Surprise Maximization

D. Borwein; J. M. Borwein; P. Marechal

The American Mathematical Monthly, Vol. 107, No. 6 (Jun. - Jul., 2000), 517-527.

Stable URL:
http://links.jstor.org/sici?sici=0002-9890%28200006%2F07%29107%3A6%3C517%3ASM%3E2.0.CO%3B2-%23

The American Mathematical Monthly is currently published by Mathematical Association of America.

Surprise Maximization

D. Borwein, J. M. Borwein, and P. Maréchal

The *Surprise Examination* or *Unexpected Hanging* Paradox has long fascinated mathematicians and philosophers, as the number of publications devoted to it attests. For an exhaustive bibliography on the subject, see [**1**].

We examine and solve the optimization problems arising from an information theoretic *avoidance* of the *Paradox*. These problems provide a very satisfactory application of both the Kuhn-Tucker theory and of various classical inequalities and estimation techniques. We assume some elementary knowledge of optimization but recall the necessary convex analytic concepts in the course of the paper. Readers unfamiliar with this background may simply skip a couple of proofs and a few technical details.

1. AN INFORMATION MEASURE OF SURPRISE. The *Paradox*, as formulated by Timothy Chow in this MONTHLY [**3**] is:

> A teacher announces in class that an examination will be held on some day during the following week, and moreover that the examination will be a surprise. The students argue that a surprise exam cannot occur. For suppose the exam were on the last day of the week. Then on the previous night, the students would be able to predict that the exam would occur on the following day, and the exam would not be a surprise. So it is impossible for a surprise exam to occur on the last day. But then a surprise exam cannot occur on the penultimate day, either, for in that case the students, knowing that the last day is an impossible day for a surprise exam, would be able to predict on the night before the exam that the exam would occur on the following day. Similarly, the students argue that a surprise exam cannot occur on any other day of the week either. Confident in this conclusion, they are of course totally surprised when the exam occurs (on Wednesday, say). The announcement is vindicated after all. Where did the students' reasoning go wrong?

We study two optimization problems arising from an entropic approach to maximizing surprise. The idea of such an approach was proposed in outline by Karl Narveson [**3**, p. 49]. We do not discuss here the various approaches to the logical resolution of the paradox itself; the interested reader may consult [**3**]. Rather we attempt to answer the question:

> What should be the probability distribution of an event occurring once every week so that it maximizes the surprise it creates?

In the first place, this requires us to define a measure of surprise. Let us start by posing an information theoretic counterpart of the paradox: during a period of m days an event (such as a test given by a teacher or a surprise tax audit) occurs with probability p_i on day i, $i = 1, \ldots, m$. We wish to find a probability distribution that maximizes the *average surprise* caused by the event when it occurs.

We consider a measure of surprise analogous to the one used in the celebrated definition of the *Shannon entropy* [6]. The surprise on day i is the negative of the logarithm of the probability that the event occurs on day i given that it has not occurred so far. As in the classical definition, $-\ln p$ is used to measure the surprise associated with an event of probability p, which is also a measure of how much we learn if it occurs. The logarithm makes the measure *additive*, in the sense that the information associated with independent events should sum up when they both occur. The use of conditional probabilities introduces some *causality* in the notion: it accounts for what is already known of the previous days.

The event '*test occurs on day i*' is denoted by i, and its probability is denoted by $P(i)$ or p_i. The event '*test does not occur on day i*' is denoted by $\sim i$. The quantity to be maximized can therefore be written as

$$-\sum_{i=1}^{m} P(i)\ln P(i|\sim 1,\ldots,\sim(i-1)). \tag{1}$$

Using *Bayes' formula* for conditional probabilities, we obtain

$$\begin{aligned} P(i|\sim 1,\ldots,\sim(i-1)) &= \frac{P(\sim 1,\ldots,\sim(i-1)|i)P(i)}{P(\sim 1,\ldots,\sim(i-1))} \\ &= \frac{P(i)}{1-(P(1)+\cdots+P(i-1))} \\ &= \frac{P(i)}{P(i)+\cdots+P(m)}. \end{aligned}$$

We are led to consider the following optimization problem:

$$(\mathscr{P}_m) \quad \inf\{S_m(\mathbf{p})|\mathbf{p}\in\mathbb{R}^m, 1=\langle\mathbf{u},\mathbf{p}\rangle\}. \tag{2}$$

