Bayesian adaptive estimation under a random cost of observation associated with each observable variable*

Janne V. Kujala†

February 26, 2008

Abstract

In this paper, we adopt a decision theoretic view to Bayesian adaptive estimation. We extend the framework to situations where each observable variable is associated with a certain random cost of observation and consider the goal of maximizing the expected utility of a sequential experiment that ends when the total cost overruns a given budget. For example, the cost could be defined as the random time taken by each trial in an experiment, and one might wish to maximize the expected total information gain over as many trials as can be completed in 15 minutes. We propose a trial placement rule that maximizes the expected immediate gain in utility divided by the expected cost of observation. This myopic rule is shown to be asymptotically optimal under certain conditions and it is expected to work well in the same situations where the greedy immediate gain maximization works in the absence of costs. However, by simple concerete examples, we also show that the ubiquitous greedy information gain maximization strategy can in fact be *arbitrarily much* worse than the optimal strategy for a certain number of trials.

*This research was supported by the Academy of Finland (grant number 121855) and by the European Commission's FP6, Marie Curie Excellence Grants (MCEXT-CE-2004-014203). The author is grateful to Matti Vihola for comments.

†Address: Agora Center, University of Jyväskylä, P.O.Box 35, FI-40014 Jyväskylä, Finland. Email address: `jvk@iki.fi`. Fax: +358 14 2604400.

Contents

1 Introduction

The topic of this paper is Bayesian estimation of an unobservable random variable Θ based on a sequence $y_{x_1}, \ldots, y_{x_T}$ of independent (given θ) realizations from some conditional densities $p(y_{x_t} \mid \theta)$ indexed by trial *placements* x_t, each of which can be adaptively chosen from some set $\mathsf{X_t} \subset \mathsf{X}$ based on the outcomes $\mathbf{y} := (y_{x_1}, \ldots, y_{x_{t-1}})$ of the earlier observations.[1]

We assume that the goal of choosing the placements is to maximize the expected value of some *utility function* (DeGroot, 1970) of the knowledge about Θ, which under the Bayesian framework can always be formulated as a function of the (random) posterior distribution $\theta \mapsto p(\theta \mid Y_{X_1}, \ldots, Y_{X_t})$. Hence, the utility after the t-th observation is given by the random variable

$$U_t := u[\, \theta \mapsto p(\theta \mid Y_{X_1}, \ldots, Y_{X_t})\,],$$

where u is the utility function, and the goal is to maximize $\mathrm{E}(U_T)$ under some constraints such as a given total number T of trials or a given total budget for the costs associated with each observation.

In the rest of this section, we review relevant literature under the decision theoretic view. In Section 2, we further illustrate the concepts with several concerete examples. Finally, in Section 3, we consider the extension of the framework to the situation where the observation of each random variable is associated with a certain random cost. In an appendix, we go through the measure theoretic technicalities that are mostly avoided in the main text. In particular, we explicitly formulate the regularity conditions that are needed for Bayesian estimation.

1.1 Psychophysics

In psychophysics, Bayesian adaptive estimation was first considered by Watson and Pelli (1983) for estimation of an observer's "threshold" α of detecting a stimulus of given intensity x. The dichotomous result of detecting (1) or not detecting (0) the stimulus is assumed to be distributed as

$$p(y_x \mid \theta) = \begin{cases} \psi_\theta(x), & y_x = 1, \\ 1 - \psi_\theta(x), & y_x = 0, \end{cases}$$

where ψ_θ is some sigmoidal function described by the four parameters $\theta = (\alpha, \beta, \gamma, \delta)$, where α is the *threshold*, β determines the *slope* at the threshold, γ is the *guessing rate*, and δ is the *lapsing rate*, see Fig. 1.

Watson and Pelli (1983) assume that only the α component is unknown and define the *loss function* (negative of a utility function) as the variance of the posterior distribution of α. Their adaptive method places each test intensity x_t at the mode of the posterior distribution $p(\alpha \mid \mathbf{y})$. However, King-Smith (1984; King-Smith et al., 1994) discovered that placement at the mean is more efficient and that it is even more efficient to use the implicit rule of choosing x_t so as to minimize the expected posterior variance after the observation of Y_{x_t}.

[1] Note that in some experiments, one can observe multiple independent (given θ) copies of the same random variable Y_x. However, instead of complicating the general notation with something like $Y_{x_t}^{(t)}$, we rely on the fact that the set X can explictly include separate indices for any identically distributed copies, for example, one might have $[Y_{(x,t)} \mid \theta] \overset{\text{i.i.d.}}{\sim} [Y_{(x,t')} \mid \theta]$ for all $t, t' \in \mathbb{N}$. Hence, we can keep the simple notation with no loss of generality.

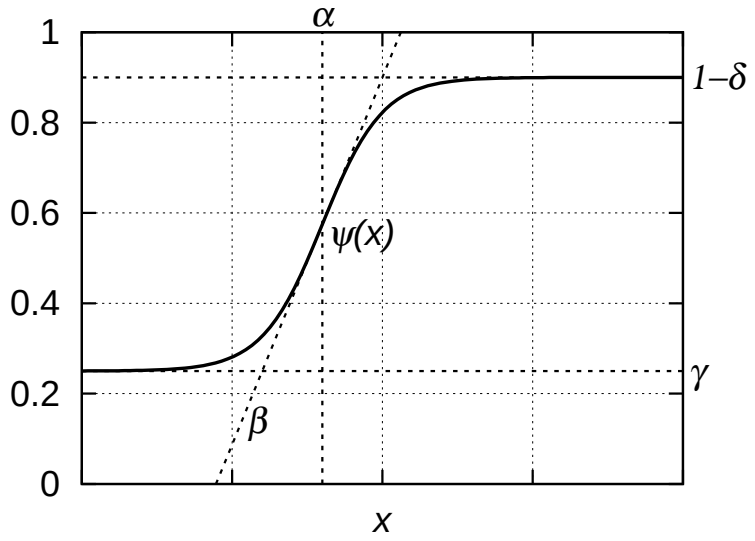


Figure 1: An illustration of a typical psychometric function relating stimulus intensity x to the oberver's probability of detecting it.

Kontsevich and Tyler (1999) consider simultaneous estimation of both the threshold α and the slope β. As the variance does not easily generalize into a loss function of multivariate uncertainty (it would require arbitrary weighting of the uncertainties along each dimension), Kontsevich and Tyler instead use the (differential[2]) Shannon (1948) entropy

$$\mathrm{H}(\Theta) = -\int p(\theta) \log p(\theta) d\theta, \tag{1}$$

which is well-defined for multivariate θ, too. Their adaptive method chooses each placement so as to minimize the expected posterior entropy of (α, β) after the observation of Y_{x_t}. This is a more elegant solution than the asymptotically justified explicit rules (e.g., King-Smith and Rose, 1997; Snoeren and Puts, 1997) that have been used before.

In recent works (Kujala and Lukka, 2006; Lesmes et al., 2006) the same algorithm is generalized to multivariate placements $x \in \mathsf{X} = \mathbb{R}^2$, too. We do not go into any details here, as in the present theoretical framework X is considered just as an unstructured set and so there is no conceptual difference to the univariate placement case.

1.2 The greedy startegy

The implicit placement rules mentioned above are special cases of a *greedy strategy*, which chooses each placement so as to optimize the expected immeadiate gain in utility after the next observation:

$$\begin{aligned} x_t &:= \arg\max_{x \in \mathsf{X}_t} \mathrm{E}(U_t \mid \mathbf{y},\ X_t = x) \\ &= \arg\max_{x \in \mathsf{X}_t} \mathrm{E}(U_t - U_{t-1} \mid \mathbf{y},\ X_t = x). \end{aligned}$$

[2]We use the same notation $\int f(\theta)p(\theta)d\theta$ for both the continuous case and the discrete case, in which it corresponds to a sum. This is measure-theoretically justified as "$d\theta$" can be considered as the counting measure in the discrete case. Thus, following Lindley (1956), even though we use the familiar notation, we are in fact working in full measure-theoretic generality, allowing the density $p(\theta)$ to be w.r.t. any measure "$d\theta$". See Sec. B for the technical details.

As the value of U_{t-1} will be known at the time of choosing x_t, maximizing the expected change of utility is equivalent to maximizing its expected value. This strategy can be applied in any model for any utility function, but it is generally not the globally optimal strategy except when the experiment consists of exactly one trial.

Remark 1.1. Strictly speaking it is possible that no maximum of the expected utility exists, in which case one should generally choose such a placement that yields an expected gain sufficiently close to the supremum (DeGroot, 1970).

1.3 The entropy loss function and mutual information

The entropy loss function (1) has some remarkable properties which sometimes go unnoticed by authors. Most imporantly, the expected difference between the prior entropy and the posterior entropy given Y_x is parametrization[3] invariant (Lindley, 1956). Hence, even though the value U_t of the utility function defined as the (negative of the) entropy as well as its change $U_t - U_{t-1}$ do depend on the parametrization chosen for θ, the expected change $\mathrm{E}(U_t - U_{t-1} \mid \mathbf{y},\ X_t = x)$ is parametrization invariant; it corresponds to the information-theoretic *mutual information*[4]

$$\mathrm{I}(\Theta; Y_x \mid \mathbf{y}) = \mathrm{H}(\Theta \mid \mathbf{y}) - \mathrm{E}[\mathrm{H}(\Theta \mid \mathbf{y},\ Y_x)] \tag{2}$$

of Θ and Y_x (given $\mathbf{y}$). Indeed, the mutual information generally defined as

$$\mathrm{I}(A; B) := \iint p(a,b) \log \frac{p(a,b)}{p(a)p(b)} da\, db$$

is insensitive to any one-to-one transformations of A and B, and as defined, it is obviously symmetric, which yields the identity

$$\mathrm{H}(A) - \mathrm{E}[\mathrm{H}(A \mid B)] = \mathrm{I}(A; B) = \mathrm{I}(B; A) = \mathrm{H}(B) - \mathrm{E}[\mathrm{H}(B \mid A)],$$

which holds whenever the differences are defined, i.e., not $\infty - \infty$ or $-\infty - (-\infty)$ (see Cover and Thomas, 1991). Thus, the mutual information represents the expected amount of information that the observation of one random variable gives about the other. This expected amount is always nonnegative even though the entropy might actually increase in case an "unexpected" outcome is observed.

1.4 Calculating expected information gain

Making use of the symmetry of the mutual information, Kujala and Lukka (2006) write the objective function (2) as

$$\mathrm{I}(\Theta; Y_x \mid \mathbf{y}) = \mathrm{H}(Y_x \mid \mathbf{y}) - \mathrm{E}[\mathrm{H}(Y_x \mid \Theta,\ \mathbf{y})]. \tag{3}$$

This formulation is usually more convenient than (2) as the distribution of Y_x given θ is typically much simpler than that of Θ given y_x. For example, in the case of dichotomous results, this yields the convenient expression

$$\mathrm{I}(\Theta; Y_x \mid \mathbf{y}) = h\left(\int \Pr\{Y_x = 1 \mid \theta\} p(\theta \mid \mathbf{y})\right) - \int h(\Pr\{Y_x = 1 \mid \theta\}) p(\theta \mid \mathbf{y}) d\theta,$$

[3] In the measure-theoretic framework, "parametrization" can be interpreted as the choice of the dominating measure "$d\theta$" w.r.t. which the density $p(\theta)$ is taken.

[4] In our notation $\mathrm{H}(A \mid \dots)$ always denotes the conditional entropy of A given the (possibly random) conditioning values $[\dots]$. Thus, $\mathrm{H}(A \mid \dots)$ will be a random variable if any of its conditioning values are random variables unlike in the unfortunate standard notation where an expectation is implicitly taken over the conditioning values. Also, $\mathrm{I}(\Theta; Y_x \mid \mathbf{y})$ denotes the mutual information of the random variables $\Theta \mid \mathbf{y}$ and $Y_x \mid \mathbf{y}$, that is, both Θ and Y_x are conditioned on $\mathbf{y}$. This is standard notation.

where[5] $h(p) := -p \log p - (1-p)\log(1-p)$ is the entropy of a binary distribution with probabilities p and $1-p$. Thus, only expectations over the (sequential) prior $p(\theta \mid \mathbf{y})$ are needed which allows for efficient computation. For example, given an (approximately) i.i.d. sample $\{\theta_i\}_{i=1}^N$ drawn from $p(\theta \mid \mathbf{y})$, the objective function can be approximated as

$$\mathrm{I}(\Theta; Y_x \mid \mathbf{y}) \approx h\left(\frac{1}{N}\sum_{i=1}^{N} \Pr\{Y_x = 1 \mid \theta_i\}\right) - \frac{1}{N}\sum_{i=1}^{N} h(\Pr\{Y_x = 1 \mid \theta_i\}).$$

This was used by Kujala and Lukka (2006) in a sequential Monte Carlo implementation of the greedy algorithm.

