Newton Institute

Information-based method in dynamic learning

Henry Wynn

London School of Economics
h.wynn@lse.ac.uk

22 July, 2011

Contents

Discussion of learning in a Bayes context

Bayesian learning: we expect to learn

Standard Bayes model:
Prior distribution: $\pi(\theta)$ Sampling distribution: $f(x, \theta)$

Prior expectation 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:

$$\mathrm{E}_\theta \log \pi(\theta)$$

More general result

$$\mathrm{E}_\theta(g(\pi(\theta))) \leq \mathrm{E}_X(\mathrm{E}_{\theta|X} g(\pi(\theta|X))$$

for a wider class of functions g, which includes Shannon information and all Renyi informations as a special cases.

Information

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

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

- Shannon: $g(u) = \log(u)$ Shannon information.
- Renyi/Tsallis: $g(u) = \frac{u^\gamma - 1}{\gamma}$, $(\gamma > -1)$

Define the preposterior information of the experiment or the *prior expectation of the posterior information* which we define as

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

The class of information functions for which we expect to learn

Theorem

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^+.

Note: equivalently $g\left(\frac{1}{u}\right)$ is convex.

Proof: "if"

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

"only if"

- Construct a special class of joint distribution assuming, assuming $I_g(\theta, X) \geq I_g(\theta)$
- Taking some limits and force $h(u) = ug(u)$ to be convex
- It is easiest to prove a 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 $u > 0$.

It is *not* true that information always increases, namely

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

Counterexample: 1

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 $pk^* \longrightarrow e, 1$ when $p \longrightarrow 0, 1$.

"If my keys are not on my desk I don't know where they are"

Counterexample 2

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:

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

Solving for a:

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

The joint distribution is

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

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

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

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

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

With prior probability equal to one half there is less Shannon information in the posterior than the prior

Majorization: or the theory of "rearrangments"

The analysis motivates the definition of a partial ordering

Definition

Define

$$\pi_1(\theta) \prec \pi_2(\theta)$$

if and only if

$$\int_\Theta h(\pi_1(\theta))d\theta \leq \int_\Theta h(\pi_2(\theta))d\theta \;\; \textit{for all convex} \;\; h(u) = ug(u) \qquad (1)$$

For Bayesian learning we may *hope* that the ordering holds when π_1 is the prior distribution and π_2 the posterior distribution.

The ordering is equivalent to majorization ordering for distributions

Discrete majorization

- 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 $\sum_i \pi_i^{(1)} = \sum_i \pi_i^{(1)} = 1$.
- Order the probabilities:

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

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

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

Equivalent conditions

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 $h(x)$.

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

Continous majorization: theory of rearrangments

Definition

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

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 $c > 0$.

First order stochastic dominance of the decreasing rearrangements

Equivalent conditions

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. $\int_\Theta (\pi_1(z) - c)_+ dz \leq \int_\Theta (\pi_z(z) - c)_+ dz$ for all $c > 0$. ("Slice condition").

Condition B2 shows that (in the univariate case) our partial ordering is equivalent to

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

in the majorization sense. (From now on just use $\preceq$.)

Condition B3 is useful in examples: the probability mass under the density but above a "slice" at height c is more for π_2 than for π_1.

Examples: Gaussian

- Univariate Gaussian:

$$\phi_{(\mu_1,\sigma_2^2)} \preceq \phi_{(\mu_2,\sigma_1^2)} \Leftrightarrow \sigma_2 \leq \sigma_1$$

- Multivariate Gaussian

$$\phi_{(\mu_1,\Sigma_1)} \preceq \phi_{(\mu_2,\Sigma_2)} \Leftrightarrow |\Sigma_2| \leq |\Sigma_1|$$

We recall that, except for constants, $-\log|\Sigma|$ is the Shannon information.

Beta distribution

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

Problem: describe $\preceq$ in terms of a partial ordering on (a, b)

Theorem

If $\pi_1 = \pi_{a_1,b_1}(z)$ *and* $\pi_2 = \pi_{a_2,b_2}(z)$ *have the same mode (or reflect through* $\frac{1}{2}$*) then*

$$\pi_1 \preceq \pi_2 \Leftrightarrow \max_{z\in[0,1]} \pi_1 \leq \max_{z\in[0,1]} \pi_2$$

Lemma

When $a_1, b_1, a_2, b_2 > 1$ *and* π_1 *and* π_2 *have the sane mode they intersect twice* at the same level.

