Bayesian information-based learning and majorization

H. P. Wynn
Department of Statistics, London School of Economics, London, UK
h.wynn@lse.ac.uk

November 27, 2007

Summary

In the context of Bayesian learning where there is a prior distribution $\pi(\theta)$ and a sampling distribution, $f(x, \theta)$, it is well known that the prior mean of the posterior Shannon information: $\mathrm{E}_X(\mathrm{E}_{\theta|X} \log(\pi(\theta|X)) = \mathrm{E}_{X,\theta} \log(\pi(\theta))$ is not less than that of the prior: $\mathrm{E}_\theta \log \pi(\theta)$. This is a special case of a more general result that $\mathrm{E}_\theta g(\pi(\theta)) \leq \mathrm{E}_X(\mathrm{E}_{\theta|X} g(\pi(\theta|X))$ for a wider class of functions g, which includes Shannon information and all Renyi informations as a special cases. The paper gives a characterisation of this class. This leads naturally to an interpretation of learning which is equivalent to the posterior being more peaked than the prior in the generalized majorization sense arising from the theory of decreasing rearrangements of Hardy, Littlewood and Polya. A detailed study of the Beta distribution is included.

1 Introduction

A Bayesian approach to the optimal design of experiments uses some measure of pre-posterior risk, or information, to assess the efficacy the experimental design or more generally the choice of sampling distribution. Various versions of this approach have been developed by Blackwell [5] and the book of Torgeson [8] is a clear account. Renyi [4], Lindley [14] and Goel and Degroot [12] use information-theoretic approaches to measuring the value of an experiment, the approach which is adopted in this paper. A lucid review of many of the ideas is by Goel [11]. The paper by Chaloner and Verdinelli [10] gives a broad discussion of Bayesian design of experiments and Wynn and Sebastiani [13] discuss the Bayes information-theoretic approach. Goldman and Shaked [1] have a version of the main characterisation in this paper, in the discrete case, and this work has been followed up notably

by Fallis and Liddell [6]. This is evidence of a wider interest in these issues in cognitive science and epistemology.

When new data arrives it tends to *on the whole* to improve, or at least not worsen, our information about an unknown parameter θ. It is our purpose to show that in particular cases, that is conditional on particular data, the information can get worse, but that on the average it improves for a wide class of information functions. This class includes many special types of information, such as Shannon information, as special cases. It also turns out that whether or not one learns for particular observations, for all information functions in this class, is related to the theory of the continuous majorization of densities based on so-called decreasing rearrangements of densities.

We shall use the classical Bayes formulation with θ as an unknown parameter with a prior distribution $\pi(\theta)$ on a parameter space Θ and a sampling distribution $f(x|\theta)$ on a sample space $\mathcal{X}$. We denote by $f_{X,\theta}(x,\theta) = f(x|\theta)\pi(\theta)$ the joint distribution of X and θ and use $f_X(x)$ for the marginal distribution of X. The nature of expectations will be clear from the notation. To make the development straightforward we shall look at the case of distributions with densities. Unless stated otherwise $\mathcal{X}$ and θ can have arbitrary dimension.

The classical formulation proceeds as follows. Let U be a random variable with density $f_U(u)$. Let $g(\cdot)$ be a function on R^+ and define a measure of information for U with respect to g as

$$I_g(U) = E_U(g(f_U(U))$$

When $g(u) = \log(u)$, we have Shannon information. When $g(u) = \frac{u^\gamma - 1}{\gamma}$, $(\gamma > -1)$, we have a version of Renyi information which is sometimes called Tsallis information.

If X represents the observable random variable, we can measure the preposterior information of the experiment (query, etc) which generates a realization of X by the prior expectation of the posterior information which we define as

$$I_g(\theta; X) = \mathrm{E}_X E_{\theta|X}(g(\pi(\theta|X)) = E_{X,\theta}(g(\pi(\theta|X))$$

In the second term the inner expectation is with respect to the posterior (conditional) distribution of θ given namely $\pi(\theta|X)$ and the outside expectation is with respect to the marginal distribution of X. In the last term the expectation is with respect to the full joint distribution of X and π.

It is a main purpose of this paper to characterise the class of functions $g(\cdot)$ such that

$$I_g(\theta; X) \geq I_g(\theta)$$

for all joint distributions $f_{X,\theta}(x,\theta)$. The condition is that $h(u) = ug(u)$ is a convex function, on R^+. This includes Shannon information and Renyi informations as special cases. We start in the next subsection with some motivating counter-examples. Section 3 gives the proof of the main theorem. Section 4 shows the close connection with generalised majorization. Section 5 give a detailed development

for the Beta distribution as a vehicle for discussing when we have full learning or only partial learning, that is learning for some $g(\cdot)$ in the class but not all. We conclude with a short discussion.

2 Counter-examples

We show first that it is *not* true that information always increases. That is, it is not true that the posterior information is always more than the prior information:

$$I_g(\theta) \leq E_{\theta|X}(g(\pi(\theta|X))$$

A simple discrete example runs as follows. I have lost my keys. With high prior probability, p, I think they are on my desk. Suppose I have a uniform prior over all k likely other locations. However, suppose when I look on the desk my keys are not there. My posterior distribution is now uniform on the other locations. Under certain condition on p and k Shannon information has gone down. For fixed p, the condition is

$$k > k^* = \frac{(1-p)^{1-\frac{1}{p}}}{p} = e\left(\frac{1}{p} - \frac{1}{2} + O(p)\right)$$

When $p = \frac{1}{2}$, for example, $k^* = 4$ and $pk^* \to e, 1$ when $p \to 0, 1$. This phenomenon is captured by the somewhat self-doubting phrase "if my keys are not on my desk I don't know where they are". Note, however, that something has improved: the support size is reduced to from $k+1$ to k.