The same idea generalizes to any finite number of outcomes:

$$\begin{aligned}\mathrm{I}(\Theta; Y_x \mid \mathbf{y}) &= \sum_{y_x} g\left(\int p(y_x \mid \theta) p(\theta \mid \mathbf{y})\right) - \int \sum_{y_x} g(p(y_x \mid \theta)) p(\theta \mid \mathbf{y}) d\theta, \\ &\approx \sum_{y_x} g\left(\frac{1}{N}\sum_{i=1}^{N} p(y_x \mid \theta_i)\right) - \frac{1}{N}\sum_{i=1}^{N}\sum_{y_x} g(p(y_x \mid \theta_i)),\end{aligned}$$

where $g(p) = -p\log(p)$. Although this formulation has not yet been used in any published works, it could be directly applied to, for example, the choice model and MCMC algorithm used in (Kujala et al., submitted).

Not only is (3) usually computationally more convenient than (2), but there is also a theoretical advantage. If Y_x is dichotomous or has a finite number of possible values, the entropies on the right side of (2) will always be finite, and the expression is well-defined unlike (2) which may come out $\infty - \infty$ or $-\infty - (-\infty)$ for some parametrizations of θ. Of course, if both Y_x and Θ have an infinite number of possible values, then either formulation may fail for some parametrization. Nonetheless, the mutual information itself is always well-defined (see Sec. B.2 for the measure-theoretic details). Hence, Kolmogorov (1956) argues that it is in fact the mutual information that is the fundamental concept of the theory of information.

Remark 1.2. Although the information gain may theoretically be infinite (e.g., $Y = \Theta \sim \mathrm{Uniform}[0,1]$ yields $\mathrm{I}(\Theta; Y) = \infty$), that will never happen in a realistic model as the observation of any real quantity Y is always subject to some measurement error.

1.5 Nuisance variables and utility weights

In some cases, one might be interested only in the value of some component Θ_1 of $\Theta = (\Theta_1, \Theta_2)$ even if its other components Θ_2 are unknown, too. In that case, one can define the utility function as the (negative of the) marginal entropy

$$U_t = -\mathrm{H}(\Theta_1 \mid Y_{X_1}, \ldots, Y_{X_t})$$

of the interesting variables. The expected change of entropy still corresponds to the mutual information

$$\mathrm{E}(U_t - U_{t-1} \mid \mathbf{y}) = \mathrm{I}(Y_x; \Theta_1 \mid \mathbf{y})$$

and hence enjoys the same parametrization invariance properties.

[5] We shall assume base e logarithm in all expressions, but we will give numerical results in bits, i.e., we define bit $:= \log 2$.

More generally, a parametrization-invariant objective function can always be defined as an arbitrary function

$$f(\mathrm{I}(Y_x; T_1(\Theta) \mid \mathbf{y}), \ldots, \mathrm{I}(Y_x; T_n(\Theta) \mid \mathbf{y}))$$

of mutual informations, where $T_1, \ldots, T_n$ can be any functions (e.g., T_k can be the component mappings $T_k(\Theta) = \Theta_k$). In particular, the maximizer x of any linear combination of marginal entropies (a class of utility functions used in the method of Tanner et al., 2005, also mentioned by Paninski, 2005)

$$U_t = -\sum_k w_k \, \mathrm{H}(\Theta_k \mid Y_{X_1}, \ldots, Y_{X_t})$$

given $\mathbf{y}$ and $X_t = x$ is insensitive to the parametrization of Θ as the expected gain in utility

$$\mathrm{E}(U_t - U_{t-1} \mid \mathbf{y}) = \sum_k w_k \, \mathrm{I}(Y_x; \Theta_k \mid \mathbf{y})$$

is a (linear) function of parametrization-invariant mutual informations.

However, the expected change of any nonlinear function of entropies no longer corresponds to a function of mutual informations and hence does not inherit the parametrization invariance. Conversely, maximization of a nonlinear combination of mutual informations generally does not correspond to the maximization of the expected value of any utility function.

While nuisance variables do not pose any conceptual problems, they can complicate the practical computations. It is usually computationally easier to apply (3) to Θ than any subset of its components, as integration out of the nuisance variables from the model at each trial interacts badly with the computational conveniences that (3) provides over (2). Furthermore, as we shall see in the following, the greedy algorithm typically works best when information about the nuisance variables is included in the utility function (although the greedy strategy can fail in that case, too, as we shall see).

1.6 Global strategies

Any adaptive (or non-adaptive) placement strategy defines a *decision function*

$$d : \mathsf{Y}_d \to \mathsf{X} \cup \{\lambda\} : (y_{x_1}, \ldots, y_{x_{t-1}}) \mapsto x_t,$$

whose domain Y_d is the set of possible sequences of trial results (including the empty sequence) and whose value is the next placement or the special value λ which flags the end of the experiment. Now we can define the random variable

$$Y_d := (Y_{X_1}, \ldots, Y_{X_T}) \in \mathsf{Y}_d$$

denoting the outcome of the whole adaptive experiment following the decision function d, where T is the possibly random time index of the trial that ends the experiment.

The fact that the whole adaptive experiment can be seen as just one observation Y_d implies that all the parametrization-invariance results of the entropy loss function apply to the whole-experiment startegies as well, regardless of whether the termination rule is adaptive or not, as long as the experiment eventually terminates (and even that is not strictly necessary if the value of the utility function converges with probability one). One could even allow randomized decision functions mapping to distributions over $\mathsf{X} \cup \{\lambda\}$ instead of deterministic values. However, as randomized decisions will generally gain nothing over deterministic ones (DeGroot, 1970), we shall only consider deterministic decision functions except for one reference to a random termination rule in Sec. 3.1 (where the randomness is outside the experimenter's control).

1.7 The globally optimal strategy

The optimal strategy is obviously to maximize the expected utility after the observation of the whole experiment result Y_d w.r.t. the decision function d. For example, in the case of dichotomous results and constant experiment length T, the decision function defines a binary decision tree with $2^T - 1$ nodes, each denoting the next placement in a certain situation. Thus, the optimal strategy can in theory be found by optimizing over these $2^T - 1$ parameters. In practice, the exponentially growing number of parameters makes the optimal strategy generally intractable very soon as T grows (although in some special cases the globally optimal strategy can be found analytically).

To improve the efficiency of the greedy algorithm in practice, one could try to apply it to the observable variables Y_d indexed by d varying in the set of decision trees of a certain small depth. For example, King-Smith et al. (1994) have implemented this strategy for a two-step "look-ahead" in a psychometric estimation procedure. However, their simulations indicated that the improvement over the one-step greedy strategy was generally small (compared to the one-step method after 16 trials, the two-step method yielded the same accuracy at 15.97 trials for $\gamma = 0.03$, $\delta = 0.01$ and at 15.6 trials for $\gamma = 0.5$, $\delta = 0.01$). Still, much larger improvements may be possible in other models.

1.8 Non-adaptive and batch strategies

Lindley (1956) considers the case where i.i.d. copies of the same random variable Y_x are observed sequentially and shows that the expected information gain $\mathrm{I}(Y_x^{(1)}, \ldots, Y_x^{(t)}; \Theta)$ over t observations is a concave function of t, i.e., the expected information gains from each additional observation are non-increasing. Note that it is possible that the first observation of some Y_x is expected to yield more information than the first observation of $Y_{x'}$, but that over repeated i.i.d. observations, $Y_{x'}$ would be more informative than the same number of observations of Y_x (this is the case in the example of Sec. 2.1).

In recent works, Müller et al. (2004) and Amzal et al. (2006) consider Monte Carlo simulation methods for optimal experiment design under a given set of adjustable parameters, such as the placements $x_1, \ldots, x_T$ of all trials in a non-adaptive experiment. These methods may also be useful for such adaptive designs where one has to choose the placements for a batch of n trials simultaneously. In that case, one could apply the optimal design algorithm for each batch within a greedy strategy that considers each batch as a single trial. In theory, the same simulation methods could be applied to the complete design d of an adaptive experiment, too, although the exponential number of parameters would become a problem soon as discussed above.

2 Examples

The greedy algorithm of optimizing the expected immediate gain is ubiquitous in practical applications of Bayesian adaptive estimation. However, little is known about its relative efficiency compared to other strategies. The only definitive result so far appears to be that under certain regularity conditions, the greedy strategy can be shown to be asymptotically more efficient than any non-adaptive strategy (Paninski, 2005).

In this section, through simple concrete examples, we demonstrate in particular that

1. the per-trial efficiency of batch strategies generally deteriorates rapidly as the batch size increases,

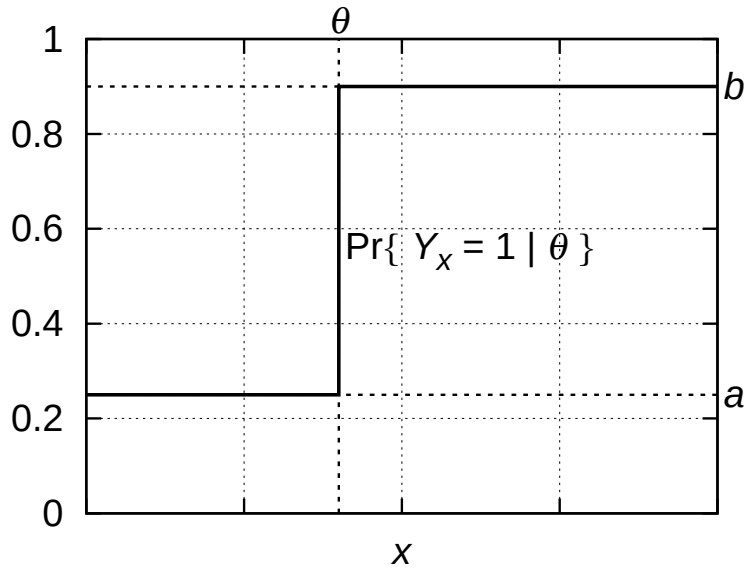


Figure 2: A simplified version of the psychometric model shown in Fig. 1 with an infinite slope at the thershold θ.

2. although the greedy algorithm usually works, it can be *arbitrarily much* worse than the globally optimal strategy, and

3. even the globally optimal strategy can apparently fail due to the fact that the entropy loss function can be inappropriate in certain situations.

2.1 A simplified psychometric model

Example 2.1. Suppose that for all $x \in \mathbb{R}$, $Y_x \in \{0, 1\}$ is a dichotomous random variable defined by

$$\Pr\{Y_x = 1 \mid \theta\} = \begin{cases} a, & x < \theta, \\ b, & x \geq \theta \end{cases}$$

for some $0 \leq a < b \leq 1$, see Fig. 2. While this model is simple, it does have the important feature that there is uncertainty of the results due to both the fact that Θ is unknown and the fact that for a given θ, the result $[Y_x \mid \theta]$ is random. This model differs from the typical psychometric model only in that, due to the infinite slope at the threshold, the obtainable information gains do not decrease over time as the scale of uncertainty reduces. However, this detail is not very important during the first few trials, and therefore this example serves to illustrate the relative efficiencies of different placement strategies that can be expected in a typical psychophysical experiment and explain the success of the greedy strategy.

2.1.1 Success of the greedy strategy

Using (3), the expected information gain of observing Y_x given any prior data $\mathbf{y} = (y_{x_1}, \ldots, y_{x_{t-1}})$ can be calculated as

$$\begin{aligned} \mathrm{I}(Y_x; \Theta \mid \mathbf{y}) &= \mathrm{H}(Y_x \mid \mathbf{y}) - \mathrm{E}[\mathrm{H}(Y_x \mid \Theta,\ \mathbf{y})] \\ &= h(a + (b-a)\Pr\{\Theta \leq x \mid \mathbf{y}\}) \\ &\quad - (h(a) + (h(b) - h(a))\Pr\{\Theta \leq x \mid \mathbf{y}\}), \end{aligned} \tag{4}$$

which depends on the placement x and the prior data $\mathbf{y}$ only through $z := \Pr\{\Theta \leq x \mid \mathbf{y}\}$. Assuming that the prior on Θ is absolutely continuous w.r.t. the Lebesgue measure, the same will be true for the posterior given $\mathbf{y}$, and one can always choose x so as to attain any value of $z \in [0,1]$. As (4) is continuous on the compact interval $z \in [0,1]$, it attains a maximum value (which is independent of any prior data). It follows that the greedy strategy yields the maximum expected total information gain over any given constant number of trials.