Beta with small integer a, b: exact computation

Example: $(a, b) = (2, 3)$

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 c and examine the solutions $z_1 \leq z_2$ to $\tilde{\pi}(z) = c$ while the considering the argument $z = z_2 - z_1$:

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

giving

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

Eliminating z_1 and z_2 gives the implicit equation for (z, c):

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

The "bow-tie"

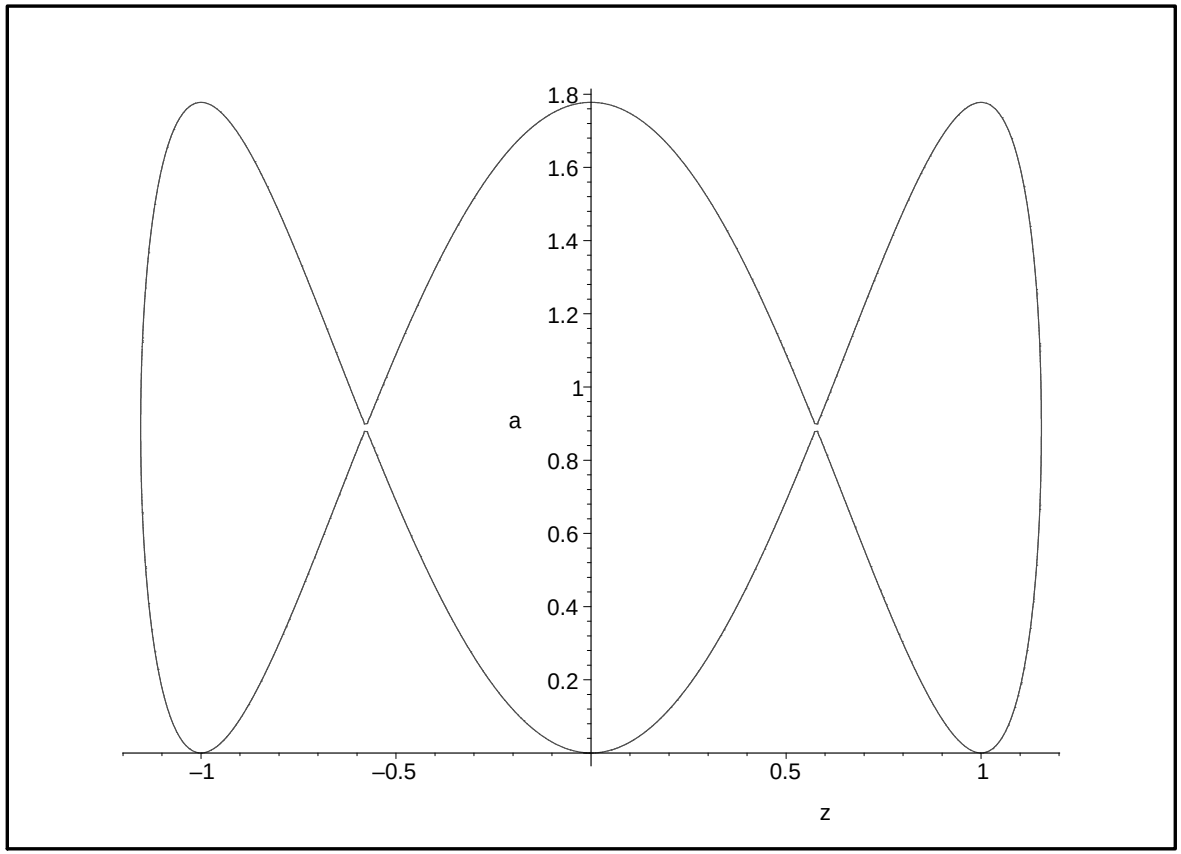


Figure: Decreasing rearrangement of the Beta(2,3) distribution

Check if $\tilde{F}_1 = \tilde{F}_2$ has any solutions:

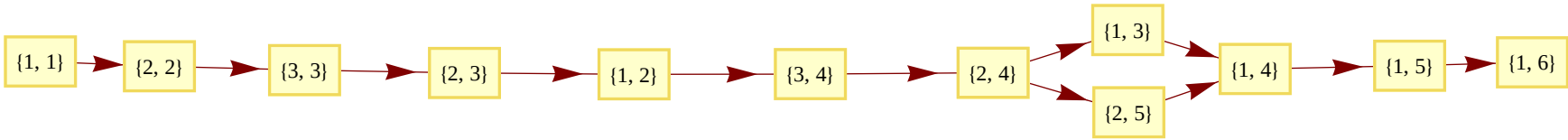
$$\tilde{F}_i = \int_{z_1}^{z_2} \pi_i(z)dz, \quad i = 1, 2$$

$$z = z_2 - z_1$$

$$T = F_2 - F1$$

1. Eliminate z_1, z_2 to find polynomial equation for $T = 0$
2. Check if $T = 0$ has any solutions in $[0, 1]$
3. If no solutions check the $\max_z \pi_i(z)$ to see which dominates