There is a simple way of obtaining a large class of examples, namely to arrange that there are x-values for which the posterior distribution is approximately uniform. Then, because the uniform distribution typically has low information, for such x we can have a decrease in information. Thus, we construct examples in which $f(x|\theta)\pi(\theta)$ happens to be approximately constant for some x. This motivates the following example.

Let $\Theta \times \mathcal{X} = [0,1]^2$ with joint distribution having support on $[0,1]^2$. Let $\pi(\theta)$ be the prior distribution and define a sampling distribution:

$$f(x|\theta) = a(\theta)(1-x) + \frac{x}{\pi(\theta)},$$

Note that we include the prior distribution into the sampling distribution as a constructive device, not as some strange new general principal. We have in mind, in giving this construction, that when $x \to 1$ the first term should approach zero and the second term, after multiplying by $\pi(\theta)$, should approach unity. Solving for $a(\theta)$ by setting $\int_0^1 f(x,\theta)dx = 1$ we have $a = \frac{2\pi(\theta)-1}{\pi(\theta)}$ so that

$$f(x|\theta) = \frac{(2\pi(\theta)-1)(1-x)+x}{\pi(\theta)}$$

The joint distribution is then

$$f(x|\theta)\pi(\theta) = (2\pi(\theta) - 1)(1 - x) + x. \tag{1}$$

The marginal distribution of X is $f_X(x) = 1$ on $[0, 1]$, since the integral of (1) is unity, so that (1) is also the posterior distribution $\pi(\theta|x)$. Note that, in order for (1) to be a proper density, we require that $\pi(\theta) \geq \frac{1}{2}$ for $0 \leq \theta \leq 1$.

The Shannon information of the prior is:

$$I_0 = \int_0^1 \pi(\theta) \log \pi(\theta) d\theta$$

and of the posterior is

$$I_1 = \int_0^1 (2\pi(\theta) - 1)(1 - x) + x) \log((2\pi(\theta) - 1)(1 - x) + x) d\theta$$

When $x = \frac{1}{2}$ the integrands of I_1 and I_0 are equal and $I_0 = I_1$. When $x = 1$ the integrand of I_1 is zero, as expected. Thus for a non-uniform prior we have less posterior information in a neighbourhood of $x = 1$, as we aimed to achieve.

Specialising to $\pi(\theta) = \frac{1}{2} + \theta$ on $[0, 1]$ gives

$$\begin{array}{rcl} I_0 & = & \frac{9}{8} \log 3 - \log 2 - 1/2 \\ I_1 & = & \frac{1}{4(1-x)}((2 - x)^2 \log(2 - x) - x^2 \log(x) + 2x) - 2 \end{array}$$

Information I_1 decreases from a maximum of $\log(2) - \frac{1}{2}$ at $x = 0$, through the value I_0 at $x = \frac{1}{2}$, to the value zero at $x = 0$. Thus $I_0 < I_1$ for $\frac{1}{2} < x \leq 1$. Since the marginal distribution of X is uniform on $[0, 1]$ we have the challenging fact that

$$\text{prob}_X\{I_1 < I_0\} = \frac{1}{2}.$$

Namely, with prior probability equal to one half there is less Shannon information in the posterior than the prior. The Renyi entropy exhibits the same phenomenon, but we omit the calculations.

2.1 The main result

Theorem 1 *For fixed $g(u)$, and the standard Bayesian set-up, the pre-posterior quantity $I_g(\theta, X)$ and posterior form $I_g(\theta)$ satisfy*

$$I_g(\theta; X) \geq I_g(\theta) = E_\theta(g(\pi(\theta)),$$

for all joint distributions $f_{X,\theta}(x, \theta)$ if and only if $h(u) = ug(u)$ is convex on R^+.

Proof. We give a version in which we assume that all densities exist, standard operation such as reversing integrals (Fubini) hold, conditional densities exist that $g(u)$ is suitably differentiable. We note again that the sample space, $\mathcal{X}$ and parameter space Θ are of arbitrary dimensions.

Assume that $ug(u)$ is convex. The posterior density for θ is

$$\pi(\theta|X) = \frac{f(x|\theta)\pi(\theta)}{f_X(x)}$$

Reversing the integration in the definition of $I_g(\theta; X)$ we have

$$\begin{aligned}
I_g(\theta; X) &= \mathrm{E}_\theta E_{X|\theta}(g(\pi(\theta|X)) \\
&= \int_\Theta \pi(\theta) \int_{\mathcal{X}} (g(\pi(\theta|X)) f(x|\theta) dx d\theta \\
&= \int_\Theta \int_{\mathcal{X}} (g(\pi(\theta|X)) \frac{f(x|\theta)\pi(\theta)}{f_X(x)} f_X(x) dx d\theta \\
&= \int_\Theta \int_{\mathcal{X}} g\left(\pi(\theta|x)\right) \pi(\theta|x) f_X(x) dx d\theta \\
&= \int_\Theta \mathrm{E}_X \left\{ g\left(\pi(\theta|x)\right) \pi(\theta|x) \right\} d\theta \\
&\geq \int_\Theta g\left(\mathrm{E}_X\{\pi(\theta|x)\}\right) \mathrm{E}_X\{\pi(\theta|x)\} d\theta \\
&= \int_\Theta g(\pi(\theta))\pi(\theta) d\theta = I_g(\theta)
\end{aligned}$$

The inequality is from Jensen's inequality using the convexity of $ug(u)$ applied to $\pi(\theta|X)$ considered as a random variable, for fixed θ.