Let us then find the value(s) of z that maximize (4). Obviously the expression is smooth and positive (assuming $a \neq b$) for $z \in (0,1)$ and zero for $z = 0$ or $z = 1$. Hence, the expression can attain a maximum value only at critical points $z \in (0,1)$. The derivative w.r.t. z is

$$h'(a + (b-a)z)(b-a) - (h(b) - h(a)),$$

where $h'(p) = \log(1/p - 1)$. The derivative is zero iff

$$\log\left(\frac{1}{a + (b-a)z} - 1\right) = \frac{h(b) - h(a)}{b - a},$$

and so the maximum value is attained at the unique point

$$z^* = \frac{\dfrac{1}{1 + \exp\left(\frac{h(b)-h(a)}{b-a}\right)} - a}{b - a}. \tag{5}$$

The value at this point is (after some algebra)

$$\mathrm{I}(Y_{x^*}; \Theta \mid \mathbf{y}) = \log\left(1 + \exp\left(\frac{h(b) - h(a)}{b - a}\right)\right) - \frac{(1-a)h(b) - (1-b)h(a)}{b - a}. \tag{6}$$

2.1.2 Non-adaptive and batch strategies

Let us then consider non-adaptive strategies with a set of n placements $x_1 \leq \cdots \leq x_n$ to be chosen before the experiment. Denoting $g(p) = -p \log p$, $z_k := \Pr\{\Theta \leq x_k\}$, $z_0 = 0$, and $z_1 = 1$, the expected information gain can be written as

$$\begin{aligned}
\mathrm{I}(Y_{x_1}, \ldots, Y_{x_n}; \Theta) &= \mathrm{H}(Y_{x_1}, \ldots, Y_{x_n}) - \mathrm{E}[\mathrm{H}(Y_{x_1}, \ldots, Y_{x_n} \mid \Theta)] \\
&= \sum_{y_{x_1}} \cdots \sum_{y_{x_n}} g\left(\int p(\theta) d\theta \prod_{j=1}^{n} p(y_{x_j} \mid \theta)\right) - \sum_{j=1}^{n} \mathrm{E}[\mathrm{H}(Y_{x_j} \mid \Theta)] \\
&= \sum_{y_1=0}^{1} \cdots \sum_{y_n=0}^{1} g\left(\sum_{k=0}^{n} (z_{k+1} - z_k) \prod_{j=1}^{n} \begin{cases} a, & j \leq k,\ y_j = 1, \\ 1-a, & j \leq k,\ y_j = 0, \\ b, & j > k,\ y_j = 1, \\ 1-b, & j > k,\ y_j = 0, \end{cases}\right) \\
&\quad - \sum_{k=1}^{n} [h(a) + (h(b) - h(a)) z_k] \\
&= -(Mz + v) \cdot \log(Mz + v) - u \cdot z - c =: f(z)
\end{aligned}$$

for some matrix $M \in \mathbb{R}^{2^n \times n}$, vectors $v \in \mathbb{R}^{2^n}$ and $u \in \mathbb{R}^n$, and constant $c \in \mathbb{R}$, where $z = (z_1, \ldots, z_n) \in \mathbb{R}^n$ and the log function is applied elementwise to $Mz + v \in [0,1]^{2^n}$.

The gradient of the function is

$$\nabla f(z) = -M^T \left(\log(Mz + v) + \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} \right) - u$$

and the Hessian matrix is

$$Hf(z) = -M^T \mathrm{diag}(\underbrace{\mathrm{inv}(Mz + v)}_{\geq 1})M,$$

where inv denotes elementwise inverse. As the Hessian is negative definite (M has full rank given $a \neq b$), the function is strictly concave and therefore has a unqiue maximizer in the compact convex set $0 \leq z_1 \leq \cdots \leq z_n \leq 1$, which can be found using, for example, Newton's iteration $z^{(t+1)} = z^{(t)} - \lambda[Hf(z^{(t)})]^{-1}\nabla f(z^{(t)})$, (where at each step t we try $\lambda = 1,\ 0.1,\ 0.01,\ \ldots$ until $z^{(t+1)}$ is within the convex feasible set).

While this linear algebraic formulation appears simple, it should be noted that the matrix M has an exponential number of rows. Thus, this deterministic approach will only work up to around $n = 20$. Beyond that, one may have to resort to a Monte Carlo approach to simulate the expected results of each possible set of placements as discussed in Sec. 1.8.

Finally, let us consider a non-adaptive design where n i.i.d. trials are conducted with the *same* placement x. Conceptually, this is just a restricted special case of the general non-adaptive design considered above, but instead of the 2^n different outcomes, one can now consider a binomial distribution of the number of 1-results. However, as this still does not lead to a closed form solution, we do not go into the details.

2.1.3 Comparison

The optimal placements of the three strategies for $n = 1, \ldots, 9$ are shown in Fig. 3. Apparently the efficiency of the non-adaptive strategies decreases rapidly relative to the optimal strategy as n increases. The reasons for this are intuitively simple: when a is close to zero and b close to one, the adaptive strategy can sequentially bisect the range down to one of 2^n distinct sections, yielding n bits of information, while the non-adaptive strategy can only divide the range to $n + 1$ distinct sections yielding at most $\log_2(n + 1)$ bits of information and the single-placement strategy with only 2 sections tops off at 1 bit. However, as the guessing and lapsing rates increase, the differences between the strategies become smaller. For example, with $a = .5$ and $b = .8$, one could present batches of 3 identical trials with only a small loss in efficiency compared to the fully adaptive strategy.

The optimal adaptive placement (5) appears to be close to the median of the distribution of $[\Theta \mid \mathbf{y}]$ under several values of a and b. With two-alternative forced choice design (i.e., guessing rate $a = 1/2$), the optimal placement tends to $z = 0.6$ as $b \to 1$, and for any a and b, the placement z is within $[1/e, 1 - 1/e] \approx [.3679, .6321]$.

2.2 Trade-off between guessing and lapsing rates

Often one can affect the guessing and lapsing rates by the design of the experiment. Obviously if both can be decreased, then the informativity of the experiment should increase. But what if the guessing rate can only be decreased at the cost of a higher lapsing rate, or vice versa? In this section we give a definitive answer under Lindley's (1956) method of comparing experiments, that is, we determine when one experiment is more informative than the other regardless of the prior information.

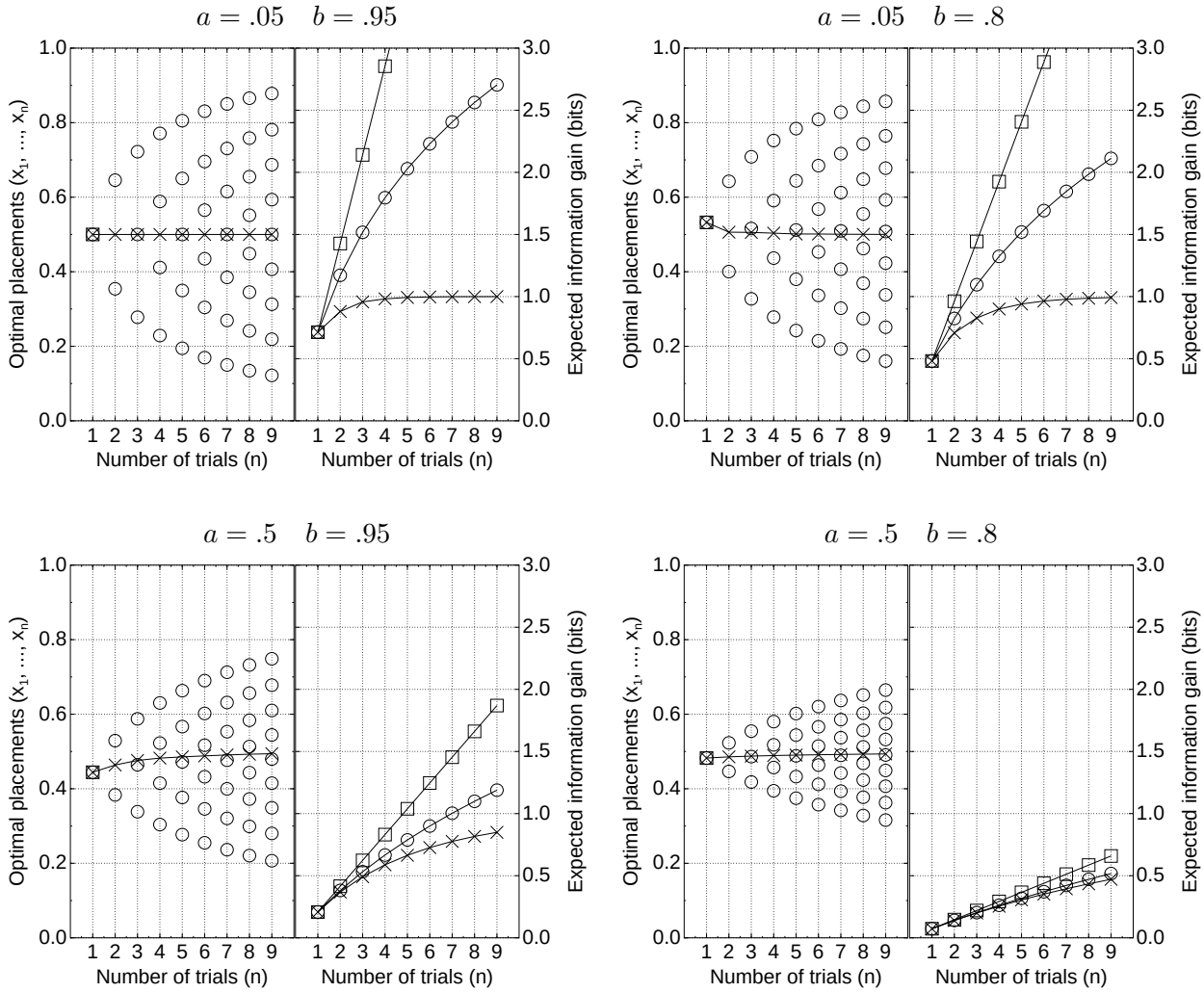


Figure 3: Illustrations of the placements and efficiency of the optimal adaptive design (squares), optimal non-adaptive design (circles), and optimal single-placement repeated observations design (crosses) in the model shown in Fig. 2. These are the optimal placements assuming a uniform prior for Θ on $[0, 1]$. For any other (absolutely continuous) prior, the placements are obtained by interpreting the values on the y-axis as the fractiles $\Pr\{\Theta \leq x_k \mid \mathbf{y}\}$, which also yields the not shown optimal adaptive placements for $n > 1$. The circles are almost but not exactly evenly spaced for each n. The crosses approach a gain of 1 bit and placement at 0.5 as $n \to \infty$.

Theorem 2.1. *Let $Y_{a,b}$ be a dichotomous random variable depending on another random variable $\Theta \in [0,1]$ through the conditional distribution*

$$\Pr\{Y_{a,b} = 1 \mid \theta\} = a + (b-a)\theta,$$

where $a, b \in [0,1]$ and $a \neq b$. If both of the inequalities

$$\frac{a(1-a)}{(b-a)^2} \leq \frac{a'(1-a')}{(b'-a')^2} \tag{7}$$

$$\frac{b(1-b)}{(b-a)^2} \leq \frac{b'(1-b')}{(b'-a')^2}, \tag{8}$$

are satisfied, then $\mathrm{I}(Y_{a',b'};\Theta) \leq \mathrm{I}(Y_{a,b};\Theta)$ regardless of the distribution of Θ. If, in addition, the inequality in (7) or (8) is strict, then $\mathrm{I}(Y_{a',b'};\Theta) < I(Y_{a,b};\Theta)$ provided that Θ is not concentrated on a single point. If (7) holds with $<$ and (8) holds with $>$ (or vice versa), then, depending on the distribution of Θ, either one of $\mathrm{I}(Y_{a,b};\Theta)$ and $\mathrm{I}(Y_{a',b'};\Theta)$ can be strictly larger than the other.

Proof. The mutual information can be written as

$$\begin{aligned} \mathrm{I}(Y_{a,b};\Theta) &= \mathrm{H}(Y_{a,b}) - \mathrm{E}[\mathrm{H}(Y_{a,b} \mid \Theta)] \\ &= h(a + (b-a)\,\mathrm{E}[\Theta]) - \mathrm{E}[h(a + (b-a)\Theta)] \\ &= h_{a,b}(\mathrm{E}[\Theta]) - \mathrm{E}[h_{a,b}(\Theta)], \end{aligned} \tag{9}$$

where

$$\begin{aligned} h(x) &= -x\log x - (1-x)\log(1-x), \\ h_{a,b}(x) &= h(a + (b-a)x). \end{aligned}$$

As $h''(x) = -1/(x(1-x))$, we obtain the expression

$$\begin{aligned} h''_{a,b}(x) &= h''(a + (b-a)x)(b-a)^2 \\ &= -\frac{(b-a)^2}{[a + (b-a)x][1 - a - (b-a)x]}, \end{aligned}$$

which is continuous on $x \in (0,1)$. Thus, Lemma A.1 implies that the conclusion $\mathrm{I}(Y_{a',b'};\Theta) \leq I(Y_{a,b};\Theta)$ follows if $h''_{a,b}(x) \leq h''_{a',b'}(x)$ for all $x \in (0,1)$. But this inequality is equivalent to

$$\frac{[a + (b-a)x][1 - a - (b-a)x]}{(b-a)^2} \leq \frac{[a' + (b'-a')x][1 - a' - (b'-a')x]}{(b'-a')^2}$$

which is linear (the second order terms cancel) and therefore holds for all $x \in (0,1)$ if and only if it holds at the end points $x \in \{0,1\}$. This is precisely the condition given by (7) and (8) in the statement of the theorem. Furthermore, if the linear inequality is strict at either end point, then it is strict at every $x \in (0,1)$ and Lemma A.1 implies $\mathrm{I}(Y_{a',b'};\Theta) < I(Y_{a,b};\Theta)$.