Here, $\mathbf{u}$ is the m-vector whose entries are all equal to 1 and S_m is the *surprise function*

$$S_m(\mathbf{p}) := \sum_{j=1}^{m} h\left(p_j, \frac{1}{m}\sum_{i=j}^{m} p_i\right), \quad \mathbf{p}\in\mathbb{R}^m, \tag{3}$$

where h is defined on $\mathbb{R}^2$ by

$$h(x,y) := \begin{cases} x\ln\dfrac{x}{y}-x & \text{if } x>0 \text{ and } y>0, \\ 0 & \text{if } x=0 \text{ and } y\geq 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{4}$$

Figure 1 displays the graph of the function h. For all $\mathbf{p}$ satisfying the constraint in (2), $S_m(\mathbf{p})$ differs from the negative of the quantity in (1) only by a constant. The factor m^{-1} was introduced into (3) to make subsequent computations more aesthetic and the limit analysis more harmonious.

We note that $S_m(\mathbf{p})$ can be regarded as a variant of the *Kullback-Leibler information measure* of $\mathbf{p}$ relative to its (normalized) *tail* $\mathbf{q}$:

$$\mathbf{q}=(q_1,\ldots,q_m) \text{ with } q_j := \frac{1}{m}\sum_{i=j}^{m} p_i, \quad j=1,\ldots,m. \tag{5}$$

The Kullback-Leibler information measure is an extension of the Boltzmann-Shannon entropy. It is also referred to as *relative information measure*, *cross-entropy*, or

I-divergence. Given two probability measures P and Q on a probability space, the *relative information of P with respect to Q* is

$$\mathcal{K}(P\|Q) := \int \left(\frac{dP}{dQ} \ln \frac{dP}{dQ} - \frac{dP}{dQ} \right) dQ = \int \left(\ln \frac{dP}{dQ} - 1 \right) dP$$

if P is *absolutely continuous* with respect to Q, and $\mathcal{K}(P\|Q) := +\infty$ otherwise. Those interested in the statistical meaning of this measure may refer to [**5**]. For an extended discussion of the *Maximum Entropy Principle*, one may consult [**4**] and references therein.

Also of interest is the following *continuous time* formulation of Problem (2). Suppose that the event occurs at some point t in the time interval $[0, T]$, with probability density $p(t)$. By analogy with the discrete case, it is reasonable to consider the following optimization problem:

$$(\mathcal{P}) \quad \inf\{\mathcal{S}(p) | p \in L_1([0, T]), 1 = \langle u, p \rangle\}, \tag{6}$$

in which the *surprise function* $\mathcal{S}$ is the functional defined on $L_1([0, T])$ by

$$\mathcal{S}(p) := \int_0^T h\left(p(t), \frac{1}{T} \int_t^T p(s)\, ds \right) dt,$$

and u denotes the function identically equal to unity on $[0, T]$.

2. SURPRISINGLY, SURPRISE IS CONCAVE. In this section, we establish the convexity of (the negative of) our measure of surprise. An extended real-valued function on $\mathbb{R}^n$ is said to be *closed* (respectively, convex) if its *epigraph* (the set of points that are above or on its graph) is closed (respectively, convex) in $\mathbb{R}^{n+1}$. If a convex function is not identically equal to $+\infty$ and is nowhere equal to $-\infty$ (such functions are said to be *proper*), then being closed is the same as being lower semi-continuous. The *domain* of a convex function f is the set of points where it is less than $+\infty$; we denote it by dom f. Given any function f on $\mathbb{R}^n$ (convex or not), the *convex conjugate* of f is the function

$$f^*(\xi) := \sup\{\langle \mathbf{x}, \boldsymbol{\xi} \rangle - f(\mathbf{x}) | \mathbf{x} \in \mathbb{R}^n\}, \quad \boldsymbol{\xi} \in \mathbb{R}^n.$$

It is easily shown that f^* is always closed and convex [**7**, Theorem 12.2]. Furthermore, if f is closed, proper, and convex, then so is f^* and the *bi-conjugate* $f^{**} := (f^*)^*$ is f itself [**7**, Theorem 12.2]. Even without this theoretical underpinning, computation of f as a *double-conjugate* provides an accessible way of establishing both convexity and semi-continuity.