The reverse statement is proved by constructing a special class of joint distribution assuming, assuming $I_g(\theta, X) \geq I_g(\theta)$, taking some limits and forcing $h(u) = ug(u)$ to be convex. We give a simple version of the result in which $h(u)$ is twice continuously differentiable so it is enough to prove that $h''(u) \geq 0$ for all $u > 0$.

Let $\mathcal{X} = \Theta = (0, 1]$. Define the piece-wise uniform joint density which has special values in the four cells of $(0, 1]^2$ defined by the lines $x = 1-\alpha$ and $\theta = 1-\beta$:

$$f(x, \theta) = \begin{cases} f_{11}, & 0 < x \leq 1-\alpha,\ 0 < \theta \leq 1-\beta \\ f_{21}, & 1-\alpha < x \leq 1,\ 0 < \theta \leq 1-\beta \\ f_{12}, & 0 < x \leq 1-\alpha,\ 1-\beta < \theta \leq 1 \\ f_{22}, & 1-\alpha < x \leq 1,\ 1-\beta < \theta \leq 1 \end{cases}$$

For convenience we add the additional condition that $f_X(x) = 1$ for x in $\mathcal{X} = (0, 1]$. Using this condition, the statement $I_g(\theta, X) \geq I_g(\theta)$ becomes:

$$\begin{aligned}
(1-\alpha)(1-\beta)g(f_{11})f_{11} + \alpha(1-\beta)g(f_{21})f_{21} + (1-\alpha)\beta g(f_{12})f_{12} + \alpha\beta g(f_{22})f_{22} \\
\geq (1-\beta)((1-\alpha)f_{11} + \alpha f_{21})g((1-\alpha)f_{11} + \alpha f_{21}) \\
+\beta((1-\alpha)f_{12} + \alpha f_{22})g((1-\alpha)f_{12} + \alpha f_{22}) \quad (2)
\end{aligned}$$

The condition $f_X(x) = 1$ has two parts:

$$\begin{aligned}
(1-\beta)f_{11} + \beta f_{12} &= 1, \quad (0 < x \leq 1-\alpha) \quad (3) \\
(1-\beta)f_{21} + \beta f_{22} &= 1, \quad (1-\alpha < x \leq 1) \quad (4)
\end{aligned}$$

Fix $f_{11} = u > 0$ and $0 \leq \alpha < 1$. It is important to note that this can be done arbitrarily, because we may always adjust the other constants so that (3) and (4) are not violated. Set $h(u) = ug(u)$ and assume that $h''(u) \neq 0$. Then set

$$f_{11} = u, \quad f_{21} = u + \delta, \quad f_{12} = v, \quad f_{22} = v + \delta^2$$

Inequality (2) becomes

$$\begin{aligned}(1-\alpha)(1-\beta)h(u) + \alpha(1-\beta)h(u+\delta) + (1-\alpha)\beta h(v) + \alpha\beta h(v+\delta^2) \\ \geq (1-\beta)h(u+\alpha\delta) + \beta h(v+\alpha\delta^2), \qquad (5)\end{aligned}$$

Also, (2) and (3) become, respectively,

$$\begin{aligned}(1-\beta)u + \beta v &= 1 \qquad (6) \\ (1-\beta)(u+\delta) + \beta(v+\delta^2) &= 1, \qquad (7)\end{aligned}$$

Solving for β and v from (6) and (7), substituting in (5) and then carrying out a Taylor expansion in δ, we find that the terms in $h(u), h'(u), h(1), h'(1)$ and $h''(1)$ cancel and we obtain

$$-\frac{\alpha(1-\alpha)}{2}\delta^3 h''(u) + \mathrm{o}(\delta^3) \geq 0$$

Letting δ tend to zero from below gives $h''(u) \geq 0$ as required.

One way of considering the second half of the proof is that if $h(u) = ug(u)$ is not convex at a particular u we may construct a counter-example in which the expected g-information decreases. Note that the condition $h(u) = ug(u)$ convex on R^+ is equivalent to $g(\frac{1}{u})$ being convex which is referred to as $g(u)$ being "reciprocally convex" by Goldman and Shaked [1]. There is also a connection, which we do not pursue, with "proper scoring rules". These are discussed by Grünwald and Dawid [7] in the context of maximum entropy priors and related topics, rather than learning directly.

3 A majorization interpretation

The analysis of the last section leads to a sense in which for two distributions $\pi_1(\theta)$ and $\pi_2(\theta)$ the first is more peaked than the second if and only

$$\int_\Theta h(\pi_1(\theta))d\theta \leq \int_\Theta h(\pi_2(\theta))d\theta \ \text{ for all convex } \ h(u) = ug(u) \ \text{ on } \ R^+ \qquad (8)$$

For Bayesian learning we may *hope* that the ordering holds when π_1 is the prior distribution and π_2 the posterior distribution. We have seen from the counter-examples that it does not hold in general but always holds in expectation, by Theorem 1. It is natural, therefore, to try to understand the partial ordering and

it is the second main purpose of this paper to reveal that the ordering is equivalent a well-known majorization ordering for distributions.