Finally, if $h''_{a,b}(0) < h''_{a',b'}(0)$ and $h''_{a,b}(1) > h''_{a',b'}(1)$ (or vice versa), then one can apply Lemma A.1 to nonsingular distributions supported on sufficiently small ranges $[0,\epsilon]$ and $[1-\epsilon,1]$ to show that both $\mathrm{I}(Y_{a,b};\Theta) > I(Y_{a',b'};\Theta)$ and $\mathrm{I}(Y_{a,b};\Theta) < I(Y_{a',b'};\Theta)$ are possible. □

Corollary 2.2. *If $a' \neq b'$ and $[a',b'] \subset [a,b] \subset [0,1]$, then $\mathrm{I}(Y_{a,b};\Theta) \geq \mathrm{I}(Y_{a',b'};\Theta)$.*

Proof. Given $b \in [b', 1]$, the inequality

$$\frac{a(1-a)}{(b-a)^2} \leq \frac{a'(1-a')}{(b'-a')^2}$$

obviously holds at $a = a'$. Differentiating the left side w.r.t. a yields

$$\frac{(1-2a)(b-a)^2 - a(1-a)(-2)(b-a)}{(b-a)^4} = \frac{(b-a) + 2a(1-b)}{(b-a)^3} \geq 0$$

and thus extends the inequality for all $a \in [0, a']$. The inequality (8) holds analogously. □

2.3 Exploration versus exploitation

Example 2.2. Suppose Y_n is a dichotomous random variable depending on the random variables Θ and M through the conditional distribution

$$\Pr\{Y_n = 1 \mid \theta, m\} = \begin{cases} \theta + (1-\theta)\frac{1}{n}, & n \leq m, \\ \frac{1}{n}, & n > m. \end{cases}$$

Here n is the number of alternatives in a multiple choice task and Θ represents the probability that the observer knows the correct answer to each question. If the correct answer is not known, the observer is assumed to guess. Increasing n decreases the probability of guessing correctly, but we also assume that there is an unknown maximum number $M \geq 2$ of choices that the observer can handle before being overwhelmed in which case the answer will be random again. This leads to a kind of trade-off between the guessing and lapsing rate.

Assuming that Θ and M are independent, we have

$$\Pr\{Y_n = 1 \mid \theta\} = \Pr\{n \leq M\}\left[\theta + (1-\theta)\frac{1}{n}\right] + \Pr\{n > M\}\frac{1}{n}$$

$$= \frac{1}{n} + \Pr\{n \leq M\}\frac{n-1}{n}\theta.$$

Using the notation of Theorem 2.1, we see that $[Y_n \mid \theta] \sim [Y_{a,b} \mid \theta]$ for $a = 1/n$ and $b - a = \Pr\{n \leq M\}(n-1)/n$. Assuming that $M \in \{2, \ldots, N\}$, we have $[Y_2 \mid \theta] \sim [Y_{1/2,1} \mid \theta]$ and so, $\mathrm{I}(Y_2; \Theta) > \mathrm{I}(Y_n; \Theta)$ if

$$1 = \frac{\frac{1}{2}(1-\frac{1}{2})}{(1-\frac{1}{2})^2} < \frac{\frac{1}{n}(1-\frac{1}{n})}{\left(\Pr\{n \leq M\}\frac{n-1}{n}\right)^2}$$

and

$$0 = \frac{1(1-1)}{(1-\frac{1}{2})^2} \leq \frac{[\ldots](1-[\ldots])}{\left(\Pr\{n \leq M\}\frac{n-1}{n}\right)^2}$$

which is equivalent to

$$\Pr\{n \leq M\} < \frac{1}{\sqrt{n-1}}.$$

This is satisfied for all $n \geq 2$, if, for example, we define the distribution of M by

$$p(m) \propto \begin{cases} 1/m, & m = 2, \ldots, 7, \\ 0, & \text{otherwise.} \end{cases}$$

For this prior distribution of M, maximization of $\mathrm{I}(Y_n;\Theta)$ yields $n=2$ regardless of the distribution of Θ. But as Y_2 does not depend on M, the posterior distribution of M given the the result y_2 remains unchanged and so $\mathrm{I}(Y_n;\Theta \mid y_2^{(1)},\dots,y_2^{(t)})$ is always maximized by $n=2$, and the greedy algorithm will only present trials with $n=2$. This is clearly suboptimal in the long run as there is a positive probability that $M>2$. Presenting some trials with $n>2$ first would allow estimating M. Once the value is known with high enough confidence, the rest of the trials can be presented with the optimal number of choices $n=M$.

It is often suggested that the objective function of the greedy algorithm should combine both the expected gain of utility as well as the expected gain of information about all unknowns as it might lead to better gains of utility in the following trials (e.g., Verdinelli and Kadane, 1992). Thus, even though M is a nuisance variable we are not interested in estimating, it turns out that it would still be more efficient in the long run to apply the greedy algorithm to minimization of $\mathrm{H}(\Theta,M \mid \mathbf{y})$ instead of $\mathrm{H}(\Theta \mid \mathbf{y})$. Indeed, this strategy appears to be asymptotically optimal in the sense that the distribution of the placements eventually converges to M provided that $\Theta>0$.[6]

However, even if the utility function is defined as the information gained about all unobservable variables, the greedy strategy can still be arbitrarily much worse than the optimal strategy:

Example 2.3. Suppose that each of $\Theta\sim\mathrm{Uniform}[0,1]$ and $\Phi_n\sim\mathrm{Uniform}\{1,\dots,n\}$ for $n\in\{1,2,\dots\}$ are independent, unobsorvable variables. The observable variables are given by $Y_{n,x,a,b}\in\{-1,0,\dots,2^{n-1}-1\}$, $n,x\in\{1,2,\dots\}$, $a,b\in[0,1]$, where

$$Y_{n,x,a,b}=\begin{cases}0, & x=\Phi_n,\ \Theta<a,\\ \left\lfloor 2^n\frac{\Theta-a}{b-a}\right\rfloor, & x=\Phi_n,\ a\le\Theta<b,\\ 2^n-1, & x=\Phi_n,\ \Theta\ge b,\\ -1, & \text{otherwise.}\end{cases}$$

In this model, the posterior of Θ given any observations $\mathbf{y}$ will always be a uniform distribution on some interval $[a,b]$. For any given n, if one knows or guesses correctly the value of Φ_n, then observation of $Y_{n,x,a,b}$ with $x=\Phi_n$ reduces the posterior interval $[a,b]$ to some of its 2^n subdivisions and thus yields n new bits of information about Θ (as well as confirms the value of Φ_n if it was uncertain). An incorrect guess yields no information about Θ, but decreases the set of possible values of Φ_n by one. However, as there is no randomness in the observed variables given the hidden state, the expected information gain of both Θ and $\Phi:=(\Phi_1,\Phi_2,\dots)$ for a given n and any untried value of x is most conveniently calculated as

$$\begin{aligned}\mathrm{I}(Y_{n,x,a,b};\Theta,\Phi \mid \mathbf{y}) &= \mathrm{H}(Y_{n,x,a,b} \mid \mathbf{y}) - \underbrace{\mathrm{E}[\mathrm{H}(Y_{n,x,a,b} \mid \Theta,\Phi,\mathbf{y})]}_{=0}\\ &= \frac{1}{n-m_n}\log 2^n + h\left(\frac{1}{n-m_n}\right), \qquad (10)\end{aligned}$$

where m_n denotes the number of incorrect values of Φ_n already tried, and where we have split the computation of the entropy into two cases according as the outcome is -1 or ≥ 0 (the entropy of the first case is zero, the entropy of the latter case is $\log 2^n$, and the

[6] We do have not have a rigorous proof of this result, but it is what happens in simulations.

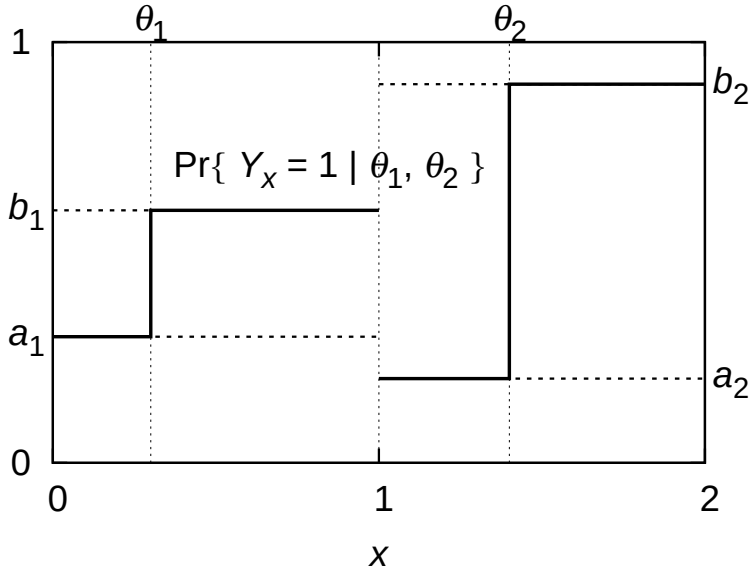


Figure 4: A two-threshold version of the model shown in Fig. 2.

uncertainty between these two cases corresponds to the latter term of the expression). Initially $m_n = 0$ for all n, in which case (10) is maximized by $n = 2$. This maximum value is 2 bits (.5 chance of guessing the correct value of Φ_2, in which case two bits of Θ are obtained, and in either case, the value of $\Phi_2 \in \{1, 2\}$ is learned which yields another bit). Observing this variable does not change the optimal informativity for $n \neq 2$ nor does it change it for $n = 2$ as the optimal gain is still 2 bits after the value of Φ_2 has been learned. Thus, if $\mathrm{I}(Y_{n,x,a,b}; \Theta, \Phi)$ is maximized at every step, only trials with $n = 2$ will ever be presented, each yielding exactly 2 bits of new information.

However, the optimal measurement strategy in this model strongly depends on the total number of trials that the experiment will include. If there are T trials, then a good strategy would be to spend (at most) $T/2$ trials to find the value of $\Phi_{T/2}$ after which the remaining trials will each yield $T/2$ bits of information. The total information gain will thus be at least $(T/2)(T/2) = T^2/4$ bits while the greedy one-step strategy will only yield $2T$ bits.

2.4 Inappropriate utility function

Example 2.4. Suppose Y_x is a dichotomous random variable depending on two thresholds $\Theta_1 \in (0, 1)$ and $\Theta_2 \in (1, 2)$ through the conditional distribution

$$\Pr\{Y_x = 1 \mid \theta_1, \theta_2\} = \begin{cases} a_1, & x < \theta_1,\ x \in (0,1), \\ b_1, & x \geq \theta_1,\ x \in (0,1), \\ a_2, & x < \theta_2,\ x \in (1,2), \\ b_2, & x \geq \theta_2,\ x \in (1,2), \end{cases}$$

see Fig. 4.

In this model, there are essentially two independent subproblems, the estimation of Θ_1 and the estimation of Θ_2. Indeed,

$$\mathrm{I}(Y_x; \Theta_1, \Theta_2 \mid \mathbf{y}) = \begin{cases} \mathrm{I}(Y_x; \Theta_1 \mid \mathbf{y}_1), & x \in (0,1), \\ \mathrm{I}(Y_x; \Theta_2 \mid \mathbf{y}_2), & x \in (1,2), \end{cases}$$

where $\mathbf{y}_1$ denotes the trial results placed on $(0, 1)$ and $\mathbf{y}_2$ those placed on $(1, 2)$, and so we have two independent instances of the single-threshold problem solved in Sec. 2.1. As the optimal gain (6) for each subproblem only depends on the values of a_1, b_1 or a_2, b_2, one of three things will happen: either the inequeality

$$\max_{x\in(1,2)} \mathrm{I}(Y_x; \Theta_1, \Theta_2 \mid \mathbf{y}) > \max_{x\in(0,1)} \mathrm{I}(Y_x; \Theta_1, \Theta_2 \mid \mathbf{y})$$

always holds, it always holds in the reverse direction, or equality always holds. In the first mentioned case, if one uses the joint entropy $\mathrm{H}(\Theta_1, \Theta_2 \mid \mathbf{y})$ as the loss function, then only the second threshold will ever be estimated. This is intuitively not the desired result.

However, even though Paninski (2005) presents this example as a failure of the greedy information maximization strategy, that is not the true cause of the problem as the greedy strategy is in fact optimal for minimizing the specified loss function. Instead, the true problem is inappropriateness of the loss function.

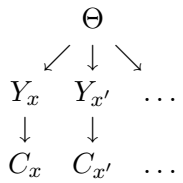
To avoid the problem, one might instead use the loss function

$$\max\{\mathrm{H}(\Theta_1 \mid \mathbf{y}), \mathrm{H}(\Theta_2 \mid \mathbf{y})\}.$$

In that case, both thresholds would be estimated to the same accuracy, but the proportion of trials spent on the more difficult-to-estimate threshold would be larger. This is intuitively the desired result.

3 Random cost of observation

In this section, we consider the situation where the observation of Y_x is associated with some random cost $C_x \geq 0$, which given the value of Y_x, is independent of Θ and the results and costs of any other observations:



The technical requirement that C_x depends on Θ only through Y_x is satisfied in particular if C_x is a component of Y_x. Thus, it leads to no loss of generality if the incurred costs are observable.

Costs of observation have been considered by DeGroot (1970) in several examples and Paninski (2005) mentions a special case where a certain price $C(x_{t-1}, x_t)$ has to be paid for each change of the state of the "observational apparatus" from x_{t-1} to x_t. What is common to all these examples is that the proposed objective function is based on some *difference* of the expected gain and the cost. However, there is the obvious problem of having to equate the units of gain with the units of the cost. In the following, we take a different approach by considering the *ratio* of the gain and cost instead.