Lemma 1. *The function h defined in* (4) *is closed and convex.*

Proof: One may show directly that h is the convex conjugate of the *indicator function*

$$\delta((\xi, \eta) | C) := \begin{cases} 0 & \text{if } (\xi, \eta) \in C, \\ +\infty & \text{otherwise,} \end{cases}$$

where C is the convex set $\{(\xi, \eta) \in \mathbb{R}^2 | \eta \leq -\exp \xi\}$. This proves that h is closed and convex. ■

Convexity of h can also be derived from the easily proven fact that, for any interval I, a function

$$(x, y) \mapsto yf(xy^{-1})$$

is convex on $I \times (0, \infty)$ if and only if f is convex on I. [A bad way of proving convexity of h is to compute the *Hessian* matrix and check that it is positive semi-definite.]

Using Lemma 1, we deduce that S_m and $\mathscr{S}$ are convex. Indeed, we have

$$S_m(\mathbf{p}) = \sum_{i=1}^{m} h(p_i, [J\mathbf{p}]_i) \qquad \text{and} \qquad S(p) = \int_0^T h(p(t), [\mathscr{J}p](t))\, dt,$$

in which J is the $(m \times m)$-matrix whose entries are m^{-1} on and above the diagonal and 0 elsewhere, and $\mathscr{J}$: $L_1([0,T]) \to \mathscr{C}([0,T])$ is the linear mapping defined by

$$[\mathscr{J}p](t) := \frac{1}{T}\int_t^T p(s)\, ds. \tag{7}$$

Composition of a convex function with a linear mapping is, of course, convex.

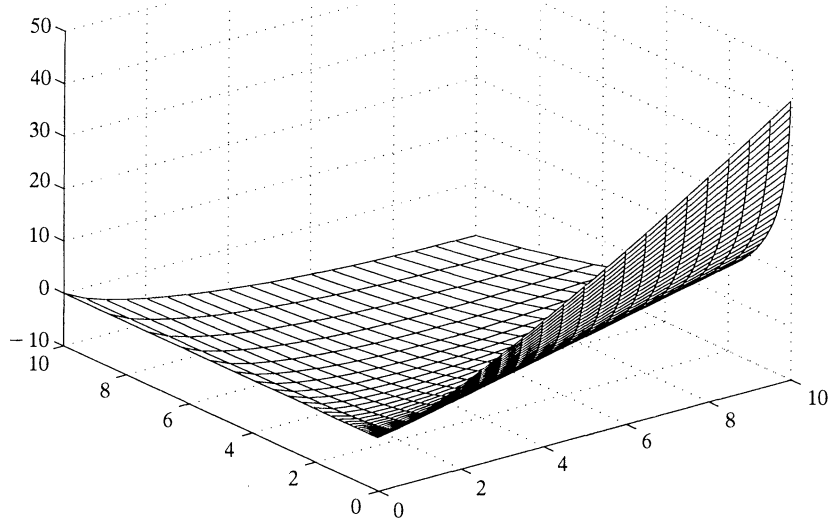

Figure 1. Graph of $(x, y) \mapsto x \ln x/y - x$.

3. DISCRETE TIME ANALYSIS. Constrained optimization problems such as (2) are traditionally approached using concepts from *duality theory*, which flows from the theory of *Lagrange multipliers*. Roughly speaking, duality theory reduces constrained optimization problems to unconstrained ones. A modern version of convex duality theory is best posed in the language of Fenchel conjugation [7, Section 31]. We recall some additional basic facts. Let f be a closed proper convex function on $\mathbb{R}^n$, let A be an $m \times n$ matrix, and let $\mathbf{y} \in \mathbb{R}^m$. We consider the linearly constrained optimization problem

$$(\mathscr{P}) \quad \inf\{f(\mathbf{x}) | \mathbf{x} \in \mathbb{R}^n, \mathbf{y} - A\mathbf{x} = \mathbf{0}\}. \tag{8}$$

We denote the optimal value of $(\mathscr{P})$ by $V(\mathscr{P})$, the *feasible set* by $F(\mathscr{P})$, and the *solution set* by $S(\mathscr{P})$. Thus, $F(\mathscr{P}) := \{\mathbf{x} | \mathbf{y} - A\mathbf{x} = \mathbf{0}\}$ and $S(\mathscr{P}) := \{\mathbf{x} \in F(\mathscr{P}) | f(\mathbf{x}) = V(\mathscr{P})\}$. The *Lagrangian* of (8) is the function

$$\mathscr{L}(\boldsymbol{\lambda}, \mathbf{x}) := f(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{y} - A\mathbf{x} \rangle, \quad \boldsymbol{\lambda} \in \mathbb{R}^m, \mathbf{x} \in \mathbb{R}^n.$$