Consider two discrete distributions with n-vectors of probabilities $\pi_1 = (\pi_1^{(1)}, \ldots, \pi_n^{(1)})$ and $\pi_2 = (\pi_1^{(2)}, \ldots, \pi_n^{(2)})$ where $\sum_i \pi_i^{(1)} = \sum_i \pi_i^{(1)} = 1$. First, order the probabilities:

$$\tilde{\pi}_1^{(1)} \geq \ldots \geq \tilde{\pi}_n^{(1)}, \;\; \tilde{\pi}_1^{(2)} \geq \ldots \geq \tilde{\pi}_n^{(2)}$$

Then π_2 is said to majorizes π_1, written $\pi_1 \preceq \pi_2$ when

$$\sum_{i=1}^{j} \tilde{\pi}_i^{(1)} \leq \sum_{i=1}^{j} \tilde{\pi}_i^{(2)}$$

for $j = 1, \ldots, n$ (equality for $j = n$). The standard reference is Marshall and Olkin [2] where one can find several equivalent conditions. Two of the best known are:

A1. there is a doubly stochastic matrix $P_{n\times n}$ such that

$$\pi_1 = P\pi_2$$

A2. $\sum_i^n h(\pi_i^{(1)}) \leq \sum_i^n h(\pi_i^{(1)})$ for all continuous convex functions $h(x)$.

Condition A2 shows that, in the discrete case, our partial ordering is equivalent to majorization of the raw probabilities.

We now extend this to the continuous case. This generalisation, which we shall also call $\preceq$, to save notation, has a long history and the area is historically referred to as the theory of the "rearrangements of functions" to respect the terminology of Hardy, Littlewood and Polya [9]. It is particularly well suited to probability density functions because of the fact that the normalisation, which plays a part, is built in via $\int \pi_1(\theta)d\theta = \int \pi_2 d\theta = 1$. The natural analogue of the ordered values in the discrete case is that every density π has a unique density $\tilde{\pi}$, called a "decreasing rearrangement", obtained by a reordering of the probability mass to be non-increasing, by direct analogy with the discrete above. In the theory π and $\tilde{\pi}$ are then referred as being *equimeasurable*, in the sense that the supports are transformed in a measure-preserving way.

There are a short sections on the topic in Marshall and Olkin [2] and in Müller and Stoyan [3]. A key paper in the development is Ryff [15]. The next paragraph is a brief summary.

Definition 2 *Let $\pi(z)$ be a density and define $m(y) = \mu\{z : \pi(z) \geq y\}$. Then $\tilde{\pi}(y) = sup\{t : m(y) > t\}$, $y > 0$ is the decreasing rearrangement of $\pi(z)$.*

Definition 3 *We say that π_2 majorizes π_1, written $\pi_1 \preceq \pi_2$ if and only if*

$$\int_0^c \tilde{\pi}(z)dz \leq \int_0^c \tilde{\pi}_2(z)z$$

for all $c > 0$.

Define a doubly stochastic kernel $P(x, y) \geq 0$ on $(0, \infty)$, that is

$$\int_x P(x, y) = \int_y P(x, y) = 1$$

There is a list of key equivalent conditions to $\preceq$ which are the continuous counterparts of the discrete majorization conditions. The first two generalize A1 and A3 above.

B1. $\pi_1(\theta) = \int_\Theta P(\theta, z)\pi_2(z)dz$ for some non-negative doubly stochastic kernel $P(x, y)$.

B2. $\int_\Theta h(\pi_1(z))dz \leq \int_\Theta h(\pi_2(z))dz$ for all continuous and convex function h

B3. $\int_\Theta(\pi_1(z) - a)_+dz \leq \int_\Theta(\pi_z(z) - a)_+dz$ for all $a > 0$

Condition B1 is the key to our discussion, for it shows that in the univariate case, if we assume that $h(u) = ug(u)$ is continuous and convex our partial ordering, namely that condition (8) holds for all convex h, is equivalent to

$$\pi_1(\theta) \preceq \pi_2(\theta),$$

in the majorization sense. From now on just use $\preceq$ to refer to the ordering.

From Definition 3 we see that $\preceq$ is equivalent to standard first order stochastic dominance of the decreasing rearrangements, since $\tilde{F}(\theta) = \int_0^\theta \tilde{\pi}(z)dz$ is the cdf corresponding to $\tilde{\pi}(\theta)$. We may write this

$$\tilde{\pi}_1(\theta) \preceq_{st} \tilde{\pi}_2(\theta),$$

and we may also write down of a number of equivalent conditions.

We note that condition B3 is useful in examples. It says that the probability mass under the density but above a "slice" at height a is more for π_2 than for π_1. It can also be used in a slightly different way: we may integrate vertically the total horizontal "length" of the density at height a. We refer to condition B3 as the "slicing condition", for ease of recall.