3.1 Obtaining the best value for money

Intuitively, we would like to maximize the unit price of the information, the amount of information given by the experiment divided by the total cost of conducting it. However, in some practically interesting situtations there may be a positive probability that the actual cost is zero (even if its expectation is positive), and therefore it generally does not

make sense to talk about the expected unit price over one trial or any predetermined number of trials as there will be a positive probability that the total cost is zero and so the expectation of the total gain divided by the total cost will be infinite.

To avoid this instability, the goal can be operationally defined either as

1. maximizing the expected amount of information given by an experiment that terminates when the total cost overruns a certain predetermined budget, or

2. minimizing the expected cost of an experiment that terminates when a predetermined amount of information has been obtained.

Both definitions are reasonable, but the first one is more elegant in that it corresponds to the plain expected information maximization goal with an adaptive termination rule as dicussed in Sec. 1.6. Hence, the optimal strategy under that goal is insensitive to the parametrization used to define the differential entropy measure of information. In contrast, simple counterexamples show that the second definition does not have this desirable property.

Remark 3.1. In the statement of the problem, we do not require the cost C_x to be observable. However, for the experiment to actually terminate when the budget is overrun, either the actual costs must be observable, or alternatively, any further trials could simply fail after the budget is overrun. In the latter case, the adaptive termination rule would be random. In either case, the actual costs are irrelevant to the final Bayesian estimates of Θ as they are assumed to depend on Θ only through the fully observable results Y_x.

Remark 3.2. Obviously exact maximization of the expected information gain under a given budget is generally intractable for the same reasons that the usual constant number of trials case is. However, even if the information gains and costs associated with each Y_x were known time-invariant constants, the problem of fitting the best value in a given constant budget would still be intractable — it is equivalent to the knapsack problem which is NP hard (see, Garey and Johnson, 1979). The heuristic we shall present is in fact similar to the heuristics used to find approximate solutions to the knapsack problem although we have the additional complications of randomness and the generally intractable sequential changes.

3.2 Heuristics

Let us define random variables denoting the gain and cost of the t-th observation:

$$G_t := U_t - U_{t-1} = u[\,\theta \mapsto p(\theta \mid Y_{X_1}, \ldots, Y_{X_t})\,] - u[\,\theta \mapsto p(\theta \mid Y_{X_1}, \ldots, Y_{X_{t-1}})\,],$$
$$C_t := C_{X_t}.$$

Assuming for a moment that the cost C_x is defined as the time taken by the observation of Y_x, one might think that choosing x so as to maximize the expected *rate* of information gain

$$\mathrm{E}\left(\frac{G_t}{C_t} \;\middle|\; \mathbf{y},\ X_t = x\right) \tag{11}$$

over the *duration* C_t of the next observation would be a good heuristic. While this formulation indeed yields the best unit price over the next trial (ignoring the potential division by zero problem), it generally falls short of this goal in a sequential experiment. Over a *constant* unit of time, repeated i.i.d. observations of Y_x are expected to result in each outcome (y_x, c_x) being observed for a total duration proportional to $p(y_x, c_x \mid \mathbf{y})c_x$. Thus,

to estimate the average rate of gain obtainable from Y_x, the expectation should be taken over the distribution

$$\frac{p(y_x, c_x \mid \mathbf{y})c_x}{\iint p(y_x, c_x \mid \mathbf{y})c_x \, dy_x \, dc_x}$$

instead. This leads to the objective function

$$\iint \left[\frac{u_{\mathbf{y},y_x} - u_{\mathbf{y}}}{c_x}\right] \frac{p(y_x, c_x \mid \mathbf{y})c_x}{\iint p(y_x, c_x \mid \mathbf{y})c_x \, dy_x \, dc_x} dy_x \, dc_x = \frac{\mathrm{E}(G_t \mid \mathbf{y},\ X_t = x)}{\mathrm{E}(C_t \mid \mathbf{y},\ X_t = x)}, \tag{12}$$

where we denote $u_{\mathbf{y}} = u[\,\theta \mapsto p(\theta \mid \mathbf{y})\,]$. With the entropy utility function (1), this can be written in the convenient form

$$\frac{\mathrm{I}(Y_x; \Theta \mid \mathbf{y})}{\mathrm{E}(C_x \mid \mathbf{y})}. \tag{13}$$

Remark 3.3. Unlike the situation in the pure information maximization case, maximization of (12) does not correspond to maximization of the immeadiate expected utility. Thus, it is not the prototypical greedy algorithm, but it is still myopic in the sense that it expects the future sets of possible expected gains and costs to be similar to those of the current trial.

3.3 Conditions for asymptotic optimality

The following proposition implies that maximization of (12) is the asymptotically optimal strategy if the set of the distributions of $[G_t \mid \mathbf{y},\ X_t = x]$ and $[C_t \mid \mathbf{y},\ X_t = x]$ over all possible values of x do not change over time as each new outcome is added to the data $\mathbf{y}$. That is, the same distributions of gain and cost are allowed to be associated with different x at different times as long as these x's at different times are in a one-to-one correspondence.

Proposition 3.1. *Suppose that the random variables G and C have finite expectations $\mathrm{E}(G)$ and $\mathrm{E}(C) \neq 0$. Then, almost surely (i.e., with probability 1)*

$$\lim_{n\to\infty} \frac{G_1 + \cdots + G_n}{C_1 + \cdots + C_n} = \frac{\mathrm{E}(G)}{\mathrm{E}(C)},$$

where G_k and C_k denote i.i.d. copies of G and C, respectively.

Proof. Having finite expectations, the i.i.d. sequences G_k and C_k satisfy the strong law of large numbers:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (G_k - \mathrm{E}(G_k)) \overset{\text{a.s.}}{=} 0,$$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} (C_k - \mathrm{E}(C_k)) \overset{\text{a.s.}}{=} 0.$$

Thus,

$$\lim_{n\to\infty} \frac{\sum_{k=1}^{n} G_k}{\sum_{k=1}^{n} C_k} = \lim_{n\to\infty} \frac{\frac{1}{n}\sum_{k=1}^{n}(G_k - \mathrm{E}(G_k)) + \mathrm{E}(G)}{\frac{1}{n}\sum_{k=1}^{n}(C_k - \mathrm{E}(C_k)) + \mathrm{E}(C)} \overset{\text{a.s.}}{=} \frac{\mathrm{E}(G)}{\mathrm{E}(C)}.$$

□

Under certain side conditions, optimality of the strategy can also be shown under the weaker assumption that the objective function always has the same maximum value α regardless of the past data:

Theorem 3.2. *Suppose that for some* $\sigma^2 < \infty$,

$$\begin{cases} \operatorname{Var}(G_t \mid \mathbf{y},\ X_t = x) < \sigma^2, \\ \operatorname{Var}(C_t \mid \mathbf{y},\ X_t = x) < \sigma^2, \end{cases}$$

for all sets $\mathbf{y}$ *of past observations and all placements* x *of the next trial and*

$$\liminf_{t\to\infty} \frac{C_1 + \cdots + C_t}{t^\rho} \overset{a.s.}{>} 0$$

for some $\rho > 1/2$. *If there exists a constant value* $\alpha > 0$ *such that*

$$\max_{x \in \mathsf{X}_t} \frac{\mathrm{E}(G_t \mid \mathbf{y},\ X_t = x)}{\mathrm{E}(C_t \mid \mathbf{y},\ X_t = x)} = \alpha \tag{14}$$

for all sets $\mathbf{y}$ *of past observations, then defining* X_t *as the maximizer of* (14) *is asymptotically optimal in the sense that with this strategy, the distribution of*

$$\frac{G_1 + \cdots + G_t}{C_1 + \cdots + C_t}$$

converges to the constant α *as* $t \to \infty$, *while with any other strategy,*

$$\lim_{t\to\infty} \Pr\left\{ \frac{G_1 + \cdots + G_t}{C_1 + \cdots + C_t} \le \alpha + \epsilon \right\} = 1$$

for all $\epsilon > 0$.

Proof. Assuming that X_t is always chosen as the maximizer of (14), the sequences $Y_t := Y_{X_t}$ and $Z_t := G_t - \alpha C_t$ satisfy the assumptions of Lemma A.2 with $\mu_k = 0$ and $\sigma_k^2 = 4\sigma^2$. Thus,

$$\operatorname{Var}(Z_1 + \cdots + Z_t) \le 4t\sigma^2,$$

whence

$$\operatorname{Var}\left(\frac{Z_1 + \cdots + Z_t}{t^\rho} \right) \le 4t^{1-2\rho}\sigma^2 \to 0$$

as $t \to \infty$, which implies that the distribution of $(Z_1 + \cdots + Z_t)/t^\rho$ converges to the constant $\mathrm{E}[(Z_1 + \cdots + Z_t)/t^\rho] = 0$. Expanding the definition of Z_t, this means

$$\frac{G_1 + \cdots + G_t}{t^\rho} - \alpha \frac{C_1 + \cdots + C_t}{t^\rho} \xrightarrow{\mathrm{P}} 0.$$

Division by $(C_1 + \cdots + C_t)/t^\rho$, which is almost surely larger than some positive constant for all but a finite number of indices t, yields

$$\frac{G_1 + \cdots + G_t}{C_1 + \cdots + C_t} \xrightarrow{\mathrm{P}} \alpha.$$

To prove the latter part of the theorem, let the logic of choosing X_t now be arbitrary. The sequence $Z'_t := G'_t - \alpha C_t$, where we use the upwards adjusted gains

$$G'_t := G_t + \underbrace{\mathrm{E}(\alpha C_t - G_t \mid Y_{X_1}, \ldots, Y_{X_t})}_{\ge 0 \text{ by } (14)},$$

satisfies the assumptions of the lemma with the same constants as the optimal case, and repeating the analogous steps, we obtain

$$\frac{G_1 + \cdots + G_t}{C_1 + \cdots + C_t} \le \frac{G'_1 + \cdots + G'_t}{C_1 + \cdots + C_t} \xrightarrow{\mathrm{P}} \alpha.$$

□

3.4 The simplified psychometric model

We shall extend the model of Sec. 2.1 with random costs C_x associated with observing Y_x. Recall that in this model, for all $x \in \mathbb{R}$, $Y_x \in \{0,1\}$ is a dichotomous random variable defined by

$$\Pr\{Y_x = 1 \mid \theta\} = \begin{cases} a, & x < \theta, \\ b, & x \geq \theta, \end{cases}$$

for some $a, b \in [0,1]$. Let us then assume that the distribution of the random cost C_x is fully determined by the value of Y_x, i.e., it does not depend directly on x. Then,

$$\mathrm{E}(C_x \mid \theta) = \begin{cases} c_a, & x < \theta, \\ c_b, & x \geq \theta \end{cases}$$

for some c_a and c_b, which we assume to be positive to avoid pathological cases.

3.4.1 Optimal strategy under varying costs

The expected information gain, as before, is given by

$$\begin{aligned} \mathrm{I}(Y_x; \Theta \mid \mathbf{y}) &= \mathrm{H}(Y_x \mid \mathbf{y}) - \mathrm{E}[\mathrm{H}(Y_x \mid \Theta,\ \mathbf{y})] \\ &= h[a + (b-a)\Pr\{\Theta \leq x \mid \mathbf{y}\}] - [h(a) + (h(b) - h(a))\Pr\{\Theta \leq x \mid \mathbf{y}\}] \end{aligned}$$

and so the objective function is

$$\frac{\mathrm{I}(Y_x; \Theta \mid \mathbf{y})}{\mathrm{E}(C_x \mid \mathbf{y})} = \frac{h(a + rz) - (h_a + h_r z)}{c_a + c_r z},$$

where we denote $z := \Pr\{\Theta \leq x \mid \mathbf{y}\}$, $r := b - a$, $h_a := h(a)$, $h_r := h(b) - h(a)$, and $c_r := c_b - c_a$. Thus, the objective function depends on the prior and the placement x only through z, which can attain any value in $[0,1]$ provided that the prior distribution is absolutely continuous w.r.t. the Lebesgue measure. The function is zero for $z \in \{0,1\}$ and positive for $z \in (0,1)$. Therefore, the optimum can be found at a critical point of the objective function.

Differentiating w.r.t. z yields

$$\frac{(h'(a+rz)d - h_r)(c_a + c_r z) - (h(a+rz) - h_a - h_r z)c_r}{(c_a + c_r z)^2}$$

The denominator is finite and positive and the numerator equals

$$\begin{aligned} & c_r h_a - c_a h_r + r h'(a+rz)(c_a + c_r z) - c_r h(a+rz) \\ &= c_r h_a - c_a h_r + r[\log(1-a-rz) - \log(a+rz)](c_a + \cancel{c_r z}) \\ &\quad + c_r[(a + \cancel{rz})\log(a+rz) + (1-a-\cancel{rz})\log(1-a-rz)] \\ &= c_r h_a - c_a h_r + (c_r a - c_a r)\log\omega + (c_a r - c_r a + c_r)\log(1-\omega) \\ &= c_b h(a) - c_a h(b) + (c_b a - c_a b)\log\omega + (c_b(1-a) - c_a(1-b))\log(1-\omega), \end{aligned}$$

where we denote $\omega := a + rz = \Pr\{Y_x = 1 \mid \mathbf{y}\}$. It is easy to verify by differentiating and simplifying that the above expression is monotone for ω between a and b and therefore has a unique zero corresponding to the maximum value of the objective function.