For a given $\boldsymbol{\lambda}$, $\mathscr{L}(\boldsymbol{\lambda}, \mathbf{x})$ can be regarded as a penalized version of f. Each component of $\boldsymbol{\lambda}$ fixes the price (positive or negative) to be paid if the corresponding constraint is violated. [This is more easily understood if the constraints $A\mathbf{x} = \mathbf{y}$ are replaced by $A\mathbf{x} \leq \mathbf{y}$, which may be handled by introducing an additional variable $\mathbf{z} \geq 0$ and writing $A\mathbf{x} + \mathbf{z} = \mathbf{y}$.] Under favourable circumstances, it is possible to find a particular value $\bar{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$ such that minimizers of $\mathscr{L}(\bar{\boldsymbol{\lambda}}, \cdot)$ also solve (8). Such a $\bar{\boldsymbol{\lambda}}$ is then called a *Lagrange multiplier* or a *shadow price*. Now minimizing $\mathscr{L}(\bar{\boldsymbol{\lambda}}, \cdot)$ is an unconstrained problem (save for any implicit constraints imposed by dom f). We can now state the Kuhn-Tucker Theorem, which provides necessary and sufficient conditions (on $\boldsymbol{\lambda}$ and $\mathbf{x}$) for $\mathbf{x}$ to be a solution of (8). A proof may be found in [7, Corollary 28.3.1].

Theorem 1 (Kuhn-Tucker). *Suppose that* $V(\mathscr{P}) \neq -\infty$ *and that* $F(\mathscr{P}) \cap \operatorname{int} \operatorname{dom} f \neq \emptyset$. *Then, the following are equivalent*:

(i) $\mathbf{x} \in S(\mathscr{P})$;
(ii) $\sup \mathscr{L}(\cdot, \mathbf{x}) = \mathscr{L}(\bar{\boldsymbol{\lambda}}, \mathbf{x}) = \inf \mathscr{L}(\bar{\boldsymbol{\lambda}}, \cdot)$ *for some* $\bar{\boldsymbol{\lambda}}$;
(iii) $\mathbf{x} \in F(\mathscr{P})$ *and* $A^*\bar{\boldsymbol{\lambda}} \in \partial f(\mathbf{x})$ *for some* $\bar{\boldsymbol{\lambda}}$.

In condition (iii), A^* is the matrix transpose of A and $\partial f(\mathbf{x})$ denotes the *subdifferential* of f at $\mathbf{x}$, i.e., the set of *subgradients* of f at $\mathbf{x}$. Precisely, a vector $\boldsymbol{\xi} \in \mathbb{R}^n$ is a *subgradient* of f at $\mathbf{x}$ if the *subgradient inequality*

$$f(\mathbf{z}) \geq g(\mathbf{z}) := f(\mathbf{x}) + \langle \boldsymbol{\xi}, \mathbf{z} - \mathbf{x} \rangle$$

holds for all $\mathbf{z} \in \mathbb{R}^n$. In the words of Rockafellar, the subgradient inequality says that "the graph of the affine function g is a non-vertical supporting hyperplane to the epigraph of f at $(\mathbf{x}, f(\mathbf{x}))$" [7, p. 214]. If f is convex and differentiable at $\mathbf{x}$, $\nabla f(\mathbf{x})$ is the unique subgradient of f at $\mathbf{x}$, and conversely. Points $(\bar{\boldsymbol{\lambda}}, \mathbf{x})$ satisfying condition (ii) are said to be *saddle points* of $\mathscr{L}$. The requirements in (iii) are a form of the *Kuhn-Tucker conditions*. Notice that, in condition (ii), $\bar{\boldsymbol{\lambda}}$ appears as the maximizer of the (concave) *dual function*

$$D(\boldsymbol{\lambda}) := \inf \mathscr{L}(\boldsymbol{\lambda}, \cdot).$$

Note finally that dom f may have an empty interior. Theorem 1 still applies, however, with the weaker assumption that $F(\mathscr{P})$ intersects the *relative interior of* dom f [7, Section 28], that is, the interior relative to the smallest affine manifold containing dom f.