Remark. The classic theory of rearrangements is for univariate distributions, whereas, as stated, we are interested in θ of arbitrary dimension. In the present paper we will simply make the claim that the interpretation of our partial ordering in terms of decreasing rearrangements can indeed be extended to the multivariate case. Heuristically this is done as follows. For a multivariate distribution we may create a univariate rearrangement by considering a decreasing threshold and "squashing" all the multivariate mass for which the density is above the threshold to univariate mass adjacent to the origin. Since we are transforming multivariate volume to area, care is needed with Jacobians. We can then use the univariate development above. It is instructive to consider the univariate decreasing rearrangement of the multivariate normal distribution, but we omit the computations here.

4 The Beta distribution

When the distribution $\pi(z)$ is symmetric and unimodal it is straightforward to construct the rearranged density, by simply shifting the mode to the origin and taking twice the right half of the density. Thus, for the normal $N(\mu, \sigma^2)$ density ϕ_{μ,σ^2}, the decreasing rearrangement is $\tilde{\phi}(z) = 2\phi_{(0,\sigma^2)}, x \geq 0$. Clearly we have the ordering $\phi_{\mu_1,\sigma_1^2} \preceq \phi_{\mu_2,\sigma_2^2}$ if and only if $\sigma_2 \leq \sigma_1$. Moreover, under the usual Bayes set up with $f(x,\theta) \sim N(\theta, \sigma^2)$, $\pi(\theta) \sim N(\nu, \tau^2)$, and known σ, ν and τ, the posterior density always dominates the prior with respect $\preceq$ because the posterior variance cannot be less than the prior variance. There is a similar statement for the multivariate normal distribution with the same dimension. With obvious notion: $\phi_{\mu_1,\Sigma_1} \preceq \phi_{\mu_2,\Sigma_2} \Leftrightarrow |\Sigma_2| \leq |\Sigma_1|$. We recall that, except for constants, $\log|\sigma|$ is the Shannon information.

For non-symmetric distributions the construction of the $\tilde{\pi}$ may be quite hard analytically. To illustrate this we discuss the case of the Beta(a, b) distribution

$$\pi_{a,b}(z) = \frac{(1-z)^{\alpha-1} z^{b-1}}{B(a,b)}$$

Our task is to find a characterisation of our ordering in terms of (a, b). We show here how to do numerical computations after first demonstrating the complexity of the problem.

Let us try to find the rearrangement $\tilde{\pi}_{2,3}(z)$. Since $\pi_{2,3}(z) = 12(1-z)z^2$ is not symmetric we have to first intersect it with a horizontal line, say at height a and examine the solutions $z_1 \leq z_2$ to $\tilde{\pi}(z) = a$ while the considering the argument $z = z_2 - z_1$. One way to resolve this is to use elimination. Thus set up the equations:

$$\pi(z_1) = c, \;\; \pi(z_2) = c, \;\; z_2 - z_1 = z,$$

giving

$$12z_1(1-z_1)^2 = c; \;\; 12z_2(1-z_2)^2 = c, \;\; z_2 - z_1 = z.$$

Eliminating z_1 and z_2 (using computer algebra methods) gives the implicit equation for (z, c):

$$48z^6 - 96z^4 + 48z^2 + 9c^2 - 16c = 0$$

The "bow-tie" curve in Figure 1 is the locus of this equation and incorporates the decreasing rearrangement $\tilde{\pi}_{2,3}$ as the decreasing section supported on $[0, 1]$ to the right hand half of the "knot" of the tie.

It is also possible to obtain an implicit algebraic equation in the (a, b) integer case for cdf corresponding to the $\pi_{a,b}$, $\tilde{F}_{a,b}$, from which we can laboriously check the conditions in Definition 3, for integer cases.

We give a nice property of the Beta distribution when both parameters are greater than 1, even when the distributions are not symmetric, two different Beta cross twice the *at the same level*.

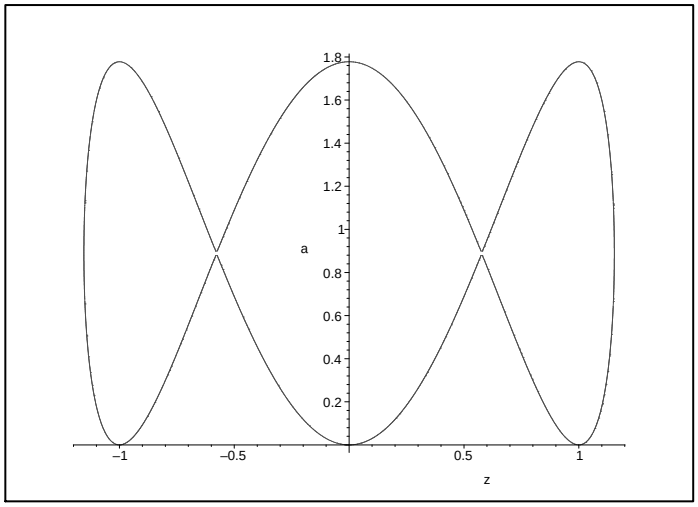

Figure 1: Decreasing rearrangement of the Beta(2,3) distribution

Lemma 4 *Let two different Beta distributions π_{a_1,b_1} and $\pi_{(a_2,b_2)}$, with all parameters be larger than 1, have the same mode so that $\frac{a_1-1}{a_1+b_1-2} = \frac{a_2-1}{a_2+b_2-2}$. Then $\pi_{a_1,b_1}(z) = \pi_{a_2,b_2}(z)$ has two solutions $z_1 < z_2$ with the property that*

$$\pi_{a_1,b_1}(z_1) = \pi_{a_2,b_2}(z_1) = \pi_{a_1,b_1}(z_2) = \pi_{a_2,b_2}(z_2)$$

Proof. The equation $\pi_{a_1,b_1}(z) = \pi_{a_2,b_2}(z)$ together with the equality of modes, gives, after a little algebra:

$$z^{a_1-1}(1-z)^{b_1-1} = \left(\frac{B(a,b)}{B(a_2,b_2)}\right)^{\frac{a_1-1}{a_1-a_2}}$$

This shows that whatever the solution the value of the left hand side namely, $z^{a_1-1}(1-z)^{b_1-1}$ is the same. But since this form appears in the density $\pi_{a_1,b_1}(z_1)$, both solutions have the same value of the density.