Placing the next trial at this optimum is by Theorem 3.2 the asymptotically optimal strategy as it yields the same maximum value of the objective function at every trial regardless of the past data (and it is easy to show that the other assumptions of the theorem hold, too).

3.4.2 Specific examples

Assuming now $C_x = 1$, i.e., constant cost, we have $c_a = c_b = 1$ and the optimality condition becomes

$$h(a) - h(b) + (a-b)\log\omega + (b-a)\log(1-\omega) = 0,$$

whence

$$\log\frac{1-\omega}{\omega} = \frac{h(b)-h(a)}{b-a},$$

which yields the optimizer

$$\omega^* = \frac{1}{1+\exp\left(\dfrac{h(b)-h(a)}{b-a}\right)}.$$

This result we have already seen in Sec. 2.1, although here we have adopted ω instead of $z = (w-a)/(b-a)$ as the parameter.

If we assume instead that $C_x = [Y_x = 0]$, i.e., each 0-result costs one unit, we have $c_a = 1-a$ and $c_b = 1-b$, and the equation becomes

$$\begin{aligned} 0 &= (1-b)h(a) - (1-a)h(b) \\ &\quad + ((1-b)a - (1-a)b)\log\omega \\ &\quad + ((1-b)(1-a) - (1-a)(1-b))\log(1-\omega) \\ &= (1-b)h(a) - (1-a)h(b) + (a-b)\log\omega, \end{aligned}$$

which yields the optimizer

$$\omega^* = \exp\left(-\frac{(1-a)h(b) - (1-b)h(a)}{b-a}\right).$$

This definition of cost was in fact used by Kujala et al. (2008) in a child-friendly mesaurement formulation, which assumes a cost on each failure of a child due the fact that failures can lower motivation. If a child can only tolerate a certain number of failures, then this formulation should yield the maximum amount of information before that limit is reached.

3.4.3 Non-adaptive and batch strategies

The discussion of Sec. 2.1 generalizes directly. Denoting

$$\mathrm{I}(Y_{x_1}, \ldots, Y_{x_n}; \Theta) = f(z) = -(Mz+v)\cdot\log(Mz+v) - u\cdot z - c,$$

and

$$E(C_x) = g(z) = c_a + (c_b - c_a)\sum_{k=1}^{n} z_k,$$

we have

$$\nabla(f/g)(z) = \frac{g(z)\nabla f(z) - f(z)\nabla g(z)}{g(z)^2}$$

and

$$H(f/g)(z) = \frac{1}{g(z)^4}\Big[(\nabla g(z)\nabla f(z)^T + g(z)Hf(z) - \nabla f(z)\nabla g(z)^T - f(z)Hg(z))g(z)^2 \\ - (g(z)\nabla f(z) - f(z)\nabla g(z))2g(z)\nabla g(z)^T\Big] \\ = \frac{Hf(z)}{g(z)} - \frac{\nabla g(z)}{g(z)}\frac{\nabla f(z)^T}{g(z)} - \frac{\nabla f(z)}{g(z)}\frac{\nabla g(z)^T}{g(z)} + 2\frac{f(z)}{g(z)}\frac{\nabla g(z)}{g(z)}\frac{\nabla g(z)^T}{g(z)},$$

where $\nabla g(z) = (c_b - c_a, \ldots, c_b - c_a)$ and $Hg(z) = 0$.

Although it is not obvious from this expression, the Hessian of the objective function f/g was in practice always negative definite and so Newton's iteration worked fine here, too. However, we have no formal proof that the found optimum is in fact the global optimum when $n > 1$.

We do not go into any details of the single placement strategy $x_1 = \cdots = x_n$ here either.

3.4.4 Comparison

Figure 5 illustrates the optimal placements under the three strategies when each 0-result costs one unit.

Comparing the optimal adaptive placements in Figs. 3 and 5 supports the characterization given in (Kujala et al., 2008): while pure information maximization works much like binary search, roughly bisecting the uncertainty distribution at each step, the cost-aware variation instead chooses the placement at a certain lower percentile closer to the easier end.

The exact percentile of the optimal placement seems to depend mostly on the lapsing rate $1 - b$ and less on the guessing rate a. This dependence was to be expected: if there is going to be a large probability of careless mistakes anyway, then it does not pay off to make the trials very easy, and conversely, if careless mistakes are unlikely, then the easiest trials will be virtually free and the placements close to that end will yield the best value for money even though the gains over one trial are smaller.

The placements of the non-adaptive strategy are no longer close to evenly spaced. Instead, they cluster near the easier end.

3.5 Discussion

So far, we have only considered discrete cost variables explicitly. An obvious topic for future work is calculation of the objective function (13) for some response time model with the cost C_x defined as the response time. In any experiments where the placement of a trial can affect its duration, this formulation can increase the efficiency per time unit over the pure information maximization greedy algorithm. In particular, in an n-choice task, the response times generally increase with n, and so a smaller value of n might turn out to be optimal even though it yields a higher guessing rate. However, the resulting time-efficiency also depends on any pre-stimulus delays, which should be included in the cost.

A Lemmas

In this section, we give proofs for some intuitively true results used in the main text.

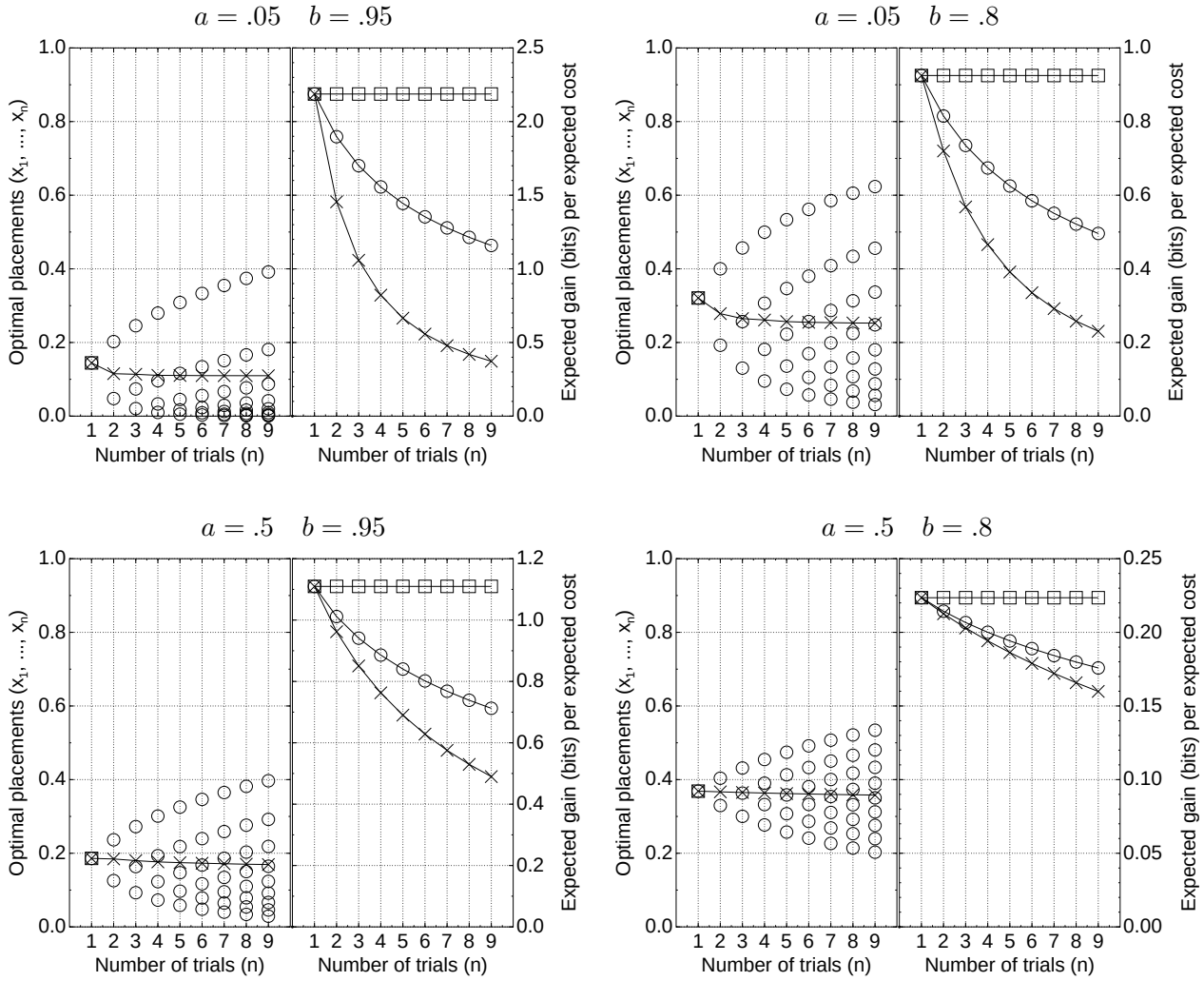


Figure 5: Optimal placements and the corresponding optimal expected gain per exepcted cost ratios for the optimal adaptive strategy (squares), optimal non-adaptive strategy (circles), and optimal single-placement repeated observations strategy (crosses) under random cost of observation defined as $C_x = [Y_x = 0]$ (i.e., each 0-result costs one unit) in the model of Figs. 2 and 3. The interpretation of the placement values is the same is in Fig. 3.

Lemma A.1. *Let X be a real-valued random variable and let the functions $f, g : [a, b] \to \mathbb{R}$ be continuous and twice continuously differentiable on (a, b). If $\Pr\{X \in [a, b]\} = 1$ and $f''(x) \leq g''(x)$ for all $x \in (a, b)$, then*

$$\mathrm{E}[f(X)] - f(\mathrm{E}[X]) \leq g[f(X)] - g(\mathrm{E}[X]).$$

The analogous result holds for "$<$" provided that X is not concentrated on a point.

Proof. Suppose $x_0 := \mathrm{E}[X] \in (a, b)$ (otherwise X is concentrated on either a or b and the result is trivial). Then, $f''(x) \leq g''(x)$ implies

$$f'(x) - f'(x_0) = \int_{x_0}^{x} f'' \leq \int_{x_0}^{x} g'' = g'(x) - g'(x_0)$$

for $x \in (x_0, b)$ and the inequality is reversed for $x \in (a, x_0)$. Integrating both sides again yields

$$\int_{x_0}^{x} [f'(t) - f'(x_0)]dt \leq \int_{x_0}^{x} [g'(t) - g'(x_0)]dt$$

for all $x \in (a, b)$, which implies

$$f(x) - f(x_0) - (x_0 - x)f'(x_0) \leq g(x) - g(x_0) - (x_0 - x)g'(x_0)$$

for all $x \in [a, b]$ (the inequeality extends to the endpoints as every term is continuous in x). Substituting the random variable X for x and taking the expectation of both sides, we obtain

$$\mathrm{E}[f(X)] - f(x_0) - (x_0 - \mathrm{E}[X])f'(x_0) \leq \mathrm{E}[g(X)] - g(x_0) - (x_0 - \mathrm{E}[X])g'(x_0),$$

which implies the statement. The "$<$" version is a simple modification. □

Lemma A.2. *Suppose Y_k and Z_k are sequences of random variables such that Z_k is independent of $Z_1, \ldots, Z_{k-1}$ given $y_1, \ldots, y_{k-1}$ and*

$$\begin{cases} \mathrm{E}(Z_k \mid y_1, \ldots, y_{k-1}) = \mu_k, \\ \mathrm{Var}(Z_k \mid y_1, \ldots, y_{k-1}) \leq \sigma_k^2, \end{cases}$$

for all values of k and $y_1, \ldots, y_{k-1}$. Then,

$$\mathrm{Var}(Z_1 + \cdots + Z_k) \leq \sigma_1^2 + \cdots + \sigma_k^2$$

for all k.

Proof. By induction: for $k = 1$, we have trivially $\mathrm{Var}(Z_1) \leq \sigma_1^2$, and for any $k > 1$, the variance can be split as

$$\begin{aligned} \mathrm{Var}(Z_1 + \cdots + Z_k) = {} & \mathrm{Var}(Z_1 + \cdots + Z_{k-1}) + \mathrm{Var}(Z_k) \\ & + 2\,\mathrm{Cov}(Z_k, Z_1 + \cdots + Z_{k-1}), \end{aligned}$$

where our induction assumption is

$$\mathrm{Var}(Z_1 + \cdots + Z_{k-1}) \leq \sigma_1^2 + \cdots + \sigma_{k-1}^2.$$

Denoting $\mathbf{y} = (y_1, \ldots, y_{k-1})$, the assumptions on the conditional statistics yield

$$\begin{aligned}\operatorname{Var}(Z_k) &= \int \mathrm{E}(Z_k^2 \mid \mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \left(\int \mathrm{E}(Z_k \mid \mathbf{y}) p(\mathbf{y}) d\mathbf{y} \right)^2 \\ &= \int \left[\operatorname{Var}(Z_k \mid \mathbf{y}) + \mu_k^2 \right] p(\mathbf{y}) d\mathbf{y} - \mu_k^2 \leq \sigma_k^2,\end{aligned}$$

which gives the induction step as $\operatorname{Cov}(Z_k, Z_1 + \cdots + Z_{k-1})$ equals

$$\begin{aligned}&\int \mathrm{E}(Z_k (Z_1 + \cdots + Z_{k-1}) \mid \mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \mathrm{E}(Z_k)\, \mathrm{E}(Z_1 + \cdots + Z_{k-1}) \\ = &\int \mathrm{E}(Z_k \mid \mathbf{y})\, \mathrm{E}(Z_1 + \cdots + Z_{k-1} \mid \mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \mathrm{E}(Z_k)\, \mathrm{E}(Z_1 + \cdots + Z_{k-1}) \\ = &\, \mu_k \int \mathrm{E}(Z_1 + \cdots + Z_{k-1} \mid \mathbf{y}) p(\mathbf{y}) d\mathbf{y} - \mu_k\, \mathrm{E}(Z_1 + \cdots + Z_{k-1}) = 0. \qquad \square\end{aligned}$$

B Technicalities

In the main text, we have avoided all measure-theoretic techinicalities and implicitly assumed certain regularity conditions. In this section we make these assumptions explicit. The same conditions must have been implicitly assumed in other works, too, as without them the mutual information $\mathrm{I}(Y_x; \Theta)$ does not make sense and even Bayes' theorem does not generally hold.

