MATH 559: BAYESIAN THEORY AND METHODS

A VARIATIONAL BOUND FOR THE KULLBACK-LEIBLER DIVERGENCE

Recall that the Kullback-Leibler divergence between densities p and q is

$$KL(p,q) = \int \log \frac{p(y)}{q(y)}\, p(y)\, dy.$$

This is a non-negative quantity, and is finite provided $q(y) > 0$ whenever $p(y) > 0$.

Lemma: Consider two distributions p and q with common support $\mathcal{Y}$, and let T be a generic scalar function of Y. Then the KL divergence admits the following dual representation

$$KL(p,q) = \sup_{T:\mathcal{Y}\to\mathbb{R}} \{\mathbb{E}_p[T] - \log \mathbb{E}_q[\exp\{T\}]\}$$

provided the two expectations are finite.

Proof. Suppose $T \equiv T(Y)$, with $\mathbb{E}_p[\exp\{T\}] < \infty$. Then as $\mathbb{E}_p[T] < \mathbb{E}_p[\exp\{T\}] < \infty$, we have

$$\begin{aligned}
KL(p,q) = \int \log \frac{p(y)}{q(y)}\, p(y)\, dy &= \int \left[\log \frac{p(y)}{q(y)} - T(y) + T(y)\right]\, p(y)\, dy \\
&= \int \left[\log \frac{p(y)}{q(y)} - T(y)\right]\, p(y)\, dy + \int T(y)\, p(y)\, dy \\
&= \mathbb{E}_p\left[\log \frac{p(Y)\exp\{-T\}}{q(Y)}\right] + \mathbb{E}_p[T] \\
&= \mathbb{E}_p[T] - \mathbb{E}_p\left[\log \frac{q(Y)}{p(Y)}\exp\{T\}\right] \\
&\geq \mathbb{E}_p[T] - \log \mathbb{E}_p\left[\frac{q(Y)}{p(Y)}\exp\{T\}\right] && \text{Jensen} \\
&= \mathbb{E}_p[T] - \log \mathbb{E}_q\left[\frac{q(Y)}{p(Y)}\exp\{T\}\frac{p(Y)}{q(Y)}\right] && \text{Change of measure} \\
&= \mathbb{E}_p[T] - \log \mathbb{E}_q\left[\exp\{T\}\right]
\end{aligned}$$

The right-hand side is therefore a lower-bound on $KL(p,q)$, but note that by Jensen's inequality this bound is achieved when

$$T = \log \frac{p(Y)}{q(Y)} + c$$

for some constant c, and this T is therefore the largest lower bound. ■

We consider this inequality applied to the prior-to-posterior calculation: we have for any distribution π on θ that

$$KL(\pi,\pi_0) \geq \mathbb{E}_\pi[g(\theta)] - \log \mathbb{E}_{\pi_0}[\exp\{g(\theta)\}]$$

with equality when

$$g(\theta) = \log \frac{\pi(\theta)}{\pi_0(\theta)} + c$$

or equivalently

$$\pi(\theta) = \exp\{c\}\exp\{g(\theta)\}\pi_0(\theta) = C\exp\{g(\theta)\}\pi_0(\theta)$$

say, where we must have

$$C^{-1} = \int \exp\{g(\theta)\}\pi_0(\theta)\, d\theta.$$

- The Bayesian posterior is obtained if

$$g(\theta) \equiv g(\mathbf{y}, \theta) = \sum_{i=1}^{n} \log f_Y(y_i; \theta) = \ell_n(\theta)$$

that is, the log-likelihood.

- If we reverse the roles of π and π_0, and instead compute $KL(\pi_0, \pi)$,

$$KL(\pi_0, \pi) \geq \mathbb{E}_{\pi_0}[h(\theta)] - \log \mathbb{E}_{\pi}[\exp\{h(\theta)\}]$$

we have that the bound is achieved when

$$h(\theta) = \log \frac{\pi_0(\theta)}{\pi(\theta)} + b$$

so that

$$\pi_0(\theta) = \exp\{b\} \exp\{h(\theta)\}\pi(\theta) = B \exp\{h(\theta)\}\pi(\theta)$$

and so deduce the identical result with $h = -g$.

- For 'optimal' inference, we need to choose the best g. We consider a Bayesian decision problem based on a *utility* (or negative loss) function u, and the choose to report the distribution π^*, a function of data $\mathbf{Y}$, such that

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi_0} \left[\mathbb{E}_{Y|\theta} \left[u(\pi, \theta); \theta \right] \right].$$

Suppose that u has the properties

(i) *Proper*: for any other π', the expected utility wrt π,

$$\int u(\pi', \theta)\pi(\theta)\, d\theta$$

is uniquely maximized when $\pi' \equiv \pi$.

(ii) *Local*: the utility function satisfies

$$u(\pi, \theta) = u(\pi(\theta), \theta)$$

that is, that the utility value depends only on the utility function evaluated at θ.

We may deduce that u takes the form

$$u(\pi, \theta) = A \log \pi(\theta) + B(\theta).$$

Under this *log* utility, we aim to choose the rule $\pi_{\mathbf{y}}(\theta)$ (i.e. a distribution indexed by data $\mathbf{y}$)

$$\begin{aligned}
\pi_{\mathbf{y}}(\theta) &= \arg\max_{\pi} \iint u(\pi, \theta)\ f_{\mathbf{Y}}(\mathbf{y}; \theta)\pi_0(\theta)\, d\theta\, d\mathbf{y} \\
&= \arg\max_{\pi} \iint \log \pi(\theta)\ f_{\mathbf{Y}}(\mathbf{y}; \theta)\pi_0(\theta)\, d\theta\, d\mathbf{y} \\
&= \arg\min_{\pi} \iint \log \frac{\pi_0(\theta)\ f_{\mathbf{Y}}(\mathbf{y}; \theta)}{\pi(\theta)}\ f_{\mathbf{Y}}(\mathbf{y}; \theta)\pi_0(\theta)\, d\theta\, d\mathbf{y}
\end{aligned}$$

which, by properties of the Kullback-Leibler divergence, implies that

$$\pi_{\mathbf{y}}(\theta)\ \propto\ f_{\mathbf{Y}}(\mathbf{y}; \theta)\pi_0(\theta).$$