Lemma 5 *Let two Beta distributions $\pi_{(a_1,b_1)}$ and $\pi_{(a_2,b_2)}$ have the same mode. Then*

$$\pi_{(a_1,b_1)} \preceq \pi_{(a_2,b_2)}$$

if and only if

$$\max_{z\in[0,1]} \pi_{(a_1,b_1)} \leq \max_{z\in[0,1]} \pi_{(a_2,b_2)}$$

Proof. This is a simple application of the slicing condition B3, above. We only the consider the case when all parameters are greater than 1. Since, by Lemma 4, the densities cross exactly twice at the same level, the difference in the width of the densities as a function of a is increasing. This forces the difference in the cumulative slices to have a single sign change and since for $a = 0$ the integral of this difference is zero, condition B3 holds.

The full characterization of the ordering in terms of (a, b) does not have a closed form but we can go a considerable way to understanding the ordering and computing its "contours". We carry out some calculations for the case $(a, b) = (2, 3)$. The faint curved lines in Figure 2 show equi-information contours, passing through $(2, 3)$, for the Renyi entropies $h(u) = -\log(u) \quad (\gamma \to -1)$ and $m = \max_z \pi(z), \quad (\gamma \to \infty)$. The bold lines are the for our ordering and we now review how they were calculated.

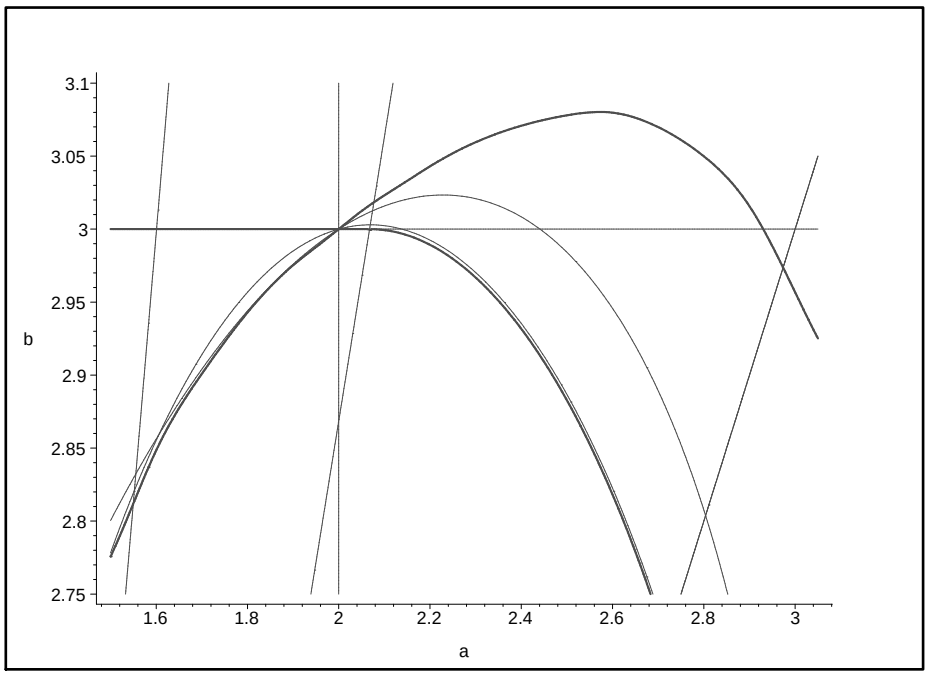

Figure 2: Information contours and the $\preceq$ ordering : $(a, b) = (2, 3)$.

First, consider the general case $a \leq b$. Locally to a given $a_0 \leq b_0$, the set of points $\preceq$-dominating (a_0, b_0) is that lying above two curves (bold in Figure 2) which cross at (a_0, b_0). The region below both lines is the set of (a, b) *dominated by* (a_0, b_0). The slopes of the curves, at fixed (a, b) are given by the supremum and infimum of over all convex $h(u)$ of the quantity

$$-\frac{\frac{\partial}{\partial a}\int_0^1 h(\pi(z))dz}{\frac{\partial}{\partial b}\int_0^1 h(\pi(z))dz} \tag{9}$$

Assuming h is differentiable this becomes

$$-\frac{\frac{\partial}{\partial a}\int_0^1 h'(\pi(z))\frac{\partial \pi(z)}{\partial a}dz}{\frac{\partial}{\partial b}\int_0^1 h'(\pi(z))\frac{\partial \pi(z)}{\partial b}dz}, \tag{10}$$