B.1 Preliminaries

Let $(\Omega, \mathcal{F}, \Pr)$ be a probability space. A *random variable* is a measurable mapping $X : \Omega \to \mathsf{X}$ to some measurable space $(\mathsf{X}, \mathcal{X})$ (usually the real line $\mathbb{R}$ equipped with the Borel σ-algebra $\mathcal{B}(\mathbb{R})$). The *distribution* of the random variable is the measure $P_X : S \mapsto \Pr(X^{-1}(S))$ induced on $\mathcal{X}$.

If $X : \Omega \to \mathsf{X}$ and $Y : \Omega \to \mathsf{Y}$ are random variables such that $P_{X,Y} \ll \mu \times \nu$ for some σ-finite measures $\mu : \mathcal{X} \to [0, \infty]$ and $\nu : \mathcal{Y} \to [0, \infty]$, then X and Y are said to have a *joint density* $p = dP_{X,Y}/d(\mu \times \nu)$. By Fubini's theorem, the marginal distributions can then be written as

$$\begin{aligned}P_X(U) &= P_{X,Y}(U \times \mathsf{Y}) = \int_{x \in U} \left[\int_y p(x, y) d\nu(y) \right] d\mu(x), \\ P_Y(V) &= P_{X,Y}(\mathsf{X} \times V) = \int_{y \in V} \left[\int_x p(x, y) d\mu(x) \right] d\nu(y),\end{aligned}$$

which implies the existence of the marginal densities

$$\begin{aligned}p_X &= \frac{dP_X}{d\mu} = \int_y p(\cdot, y) d\nu(y), \\ p_Y &= \frac{dP_Y}{d\nu} = \int_x p(x, \cdot) d\mu(x).\end{aligned}$$

For brevity, we leave out the subscript of the density when it matches the arguments, i.e, instead of $p_X(x)$, we write simply $p(x)$. When there is no confusion about the dominating measure μ, we will write dx instead of $d\mu(x)$ (this is done everywhere in the main text, in particular, the differential entropy can be w.r.t. any measure, not just the Lebesgue measure).

A *transition measure* from $(\mathsf{Y}, \mathcal{Y})$ to $(\mathsf{X}, \mathcal{X})$ is any function $\mu : \mathsf{Y} \times \mathcal{X} \to [0, \infty]$ satisfying the following axioms:

1. for any $y \in \mathsf{Y}$, the function $S \mapsto \mu(y, S)$ is a measure on $\mathcal{X}$,
2. for any $S \in \mathcal{X}$, the function $y \mapsto \mu(y, S)$ is $\mathcal{Y}$-measureable.

The product of a transition measure $\mu : \mathsf{Y} \times \mathcal{X} \to [0, \infty]$ and a σ-finite measure $\nu : \mathcal{Y} \to [0, \infty]$ is given by

$$(\mu \times \nu)(S) := \int \mu(y, S_y) d\nu(y)$$

for all $S \in \mathcal{X} \otimes \mathcal{Y}$, where $S_y := \{x : (x, y) \in S\}$. The product is a measure on $\mathcal{X} \otimes \mathcal{Y}$. If a transition measure $P_{X|Y} : \mathsf{Y} \times \mathcal{X} \to [0, \infty]$ satisfying

$$P_{X,Y} = P_{X|Y} \times P_Y$$

exists, then it is called a *conditional distribution* of X given Y. We will also use the shorthand $P_{X|y} := P_{X|Y}(y, \cdot)$. Note that a conditional distribution always exists for a random variable in $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, $(\mathbb{R}^\infty, \mathcal{B}(\mathbb{R}^\infty))$, or any other complete separable metric space, but there are spaces where its existence is not guaranteed (Shiryaev, 1996).

If a conditiondal distribution $P_{X|Y}$ exists and satisfies

$$P_{X|y}(S) = \int_S p(x \mid y) d\mu(x)$$

for all $S \in \mathcal{X}$, $y \in \mathsf{Y}$ for some measurable function $(x, y) \mapsto p(x \mid y)$ and some measure μ (which need not be σ-finite), then $p(x \mid y)$ is called a *conditional density* of X given y. If the joint density $p = dP_{X,Y}/d(\mu \times \nu)$ exists for some σ-finite μ and ν, then a conditional density can always be obtained by

$$p(x \mid y) := \begin{cases} \dfrac{p(x, y)}{p(y)}, & p(y) > 0, \\ 0, & p(y) = 0. \end{cases}$$

(The value chosen for $p(y) = 0$ is immaterial as the conditional density is only determined $\mu \times P_Y$-a.e.)

B.2 Regularity conditions for Bayesian estimation

The following theorem gives a set of equivalent conditions under which we can avoid the potential problems of nonexistent distributions or densities.

Theorem B.1. *Let $X : \Omega \to \mathsf{X}$ and $Y : \Omega \to \mathsf{Y}$ be random variables. Then, the following are equivalent:*

1. *X and Y have a joint density (which is by definition w.r.t. a σ-finite product measure $\mu \times \nu$),*
2. *$P_{X,Y} \ll P_X \times P_Y$,*
3. *X has a conditional density $p(x \mid y)$ w.r.t. a σ-finite measure μ,*
4. *X has a conditional distribution $P_{X|Y}$ such that $P_{X|y} \ll P_X$ for all y,*

5. *X has a conditional distribution $P_{X|Y}$ and a marginal density $p(x)$ w.r.t. a (not necessarily σ-finite) measure μ such that $P_{X|y} \ll \mu$ for all y.*

Furthermore,

6. *if the above conditions hold for X and Y, then they also hold for $X' = F(X)$ and $Y' = G(Y)$ where $F : \mathsf{X} \to \mathsf{X}'$ and $G : \mathsf{Y} \to \mathsf{Y}'$ are any measurable functions.*

Remark B.1. The conditions of this theorem are precisely those under which Bayes' formula

$$p(x \mid y) = \frac{p(y \mid x)p(x)}{\int p(y \mid x)p(x)dx}$$

for conditional densities holds (it derives from condition 1 and implies condition 4).

Remark B.2. These conditions are also precisely those under which the Radon-Nikodým derivative in the measure-theoretic definition of the mutual information

$$\mathrm{I}(X;Y) = \int dP_{X,Y} \log \frac{dP_{X,Y}}{d(P_X \times P_Y)}$$

exists[7] (condition 2). Therefore, even though one might be able to work with conditional distributions directly, if one's utility function is the information gain, then Y_x and Θ in the main text must still satisfy these conditions for all x for the problem to be well-defined.

Remark B.3. If Y_x and Θ satisfy these conditions for all x, then one can apply Bayes' formula to any finite set $\mathbf{y} = \{y_{x_t}\}_{t=1}^{T}$ of results sequentially. This implies that $P_{\Theta|\mathbf{y}} \ll P_\Theta$ for all $\mathbf{y}$ (condition 4) and as this condition makes no reference to the distribution of $\mathbf{y}$, it follows that regardless of the decision function d (which determines the distribution of the placements $X_1, \ldots, X_T$), the whole-experiment outcome variable Y_d (which has a generally random number T of individual observations) has a joint density with Θ as long as the experiment terminates with probability one. Thus, the expected information gain $\mathrm{I}(\Theta; Y_d)$ for the whole experiment is formally well-defined (although its value may still come out ∞). However, if there is a positive probability that the experiment does not terminate, then it is possible that no joint density of Θ and Y_d exists, even for constant placements (Example B.2 below).

Proof. **2 ⇒ 5:** Using the joint density $p := dP_{X,Y}/d(P_X \times P_Y)$, we obtain the induced marginal density $p(x)$ w.r.t. the measure P_X and the conditional density $p(x \mid y)$, which induces a conditional distribution $P_{X|y} \ll P_X$.

5 ⇒ 4: Denoting $N := \{\, x \in \mathsf{X} : p(x) = 0 \,\}$, we have

$$0 = \int_N p(x)d\mu(x) = P_X(N) = \int P_{X|y}(N)dP_Y(y),$$

which implies $P_{X|y}(N) = 0$ for P_Y-a.e. y. However, as $P_{X|y}$ is only determined for P_Y-a.e. y, we are free to modify it so that $P_{X|y}(N) = 0$ for all y. We will show that this $P_{X|y}$ is dominated by P_X for all y. Let $S \in \mathcal{X}$ be such that $P_X(S) = 0$. Then, we have

$$0 = P_X(S \setminus N) = \int_{S\setminus N} \underbrace{p(x)}_{>0} d\mu(x),$$

which implies $\mu(S \setminus N) = 0$. As $P_{X|y} \ll \mu$, we have $P_{X|y}(S \setminus N) = 0$, but as also $P_{X|y}(N) = 0$, we obtain $P_{X|y}(S) = 0$. Thus, $P_{X|y} \ll P_X$ for all y.

[7] In case $P_{X,Y}$ is singular w.r.t. $P_X \times P_Y$, Kolmogorov (1956) defines $\mathrm{I}(X;Y) = \infty$.

4 ⇒ 3: Choose $\mu = P_X$.

3 ⇒ 1: By the definition of conditional density, we have

$$P_{X,Y}(S) = \int_S p(x \mid y) d\mu(x) dP_Y(y).$$

Thus, $p(x \mid y)$ is a joint density of X and Y w.r.t. the σ-finite measure $\mu \times P_Y$.

1 ⇒ 2: Suppose that $p = dP_{X,Y}/d(\mu \times \nu)$ exists for some σ-finite measures μ and ν and let $S \in \mathcal{X} \otimes \mathcal{Y}$ be an arbitrary measurable set such that $(P_X \times P_Y)(S) = 0$. We will show that then $P_{X,Y}(S) = 0$. Denoting

$$\begin{aligned} U &:= \{\, x \in \mathsf{X} : p(x) = 0 \,\}, \\ V &:= \{\, y \in \mathsf{Y} : p(y) = 0 \,\}, \\ N &:= (U \times \mathsf{Y}) \cup (\mathsf{X} \times V), \end{aligned}$$

we have $P_X(U) = 0$ and $P_Y(V) = 0$. Furthermore,

$$0 = (P_X \times P_Y)(S \setminus N) = \int_{S \setminus N} \underbrace{p(x)}_{>0} \underbrace{p(y)}_{>0} d(\mu \times \nu)(x, y)$$

implies that $(\mu \times \nu)(S \setminus N) = 0$, whence $P_{X,Y}(S \setminus N) = 0$. Thus,

$$P_{X,Y}(S) \leq P_{X,Y}(S \setminus N) + \underbrace{P_{X,Y}(U \times \mathsf{Y})}_{=P_X(U)} + \underbrace{P_{X,Y}(\mathsf{X} \times V)}_{=P_Y(V)} = 0.$$

2 ⇒ 6: Suppose that $F : \mathsf{X} \to \mathsf{X}'$ and $G : \mathsf{Y} \to \mathsf{Y}'$ are arbitrary measurable mappings. We show that $P_{X,Y} \ll P_X \times P_Y$ implies $P_{F(X),G(Y)} \ll P_{F(X)} \times P_{G(Y)}$. For any $S \in \mathcal{X} \otimes \mathcal{Y}$,

$$0 = (P_{F(X)} \times P_{G(Y)})(S) = (P_X \times P_Y)(\{(F^{-1}(x), G^{-1}(y)) : (x, y) \in S\})$$

implies

$$0 = P_{X,Y}(\{(F^{-1}(x), G^{-1}(y)) : (x, y) \in S\}) = P_{F(X),G(Y)}(S),$$

where F^{-1} and G^{-1} denote the preimage sets. □

The conditions of the theorem are mild, being satisfied whenever either X or Y is discrete as well as in most practical situations with continuous random variables. However, it precludes in particular the following example.

Example B.1. Suppose that $X = Y \sim \text{Uniform}[0, 1]$. The conditional distribution $P_{X|y}(S) = [y \in S]$ is singular w.r.t. $dP_X(x) = dm_{[0,1]}(x)$, where $m_{[0,1]}$ denotes the restriction of the Lebesgue measure to $[0, 1]$, and thus condition 4 of Theorem B.1 is not satisfied. The conditional density

$$p(x \mid y) = [x = y] := \begin{cases} 1, & x = y \\ 0, & x \neq y \end{cases}$$

exists w.r.t. the counting measure, but this measure is not σ-finite when defined on the measurable subsets of $[0, 1]$, and so this density does not satisfy condition 3. Even though the joint distribution can be written as

$$P_{X,Y}(S) = \int \left[\int_{S_y} [x = y] d\#(x) \right] dm_{[0,1]}(y),$$

where # is the counting measure, the integrand $[x = y]$ does not yield the joint density of condition 1 as the product $\# \times m_{[0,1]}$ does not exist (the product measure is only defined for σ-finite measures). Some authors do define the product measure more generally, but even then, a joint density cannot be obtained as the function $[x = y]$ is not integrable w.r.t. $\# \times m_{[0,1]}$ and so Fubini's theorem does not hold for the iterated integral above.

