Bayesian Analysis (2012) **7**, Number 1, pp. 47–50

Rejoinder

Alessio Sancetta[*]

I thank Professor Clarke for his sharp comments. He stresses the important fact that in most regular cases,

$$\mathbb{E}^{\theta} D_t \left(P_\theta \| P_w\right) = O\left(d/t\right),$$

where d is the dimension of θ (I am using the notation in my text). The intuition is based on his eq. (1), which is one of the results in Clarke and Barron (1994). The analytical case for the exponential family with conjugate priors shows that this is indeed the case. The crucial ingredient is eq. (12). Effectively, the argument should fail as soon as the posterior does not satisfy the Bernstein–von Mises Theorem. That is, the argument relies on asymptotic normality of the posterior distribution, or more precisely on the posterior concentrating around the MLE when the MLE is asymptotically normal. This is clearly stated by Professor Clarke in the last paragraph of his Section 2.

Of course, the behaviour of the t^{th} stage risk is what really matters for the practical problem of prediction. This is not directly addressed in the paper. The paper points out that in most regular cases, the cumulative expected risk is $O(\ln T)$, however, this begs the question of one example when this is not the case. As remarked by Professor Liang, one does not need to look at uncommon circumstances for the supremum of the resolvability index to be infinite. Hence, it is of interest to find examples where this is finite, but grows faster than $\ln T$. One such case is when the prior gives too little weight to some regions in the parameter space. The following is rather artificial: $\Theta = [0,1]$, $w(d\theta) = C\exp\{-\theta^{-c}\}$ for $C, c > 0$. When

$$\mathbb{E}^{\theta}\left[\ln p_{\theta'}\left(Z_t|\mathcal{F}_{t-1}\right) - \ln p_{\theta}\left(Z_t|\mathcal{F}_{t-1}\right)\right] = O\left(|\theta-\theta'|\right),$$

we need $|\theta-\theta'| \leq \delta/T$, but the prior gives weight

$$w\left(B_T(\theta)\right) = O\left(\exp\left\{-\left(\frac{T}{\delta}\right)^c\right\}\left(\frac{\delta}{T}\right)\right)$$

when $\theta = 0$. Hence, taking logs, the resolvability index at 0 is

$$R_T(0) = O\left(\inf_{\delta>0}\left\{\delta + \left(\frac{T}{\delta}\right)^c - \ln\delta + \ln T\right\}\right)$$

which grows faster than $\ln T$, but is still $o(T)$, so that universality holds.

Universality may of course fail, but the resolvability index still be finite uniformly in the parameter space. Consider an AR(1) with autoregressive coefficient $\theta \in [0, 1+\epsilon]$

[*] The author was lecturer at the University of Cambridge until 2008. He has then worked on several algorithmic trading positions in the City of London. He now works as a freelancer. http://sites.google.com/site/wwwsancetta/

DOI:10.1214/12-BA701REJ

for some $\epsilon > 0$ and Lebesgue measure as prior. In this case, from calculations as in Example 2 (taking expectations), when $\theta = 1 + \epsilon$,

$$w\left(B_T\left(1+\epsilon\right)\right) = O\left(\frac{\delta}{\left(1+\epsilon\right)^T}\right)$$

implying that the resolvability index at $1 + \epsilon$ is

$$R_T\left(1+\epsilon\right) = O\left(\inf_{\delta>0}\left\{\delta - \ln\delta + T\ln\left(1+\epsilon\right)\right\}\right) = O\left(T\right),$$

so that, as perhaps expected, the Bayesian predictive density is not a good estimator when we have an explosive root.

I shall now attempt to answer Professor Liang who I thank for her stimulating comments. The first of these comments points to some crucial shortcomings of universality. Indeed, the lack of a compact parameter space bounded away from certain values can lead to infinite loss in common examples like location and scale problems, in which case a diffuse prior solves the problem (Liang and Barron, 2004). Liang and Barron (2004) use some invariance properties for location scale models. One may wish to look at the more general problem when — for example — the regression variable is endogenous, e.g. an AR(1). This is one case where conditioning is needed. However, in this case the parameter space needs to be restricted to $\Theta \in [-1, 1]$, ruling out altogether the need for an unbounded parameter space.

The second question stresses the importance of looking at the loss per observation, rather than the Cesaro sum, as pointed out by Professor Clarke. The estimator $\tilde{p}_w$ derived from the average of the predictive density does not rely on special regularity conditions, as long as (3) in Professor Liang's discussion holds. Consequently, if (4) — in the discussion — were to hold, then we would have the equivalent result uniformly in Θ.

To answer the third question, I give a special example that satisfies universality by directly using the results in the paper. For simplicity, start with one model and keep adding one every new observation. Hence, the support of the posterior increases linearly with the sample size. Using the same notation as in Section 3 in the paper, at $t = 0$,

$$m\left(k|\mathcal{F}_0\right) = m\left(k\right) = \left\{k = 1\right\},$$

as we start with one model only and for $t > 0$,

$$m\left(k|\mathcal{F}_t\right) = \left(1-\lambda\left(t\right)\right)m'\left(k|\mathcal{F}_t\right)\left\{k \le t\right\} + \lambda\left(t\right)\left\{k = t+1\right\},$$

$$m'\left(k|\mathcal{F}_t\right) = \frac{p_{w_k}\left(Z_t|\mathcal{F}_{t-1}\right)m\left(k|\mathcal{F}_{t-1}\right)}{\sum_{k=1}^t p_{w_k}\left(Z_t|\mathcal{F}_{t-1}\right)m\left(k|\mathcal{F}_{t-1}\right)},$$

where $m'\left(k|\mathcal{F}_t\right)$ has support $\{1, 2, ...t\}$. With this choice, the predictive density is

$$p_m\left(Z_t|\mathcal{F}_{t-1}\right) := \sum_{k=1}^t p_{w_k}\left(Z_t|\mathcal{F}_{t-1}\right)m\left(k|\mathcal{F}_{t-1}\right).$$

The above updating scheme is similar in flavour to the one in Section 4.1 of the paper. Hence, it can be given a Bayesian interpretation: at each time t, the number of models stays the same with probability $(1-\lambda(t))$ and with probability $\lambda(t)$ a new model $p_{w_t}(Z_t|\mathcal{F}_{t-1})$ is included in the posterior. Arguments similar to the ones in the proof of Lemma 9 give, for $k \in \{1, 2...., T\}$,

$$D_{1,T}\left(P_{w_k} \| P_m\right) \leq -\ln \lambda(k) - \sum_{l=k+1}^{T} \ln(1-\lambda(k)).$$

Choosing $\lambda(k) = \lambda/k$, from Lemma 12 we deduce the above display is $O(\ln k + \ln T)$. We can then bound $D_{1,T}(P \| P_m)$ using the triangle inequality and the above display. Universality holds as long as it holds for each single model. As the resolvability index for each single model is $O(\ln T)$, in most regular cases, this model averaging procedure does not deteriorate the bound in order of magnitude.