j = \sum_{j=1}^{d} u_j = 1 \right)$$

$$= KL(w\|u)$$

Property (scaling): $D_{af}(w\|u) = a D_f(w\|u)$.

[Theorem 31] regret of OMD using Bregman divergence

(Theorem 31) regret of OMD using Bregman divergence.

OMD obtains the following regret bound:

$$\text{Regret}(u) \le [\psi(u) - \psi(w_1)] + \sum_{t=1}^{T} D_{\psi^*}(\theta_{t+1} \| \theta_t) \quad (528)$$

Pf: Assume all loss functions are linear.

Recall:

* Learner's loss is $\sum_{t=1}^{T} w_t \cdot z_t$
* Expert's loss is $\sum_{t=1}^{T} u \cdot z_t$
* Regret $= \sum_{t=1}^{T} w_t z_t - \sum_{t=1}^{T} u \cdot z_t$

The expert:

$$\begin{aligned} \psi^*(\theta_{T+1}) &\ge u \cdot \theta_{T+1} - \psi(u) \\ &= \sum_{t=1}^{T} (-u \cdot z_t) - \psi(u) \quad ① \end{aligned}$$

Note that we have equality if $u$ is the best expert.

The learner:

$$\begin{aligned} \psi^*(\theta_{T+1}) &= \psi^*(\theta_1) + \sum_{t=1}^{T} [\psi^*(\theta_{t+1}) - \psi^*(\theta_t)] \\ &= \psi^*(\theta_1) + \sum_{t=1}^{T} [\nabla \psi^*(\theta_t)(\theta_{t+1} - \theta_t) + D_{\psi^*}(\theta_{t+1} \| \theta_t)] \\ &= \psi^*(\theta_1) + \sum_{t=1}^{T} [-w_t \cdot z_t + D_{\psi^*}(\theta_{t+1} \| \theta_t)] \quad ② \end{aligned}$$

Note that $\psi^*(\theta_1) = -\psi(w_1)$ $(\theta_1 = 0,\ \psi^*(0) = \sup_w -\psi(w))$

Combining ① and ②, we have:

$$\sum_{t=1}^{T} w_t \cdot z_t - \sum_{t=1}^{T} u \cdot z_t \le \psi(u) - \psi(w_1) + \sum_{t=1}^{T} D_{\psi^*}(\theta_{t+1} \| \theta_t)$$

In OGD we know for any convex loss, linear function is the worst, which finishes the proof □

Note: Th. 31 is a generalization of Th. 30, where $\psi(w) = \frac{1}{2\eta}\|w\|_2^2$ and $D_{\psi^*}(\theta_{t+1} \| \theta_t) = \frac{\eta}{2}\|\theta_{t+1} - \theta_t\|_2^2 = \frac{\eta}{2}\|z_t\|_2^2$.