Results for small a, b



Smallest incomparable (a, b) are

$$(1, 3), \quad (2, 5)$$

	(1,3)	(2,5)
Shannon (I)	0.43	0.484
max (M)	3.0	0.257

Shannon with max: not enough for $\preceq$

It is not true that $\pi_1 \preceq \pi_2$ is equivalent to Shannon and max together $I_1 \leq I_2$ and $M_1 \leq M_2$

Counter example:

	(3,4)	(4,4)
I	2.073 ...	2.187...
M	0.344 ...	0.384 ...
I_h	0.226...	0.207...

,

where $h(z) = 1 - \frac{z}{2}$

Constructing the partial ordering

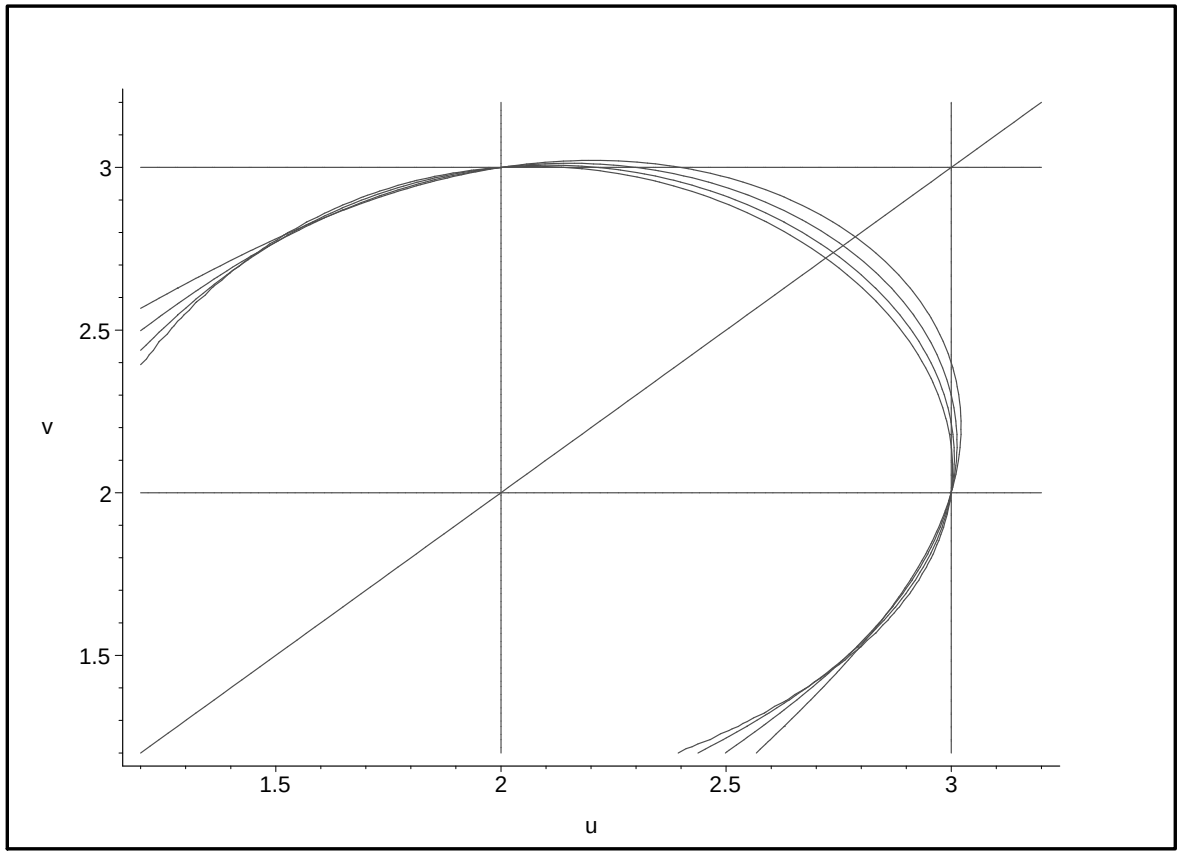


Figure: Equi-information lines from the Beta(2,3) distribution

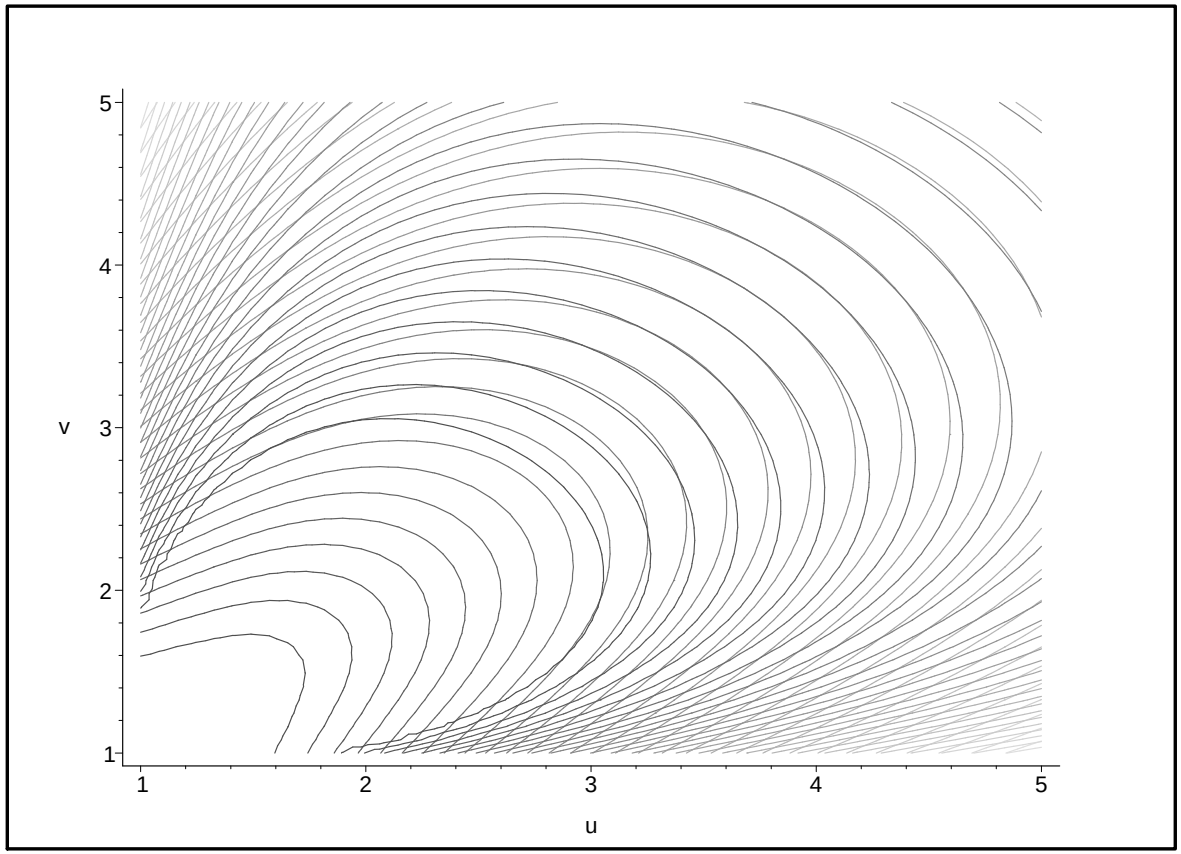


Figure: More equi-information lines distribution

Strong and partial Learning

Definition

We have strong learning if $\pi(\theta) \prec \pi(\theta|X)$, otherwise partial learning. In the latter case we may have learning wrt some information function but not others

The counterexamples are cases of partial learning. Even for the "keys" example we have partial learning in that the support size decreases (exercise: which information function gives this?).

Theorem

In the Beta-Binomial case given a Beta prior $\pi_{a,b}(\theta)$ and and $X \sim$ Binomial (n, θ) there is a constant n^* such that we have strong learning for any data X if $n \geq n^*$.

The computational method for $\preceq$

Select one of the definition of $\preceq$ and compute the upper and lower envelopes

Take the "slice" condition: B3. Let $p_{a,b}(c)$ be the slice probability. fix a. Upper/lower envelope of curves through (a_0, b_0), $a_0 < b_0$:

$$\overline{b}(a) = \sup_c \{b : p_{a,b}(c) = p_{a_0,b_0}(c)\}$$

$$\underline{b}(a) = \inf_c \{b : p_{a,b}(c) = p_{a_0,b_0}(c)\}$$

A cheap result for max

Or we can pick our favourite learning function and ask how big does n have to be to guarantee partial learning wrt to that function.