Now, since h is convex, h' is nondecreasing. Moreover, since $\int_0^1 \pi(z)dz = 0$, we have

$$\int_0^1 \frac{\partial}{\partial a}\pi(z)dz = \int_0^1 \frac{\partial}{\partial b}\pi(z)dz = 0,$$

which means that without loss of generality we can take h' to be non-negative. This implies that h', can be expressed up to scalar multiplier (which cancels in the ratio (10) above) as a non-negative convex mixture of indicator functions $I_{[0,\infty)}$. After simple geometric considerations, the problem of finding the supremum and infimum of (11) then reduces to finding when a line through the origin touches the convex hull of the points

$$\left(\int_{\pi\geq c}\frac{\partial\pi(z)}{\partial a}dz, \int_{\pi\geq c}\frac{\partial\pi(z)}{\partial a}dz\right).$$

The conclusion is that this will occur at one of the extreme points and so it is enough to compute the supremum and infimum over $c \geq 0$ of

$$-\frac{\int_{\pi\geq c}\frac{\partial\pi(z)}{\partial a}dz}{\int_{\pi\geq c}\frac{\partial\pi(z)}{\partial a}dz} = -\frac{\int_{z_1}^{z_2}\frac{\partial\pi(z)}{\partial a}dz}{\int_{z_1}^{z_2}\frac{\partial\pi(z)}{\partial a}dz}, \tag{11}$$

under the constraint on $\{z_1, z_2\}$, namely $\pi(z_1) = \pi(z_2) = c$.

The following useful transformation captures this constraint

$$u = \frac{z_1}{z_2}, \quad v = \frac{1-z_1}{1-z_2}, \quad v = u^{\frac{1-a}{b-1}}.$$

Using computer algebra we are able to derive a closed form for (12) as a function of $u \in [0,1]$, for small integer (a,b), and work numerically for non-integer values. There are several different situations, depending on whether the maximum and minimum of (12) occurs at an end point or internal point of $[0,1]$. The value at $u = 0$ is typically zero. The value $u = 1$ is corresponds to using the maximum:

$$m(a,b) = \max_{[0,1]}\pi(z) = \frac{1}{B(a,b)}\left(\frac{b-1}{a+b-2}\right)^{a-1}\left(\frac{a-1}{a+b-2}\right)^{b-1},$$

which, to recall, is the limiting value of Renyi entropy as $\gamma \to \infty$. Realizing this, the required value of (12) at $u = 1$ can then be computed explicitly as:

$$-\frac{\frac{\partial m}{\partial a}}{\frac{\partial m}{\partial m}} = -\frac{\log\left(\frac{a-1}{a+b-2}\right) + \Psi(a+b) - \Psi(a)}{\log\left(\frac{b-1}{a+b-2}\right) + \Psi(a+b) - \Psi(b)}, \tag{12}$$

where Ψ is the digamma function. At internal points of $[0,1]$, we need to compute the values numerically, which turns out to be a quite delicate calculation.

The role of m, the maximum, is critical. Our first conjecture was that for $a_1 \leq b_1,\ a_2 \leq b_2$ it was necessary and sufficient that $b_1 \leq b_2$ and $m(a_1,b_1) \leq m(a_2,b_2)$. But this turns out to be true only for restricted values of the parameters. The true situation is more complex. For $a \leq b$, in addition to the line of symmetry $a = b$ there are two special curves. These are drawn as rather steep, almost straight, curves in Figure 2. The one closest to $a = b$ corresponds to when the formula (13) is zero, namely

$$\log\left(\frac{a-1}{a+b-2}\right) + \Psi(a+b) - \Psi(a) = 0 \tag{13}$$

As $a, b \to \infty$ this tends to the line $b = 2a$. The second special curve (on the left) is the locus of points at which the maximum of (12) occurs at $u = 1$, that is to say no longer at an interior point. Finding this locus is harder. It is obtained by checking the second derivative of (12) at $u = 1$. The locus is given by:

$$\begin{array}{l} -(-\log(M) + \Psi(a))(a-1)(a-2b+1) \\ -(-\log(1-M) + \Psi(b))(b-1)(-b+2a-1) \ , \\ +\Psi(a+b)(a+b-2)(a-b) = 0 \end{array} \tag{14}$$

where $M = \frac{a-1}{a+b-2}$ is the mode. Intriguingly, this special curve tends to the line $b = ra$ where $r = \frac{4}{9-\sqrt{65}} = 4.26556$. There is a rule that to the left of the line (15), namely for quite large values of b relative to a the conjecture mentioned above holds and the upper (right) boundary of the dominating region is given by the contour of the maximum, m. A second rule is that to the right of the first line (14) m gives the lower, not the upper, (right) boundary, and corresponds to where the lower boundary starts to have negative slope. To roughly summarize: the maximum m starts as an upper (right) boundary and finishes as a lower (right) boundary with a transitional region between the two special lines.

Consider the point $(2, 3)$. The upper right boundary of $\preceq$ is given by the maximum of (12), as function of u, at a point internal to $[0, 1]$ and the maximum value is $0.24372\ldots$. The minimum is at $u = 0$ with value zero. Locally, then the upper and lower boundaries are respectively:

$$b - 3 \simeq 0.24372(a - 2), \quad b = 3$$

Proceeding along the line $b = 3$ in a positive direction $(2, 3)$ we quickly meet a tangent curve at the point $(2.06758\ldots, 3)$, where $b = 3$ meets line (13). The lower boundary which falls away to the right from this point and is the equi-information trajectory of for m.