Example B.2. Suppose that $X \sim \text{Uniform}[0, 1]$ and the random variables $Y_t \in \{0, 1\}$ for $t = 1, 2, \ldots$ are defined as the binary representation of X. Then, although the conditional density $p(x \mid y_1, \ldots, y_T)$ w.r.t. the Lebesgue measure is well-defined for any finite set of observations, the full sequence of results $Y := \{Y_t\}_{t=1}^{\infty}$ cannot have any joint density with X, because by condition 6 of the theorem, that would imply that also the transformed variable

$$Y' := F(Y) := \sum_{t=1}^{\infty} 2^{-t} Y_t$$

would have a joint density with $X = Y'$, which contradicts the negative result of the previous example.

B.3 Generalization

For completeness, we present a generalization of Theorem B.1 to more than two random variables. Although we make no use of it here, it may be useful elsewhere.

To state the generalization, we need another definition.

Definition B.1. A *Bayes network* is a directed acyclic graph representing a dependency structure of a set $X_1, \ldots, X_n$ of random variables. Each random variable X_k is represented by a node whose parents are its conditioning variables $X_{j(k,1)}, \ldots, X_{j(k,n_k)}$, where we can assume WLOG that $j(k, i) < k$ for all $i = 1, \ldots, n_k$ (topological sorting), so that the joint distribution of $X_1, \ldots, X_n$ is given by the product

$$P_{X_1,\ldots,X_n} = \prod_k P_{X_k | X_{j(k,1)},\ldots,X_{j(k,n_k)}},$$

where one can interpret, e.g., $P_{X_k | X_{j(k,1)},\ldots,X_{j(k,n_k)}} = P_{X_k | X_1,\ldots,X_{k-1}}$ and then apply the transition measure product operator.

Theorem B.2. *Let $X_1, \ldots, X_n$ be random variables. Then, the following are equivalent:*

1. *$X_1, \ldots, X_n$ have a joint density (which is by definition w.r.t. a σ-finite product measure $\mu_1 \times \cdots \times \mu_n$),*
2. *$P_{X_1,\ldots,X_n} \ll P_{X_1} \times \cdots \times P_{X_n}$,*
3. *$P_{X_1,\ldots,X_n}$ is representable as a Bayes network where each conditional distribution $P_{X_k | x_{j(k,1)},\ldots,x_{j(k,n_k)}}$ has a density w.r.t. a σ-finite measure μ_k,*
4. *$P_{X_1,\ldots,X_n}$ is representable as a Bayes network where each conditional distribution $P_{X_k | x_{j(k,1)},\ldots,x_{j(k,n_k)}}$ is absolutely continuous w.r.t. P_{X_k}.*
5. *$P_{X_1,\ldots,X_n}$ is representable as a Bayes network where each conditional distribution $P_{X_k | x_{j(k,1)},\ldots,x_{j(k,n_k)}}$ is dominated by a (not necessarily σ-finite) measure μ_k w.r.t. which there exists a marginal density $p(x_k)$.*

Furthermore,

6. *if the above conditions hold for $X_1, \ldots, X_n$, then they also hold for $X'_k = F_k(X_k)$, where $F_k : \mathsf{X}_k \to \mathsf{X}'_k$ are any measurable functions.*

Proof. This proof is a straightforward generalization of the proof of Theorem B.1.

For brevity, we shall denote the parents of x_k by $x_{<k} := (x_{j(k,1)}, \ldots, x_{j(k,n_k)})$.

2 ⇒ 5: The joint density $p := dP_{X_1,\ldots,X_n}/d(P_{X_1} \times \cdots \times P_{X_n})$ induces for each k the conditional density $p(x_k \mid x_1, \ldots, x_{k-1})$ w.r.t. the marginal distribution P_{X_k}. Thus, the required Bayes network is given by $x_{<k} := (x_1, \ldots, x_{k-1})$ for all $k = 1, \ldots, n$.

5 ⇒ 4: Let $k \in \{1, \ldots, n\}$ be arbitrary. Denoting $N := \{\, x_k \in \mathsf{X}_k : p(x_k) = 0 \,\}$, we have by the definition of conditional distribution

$$0 = \int_N p(x_k) d\mu(x_k) = P_{X_k}(N) = \int_{\mathsf{X}_{<k}} P_{X_k|x_{<k}}(N) dP_{X_{<k}}(x_{<k}),$$

which implies $P_{X_k|x_{<k}}(N) = 0$ for $P_{X_{<k}}$-a.e. $x_{<k}$. However, as $P_{X_k|x_{<k}}$ is only determined for $P_{x_{<k}}$-a.e. $x_{<k}$, we are free to modify it so that $P_{X_k|x_{<k}}(N) = 0$ for all $x_{<k}$. We will show that this $P_{X_k|x_{<k}}$ is dominated by P_{X_k} for all $x_{<k}$. Let $S \in \mathcal{X}_k$ be such that $P_{X_k}(S) = 0$. Then, we have

$$0 = P_{X_k}(S \setminus N) = \int_{S \setminus N} \underbrace{p(x_k)}_{>0} d\mu(x_k),$$

which implies $\mu_k(S \setminus N) = 0$. As $P_{X_k|x_{<k}} \ll \mu_k$, we have $P_{X_k|x_{<k}}(S \setminus N) = 0$, but as also $P_{X_k|x_{<k}}(N) = 0$, we obtain $P_{X_k|x_{<k}}(S) = 0$. Thus, $P_{X_k|x_{<k}} \ll P_{X_k}$ for all $x_{<k}$.

4 ⇒ 3: Choose $\mu_k = P_{X_k}$.

3 ⇒ 1: By the definition of the conditional densities and Fubini's theorem, we have

$$\begin{aligned} P_{X_1,\ldots,X_n}(S) &= \int_S \prod_k p(x_k \mid x_{<k}) d\mu_k(x_k) \\ &= \int_S \left[\prod_k p(x_k \mid x_{<k}) \right] d(\mu_1 \times \cdots \times \mu_n)(x). \end{aligned}$$

Thus, $\prod_k p(x_k \mid x_{<k})$ is a joint density of $X_1, \ldots, X_n$ w.r.t. the σ-finite measure $\mu_1 \times \cdots \times \mu_k$.

1 ⇒ 2: Suppose that $p = dP_{X_1,\ldots,X_n}/d(\mu_1 \ldots \mu_n)$ exists for some σ-finite measures $\mu_1, \ldots, \mu_n$ and let $S \in \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n$ be an arbitrary measurable set such that $(P_{X_1} \times \cdots \times P_{X_n})(S) = 0$. We will show that then $P_{X_1,\ldots,X_n}(S) = 0$. Denoting

$$\begin{aligned} N_k &:= \{\, x_k \in \mathsf{X}_k : p(x_k) = 0 \,\}, \\ N &:= \bigcup_k \mathsf{X}_1 \times \cdots \times \mathsf{X}_{k-1} \times N_k \times \mathsf{X}_{k+1} \times \cdots \times \mathsf{X}_n, \end{aligned}$$

we have $P_{X_k}(N_k) = 0$ for all k. Furthermore,

$$\begin{aligned} 0 = (P_{X_1} \times \cdots \times P_{X_n})(S \setminus N) &= \int_{S \setminus N} \prod_k p(x_k) d\mu_k(x_k) \\ &= \int_{S \setminus N} \underbrace{\left[\prod_k p(x_k) \right]}_{>0} d(\mu_1 \times \cdots \times \mu_k)(x) \end{aligned}$$

implies that $(\mu_1 \times \ldots \mu_n)(S \setminus N) = 0$, whence $P_{X_1,\ldots,X_n}(S \setminus N) = 0$. Thus,

$$P_{X_1,\ldots,X_n}(S) \leq P_{X_1,\ldots,X_n}(S \setminus N) + \sum_k P_{X_k}(N_k) = 0.$$

2 $\Rightarrow$ 6: Suppose that $F_k : \mathsf{X}_k \to \mathsf{X}'_k$ are arbitrary measurable mappings. We show that $P_{X_1,\ldots,X_n} \ll P_{X_1} \times \cdots \times P_{X_n}$ implies $P_{F_1(X_1),\ldots,F_n(X_n)} \ll P_{F_1(X_1)} \times \cdots \times P_{F_n(X_n)}$. For any $S \in \mathcal{X}_1 \otimes \cdots \otimes \mathcal{X}_n$,

$$\begin{aligned} 0 &= (P_{F_1(X_1)} \times \cdots \times P_{F_n(X_n)})(S) \\ &= (P_{X_1} \times \cdots \times P_{X_n})(\{\, (F_1^{-1}(x_1), \ldots, F_n^{-1}(x_n)) : x \in S \,\}) \end{aligned}$$

implies

$$0 = P_{X_1,\ldots,X_n}(\{\, (F_1^{-1}(x_1), \ldots, F_n^{-1}(x_n)) : x \in S \,\}) = P_{F_1(X_1),\ldots,F_n(X_n)}(S),$$

where F_k^{-1} denotes the preimage set. □

References

Billy Amzal, Frédéric Y. Bois, Eric Parent, and Christian P. Robert. Bayesian-optimal design via interacting particle systems. *Journal of the American Statistical Association*, 101(474):773–785, 2006.

Thomas M. Cover and Joy A. Thomas. *Elements of Information Theory*. John Wiley, 1991.

Morris H. DeGroot. *Optimal Statistical Decisions*. McGraw-Hill, 1970.

Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W. H. Freeman, 1979.

P. Ewen King-Smith. Efficient threshold estimates from yes-no procedures using few (about 10) trials. *American Journal of Optometry and Physiological Optics*, 61:119P, 1984. Abstract.

P. Ewen King-Smith and David Rose. Principles of an adaptive method for measuring the slope of the psychometric function. *Vision Research*, 37(12):1595–1604, 1997.

P. Ewen King-Smith, Scott S. Grigsby, Algis J. Vingrys, Susan C. Benes, and Aaron Supowit. Efficient and unbiased modifications of the QUEST threshold method: theory, simulations, experimental evaluation and practical implementation. *Vision Research*, 34 (7):885–912, 1994.

Andrei N. Kolmogorov. On the Shannon theory of information transmission in the case of continuous signals. *IEEE Transactions on Information Theory*, 2(4):102–108, 1956.

Leonid L. Kontsevich and Christopher W. Tyler. Bayesian adaptive estimation of psychometric slope and threshold. *Vision Research*, 39(16):2729–2737, 1999.

Janne V. Kujala and Tuomas J. Lukka. Bayesian adaptive estimation: The next dimension. *Journal of Mathematical Psychology*, 50(4):369–389, 2006.

Janne V. Kujala, Ulla Richardson, and Heikki Lyytinen. A Bayesian-optimal principle for child-friendly adaptation in learning games. University of Jyväskylä, Department of Mathematics and Statistics, Preprint 363, 2008.

Janne V. Kujala, Ulla Richardson, and Heikki Lyytinen. Estimation and visualization of confusability matrices from adaptive measurement data. submitted.

Luis Andres Lesmes, Seong-Taek Jeon, Zhong-Lin Lu, and Barbara Anne Dosher. Bayesian adaptive estimation of threshold versus contrast external noise functions: The quick TvC method. *Vision Research*, 46(19):3160–3176, 2006.

D. V. Lindley. On a measure of the information provided by an experiment. *The Annals of Mathematical Statistics*, 27(4):986–1005, 1956.

Peter Müller, Bruno Sansó, and Maria De Iorio. Optimal Bayesian design by inhomogeneous Markov chain simulation. *Journal of the American Statistical Association*, 99 (467):788–798, 2004.

Liam Paninski. Asymptotic theory of information-theoretic experimental design. *Neural Computation*, 17(7):1480–1507, 2005.

C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423, 623–656, 1948.

A.N. Shiryaev. *Probability*. Springer, 2nd edition, 1996.

Peter R. Snoeren and Marco J. H. Puts. Multiple parameter estimation in an adaptive psychometric method: MUEST, an extension of the QUEST method. *Journal of Mathematical Psychology*, 41(4):431–439, 1997.

Thomas G. Tanner, N. Jeremy Hill, Carl E. Rasmussen, and Felix A. Wichmann. Efficient adaptive sampling of the psychometric function by maximizing information gain. In Heinrich H. Bülthoff, Hanspeter A. Mallot, Rolf D. Ulrich, and Felix A. Wichmann, editors, *Proceedings of the 8th Tübinger Perception Conference*, volume 106, page 109, 2005.

Isabella Verdinelli and Joseph B. Kadane. Bayesian designs for maximizing information and outcome. *Journal of the American Statistical Association*, 87(418):510–515, 1992.

Andrew B. Watson and Denis G. Pelli. QUEST: A Bayesian adaptive psychometric method. *Perception & Psychophysics*, 33(2):113–120, 1983.