MLT Notes 1

Wouter M. Koolen

September 13, 2017

1 Notation and Definitions

Definition 1. Fix a differentiable convex function $\phi : \mathbb{R}^k \to \mathbb{R}$. The *Bregman divergence* from $x \in \mathbb{R}^k$ to $y \in \mathbb{R}^k$ generated by ϕ is

$$B_\phi(x, y) \;=\; \phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle$$

where $\langle x, y\rangle$ denotes the dot product $\sum_{i=1}^k x_i y_i$, and $\nabla\phi(y)$ is the gradient (vector of partial derivatives) of ϕ at y.

Definition 2. Let us write $[k] = \{1, \ldots, k\}$, and let us denote the probability simplex by $\triangle_k = \{x \in \mathbb{R}^k \mid \sum_{i=1}^k x_i = 1 \text{ and } \forall i\, x_i \geq 0\}$. Fix a loss function $\ell : [k] \times \triangle_k \to \mathbb{R}$, and let

$$L(p, q) \;=\; \sum_{i=1}^k p_i \ell(i, q)$$

be its associated *risk*, and let

$$\underline{L}(p) \;=\; \inf_q L(p, q)$$

be its *entropy*. A loss function is called *proper* if for all $p, q \in \triangle_k$

$$L(p, p) \;\leq\; L(p, q).$$

2 Results

Here are two results about proper losses and Bregman divergences.

Theorem 1 (Savage'75)**.**

1. *For any loss, the entropy $\underline{L}$ is concave.*
2. *For every differentiable concave $\Lambda : \triangle_k \to \mathbb{R}$, there is a proper loss with entropy $\underline{L}(p) = \Lambda(p)$.*

Proof. 1. Minimum of linear is concave.

2. *In this proof we are following the common special-case notation for* 2 *outcomes, where outcomes are* $\{0,1\}$ *and distributions on these* 2 *outcomes are parametrised by the probability* $q \in [0,1]$ *of observing the outcome* 1. Let
$$\ell(y,q) \;=\; \Lambda(q) + (y-q)\Lambda'(q)$$
Then
$$L(p,q) \;=\; \Lambda(q) + (p-q)\Lambda'(q)$$
Using concavity, we hence find that
$$L(p,q) \;\geq\; \Lambda(p) = L(p,p)$$
and hence ℓ is proper and $\underline{\mathrm{L}}(p) = \Lambda(p)$.

□

Theorem 2. *Fix a differentiable convex* $\phi : \mathbb{R}^d \to \mathbb{R}$*, and let* B_ϕ *be the associated Bregman divergence. Then*

1. Reflexivity: $B_\phi(x,x) = 0$
2. Non-negativity: $B_\phi(x,y) \geq 0$
3. Convexity: $B_\phi(x,y)$ *is convex in* x *for each* y.
4. Generalised Properness: $\mathbb{E}[X] \;=\; \arg\min_y \mathbb{E}\left[B_\phi(X,y)\right]$.
5. Generalised Pythagorean Inequality: *Fix a convex set* $C \subseteq R^d$ *and* $y \in \mathbb{R}^d$*, and let*
$$\hat{y} \;=\; \arg\min_{\hat{y}\in C} B_\phi(\hat{y},y)$$
Then for any $x \in C$ *we have*
$$B_\phi(x,\hat{y}) + B_\phi(\hat{y},y) \;\leq\; B_\phi(x,y)$$

Proof. 1. Homework 1, question 2(a).

2. Homework 1, question 2(b).
3. Homework 1, question 2(c).
4. Homework 1, question 4.
5. By the first order optimality condition for $\hat{y}$, we know that for all $x \in C$
$$\langle x - \hat{y}, \nabla_{\hat{y}} B_\phi(\hat{y},y)\rangle \;\geq\; 0 \qquad (1)$$
and since $\nabla_{\hat{y}} B_\phi(\hat{y},y) = \nabla\phi(\hat{y}) - \nabla\phi(y)$ we have
$$\langle x - \hat{y}, \nabla\phi(\hat{y}) - \nabla\phi(y)\rangle \;\geq\; 0$$

It remains to show

$$\underbrace{\phi(x) - \phi(\hat{y}) - \langle x - \hat{y}, \nabla\phi(\hat{y})\rangle}_{B_\phi(x,\hat{y})} + \underbrace{\phi(\hat{y}) - \phi(y) - \langle \hat{y} - y, \nabla\phi(y)\rangle}_{B_\phi(\hat{y},y)} \;\leq\; \underbrace{\phi(x) - \phi(y) - \langle x - y, \nabla\phi(y)\rangle}_{B_\phi(x,y)}$$

that is

$$\langle x - \hat{y}, \nabla\phi(y) - \nabla\phi(\hat{y})\rangle \;\leq\; 0$$

which is equivalent to (1).

$\square$