As a check, consider the version of Renyi information as $\gamma \to -1$ which is equivalent to taking $h(u) = -\log u$ (the upper faint curve). The slope of the information contour in that case at $(2, 3)$ is

$$-\frac{\Psi(a) - \Psi(a+b) + 1}{\Psi(b) - \Psi(a+b) + 1},$$

where Ψ is the digamma function, and it turns out that at $(2, 3)$ this is the largest slope among all Renyi informations. The value at $(2, 3)$ is $\frac{1}{5} < .24372\ldots$ confirming that the our ordering is stronger.

To the right of the special curve (15) the upper (right) boundary is given by the lower envelope of the tangents to iso-information contours for typically *different* $h(u)$ for each (a, b) on the contour. Each such tangent comes from a primitive piece-wise linear $h(u)$, with a single non-constant piece. The envelope is concave and we as move along it for increasing a the value of b increases, until it reaches a maximum and then decreases towards the line $a = b$. The computation of the envelope is a somewhat delicate exercise in solving a differential equation whose

differentials are themselves only expressed numerically. Here we have used a rough cubic spline approximation.

We have only discussed the case $a \leq b$. The full diagram, is symmetric about the line $a = b$, and part of this line is also drawn in Figure 2. Note that the upper boundary crosses this line at $a = b \simeq 2.9$ a value larger than for any Renyi information, again reinforcing that $\preceq$ is a stricter condition.

Drawing on this situation we define *strong learning* as the situation in which $\pi(\theta) \preceq \pi(\theta|X)$. When (8) holds for some for $h(u) = ug(u)$ but for not for others, we can call the region the *partial learning* region. For $(a, b$

$)= (2,3) the region above both bold lines is the strong learning region and between the selines (on both si$

5 Discussion

The paper could be taken as a critique of Bayesian methods. The argument would go as follows. There can be situations where the data is radically at odds with the prior distribution to the extent that *even under the Bayesian analysis* the posterior information is less that the prior. One may go on to argue that a strong prior may lead to the conduct of an inappropriate experiment, as in the lost keys example; one can be led to look in the wrong place. However such a strong critique is not the intention. A counter argument would be that, without some idea of where to look, looking may be vastly inefficient. We may have to methodically search every location.

The purpose has been to understand what learning means and the Bayesian information theoretic approach gives pleasantly intuitive answers. The Beta case shows clearly when, for a particular situation, that is a for particular posterior, one has strong learning and weak learning. One can go on to look at many other standard cases both univariate and multi-variate: two-parameter Gamma, Dirichelet, Gaussian with priors on variances and so on. For a given parameter values each will have its separation into strong and partial learning regions. The challenge is the full characterization of $\preceq$, in terms of parameters. If this cannot be done in closed form then computational methods, similar to the above, may be developed.

This author is particularly interested in the implications for Bayesian experimental design which can be seen as allowing some choice, typically via setting the levels of independent variables, of the sampling distribution $f(x, \theta)$. One may investigate designs in which the expected information is largest for one or more g, which is the classical approach. But also there is hope of finding designs for which the posterior distribution is always more peaked than the prior in terms of a particular g in our class or even for all g in our class. This is easy if the prior has minimal information since then typically we always have strong learning. It is more of a challenge if the prior is informative.

References

[1] Goldman A. and Shaked M., *Goldman on probabilistic inference*, Philosophical studies **63** (1991), 31–55.

[2] Marshall A. and Olkin I, *Inequalities: Theory of majorization and its applications*, Academic Press, New York, 1979.

[3] Müller A. and Stoyan D., *Comparison methods for stochastic models and risk*, Wiley, New York, 2002.

[4] Renyi A., *On a measure of entropy and information provided by an experiment*, Proc. 4th Berkeley Symp. Math. Statist. Prob. **1** (1961), 547–561.

[5] Blackwell D., *Comparison of experiments*, Proc. Second Berkeley Symp. on Math. Statist. Prob. (1951), 93–102.

[6] Fallis D., *Goldman on probabilistic inference*, Philosophical studies **109** (2002), 223–240.

[7] Grünwald P. D. and Dawid A. P., *Game theory, maximum entropy, minimum discrepancy and robust bayesian decision theory*, Ann, Statist. **21** (2004), 1367–1433.

[8] Torgersen E., *Comparison of statistical experiments*, Cambridge University Press, Cambridge, 1991.

[9] Littlewood J. E. Hardy G.H. and Polya G., *Inequalities*, Cambridge University Press, Cambridge, 1952.

[10] Chaloner K. and Verdinelli I., *Bayesian experimental design: a review*, Statistical Science **10** (1995), 273–304.

[11] Goel P. K. and Josep G., *When one experiment is 'always better' than another*, The Statistician **52** (2003), 515–537.

[12] Goel P. K. and Degroot M. H., *Comparison of experiments and information measures*, Ann. Statist. **7** (1979), 1066–1077.

[13] Sebastiani P. and Wynn H.P., *Maximum entropy sampling and bayesian optimal experimental design*, J. Roy. Statist. Soc. B **62** (2000), 145–157.

[14] Lindley D. V., *On a measure of information provided in an experiment*, Ann. Math. Statist. **27** (1956), 986–1005.

[15] Ryff J. V., *Orbits of l^1-functions under doubly stochastic transformations*, Trans. Amer. Math. Soc. **117** (1965), 92–100.