Lemma

In the Beta-binomial case and for large a, b we learn with respect to $\max \pi(\theta)$ *if*

$$n \geq \frac{(\mu - \frac{1}{2})^2}{\sigma^2}$$

Comparison of experiments

Suppose that θ has scientific meaning and that $\pi(\theta)$ does not depend on the experiment (within some class). Then by an experimental design, D, we mean a choice of $f(x, \theta)$

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

- $I_g(\theta, X|D_1) \leq I_g(\theta, X|D_2)$, for *all* g in our class
- $I_g(\theta, X|D_1) \leq I_g(\theta, X|D_2)$, for *one* g

Using Shannon 1: Maximum Entropy Sampling

Let $(Y, \tilde{Y})$ have a joint distribution and we want to observe Y and predict $\tilde{Y}$. Eg random field: $(Y, \tilde{Y})$ is the *whole population* and Y is the sampled values chosen by *design*. Then

$$\mathrm{Ent}(Y, \tilde{Y}) = \mathrm{Ent}(Y) + \mathrm{E}_Y\{\mathrm{Ent}(\tilde{Y}|Y)\}$$

Bayes learning says we want to minimise second term. But since lhs is fixed this is equivalent to maximising $\mathrm{Ent}(Y)$

Using Shannon 2: modelling and design

Let (Y, θ) have a joint distribution and we alkwant to observe Y and "predict" θ. Then

$$\text{Ent}(Y, \theta) = \text{Ent}(Y) + \text{E}_Y\{\text{Ent}(\theta|Y)\}$$

Again, Bayes learning says we want to minimise second term. So, again, if we can establish that the lhs does not depend on the design, then we have an equivalence:

MES is best for prediction and estimation

Use Shannon again, other way round:

$$\text{Ent}(Y, \theta) = \text{Ent}(\theta) + \text{E}_Y\{\text{Ent}(Y|\theta)\}$$

First term on rhs is just entropy of the prior: so want second term to be design independent: True for our standard model!

$$Y_i = f(\theta, x_i) + \epsilon_i,$$

ϵ_i independent of θ

Using Shannon 3: Information as a "valuation" (potential)

- Independence: take logs and expectations

$$f_{12} = f_1 f_2 \Rightarrow \log f_{12} = \log f_1 + \log f_2$$

$$I_{12} = I_1 + I_2$$

- Conditional independence

$$f_{123} = \frac{f_{13} f_{23}}{f_3}$$

$$\log f_{123} = \log f_{13} + \log f_{23} - \log f_3$$

$$I_{123} = I_{13} + I_{23} - I_3$$

- Example: decomposable graphical model

$$f_{12345} = \frac{f_{123} f_{234} f_{345}}{f_{23} f_{34}}$$

$$\log f_{12345} = \log f_{123} + \log f_{234} + \log f_{345} - \log f_{23} - \log f_{34}$$

Notice that I satifies "inclusion exclusion".

Lattice conditional independence, Gaussian case: LCI

White noise representation: $Y = A\epsilon$
Independence: $P_{12} = P_1 + P_2$
Conditional independence: $P_{123} = P_{13} + P_{23} - P_3$
Equivalent to commutativity: $P_{13}P_{23} = P_{23}P_{13}$
Decomposable:

$$P_{12345} = P_{123} + P_{234} + P_{345} - P_{23} - P_{34}$$

All the relevant projectors commute: simultaneously diagonalise. Boolean operation on the spectrum:

$$1234 = 123 + 234 + 345 - 23 - 34$$

Quantum information theory: the non-commutative case

Selective sampling: a generalisation of design

Accept n observations from $f(x,\theta)$ in a "window" A

$$f_A(x,\theta) = C\, f(x,\theta)\, \mathbb{I}_A(x)$$

- $I_g(\theta, X|A_1) \leq I_g(\theta, X|A_2)$, for *all* g in our class
- $I_g(\theta, X|A_1) \leq I_g(\theta, X|A_2)$, for *one* g

We could say: $A_1 \leq A_2$ or $A_1 \leq_g A_2$.

Example.

$$X \sim N(\theta, \sigma^2), \quad \theta \sim N(\mu, \tau^2)$$

$$A_K = (-\infty, \mu - K] \cup [\mu + K, \infty)$$

$$K_1 < K_2 \quad \Rightarrow \quad A_{K_1} \leq A_{K_2}$$

Observe where the prior is lowest!

Extend information-based methods to general design problems: only use data which is informative