We now return to the study of Problem (2), which must have a solution since we are minimizing a closed function over a compact set. The *Lagrangian* of (2) is the function

$$\mathscr{L}(\mathbf{p}, \lambda) := S_m(\mathbf{p}) + \lambda(1 - \langle \mathbf{u}, \mathbf{p} \rangle), \quad \mathbf{p} \in \mathbb{R}^m, \ \lambda \in \mathbb{R}.$$

Theorem 1 tells us that $\mathbf{p}$ is a solution for (2) if and only if

(α) $0 = 1 - \langle \mathbf{u}, \mathbf{p} \rangle$;
(β) there exists $\bar{\lambda} \in \mathbb{R}$ such that $\mathbf{0} \in \partial S_m(\mathbf{p}) + \bar{\lambda}\partial[1 - \langle \mathbf{u}, \cdot \rangle](\mathbf{p})$.

Indeed, one can check that $V(\mathscr{P}_m) \neq -\infty$ and that $(\mathscr{P}_m)$ has a feasible solution in int dom $S_m = \{\mathbf{p} \in \mathbb{R}^m \mid \mathbf{p} > \mathbf{0}\}$. Furthermore, S_m is differentiable in the interior of its domain, and we have

$$\frac{\partial S_m}{\partial p_k}(\mathbf{p}) = \ln m\mu_k - \sum_{i \leq k} \mu_i, \quad \text{where} \quad \mu_k := \frac{p_k}{\sum_{j \geq k} p_j}. \tag{9}$$

Consequently, for a strictly positive distribution, condition (β) becomes

$$0 = \ln m\mu_k - \sum_{i \le k} \mu_i - \lambda, \quad k = 1, \ldots, m. \tag{10}$$

Now, by definition, $\mu_m = 1$, so setting $k = m$ in (10) gives $\lambda = \ln m - \Sigma \mu_i$, from which we obtain the recursion

$$\mu_m = 1, \quad \mu_k = \exp\left(- \sum_{j=k+1}^{m} \mu_j \right), \quad k = m - 1, \ldots, 1. \tag{11}$$

Since

$$\mu_{k-1} = \exp\left(- \sum_{j=k}^{m} \mu_j \right) = \exp(-\mu_k) \exp\left(- \sum_{j=k+1}^{m} \mu_j \right),$$

the *backward recursion* (11) can be rewritten as

$$\mu_m = 1, \quad \mu_{k-1} = \mu_k \exp(-\mu_k), \quad k = m, \ldots, 2. \tag{12}$$

The values of the μ_k's can be obtained as illustrated in Figure 2. Figure 3 shows examples of optimal probability distributions, for $m = 7$ and $m = 50$.

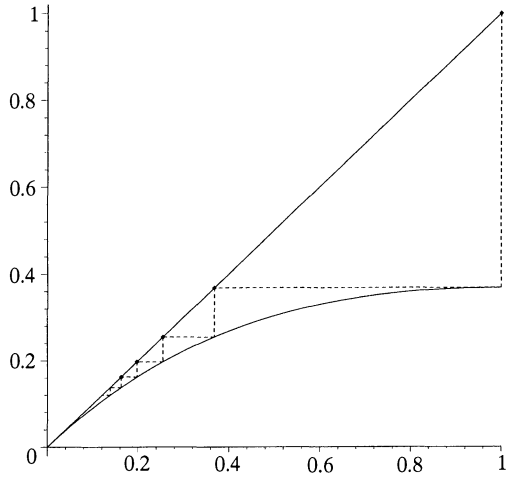

Figure 2. Recursion for the μ_k's.

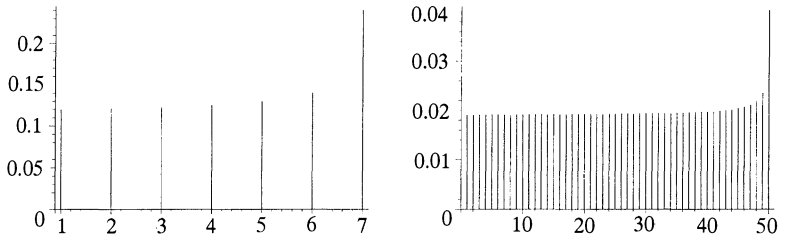

Figure 3. Optimal distributions for $m = 7$ (left) and $m = 50$ (right).

Finally, from condition (α) and the values of the μ_k's, we see that the components of $\mathbf{p}$ must obey the following *forward recursion*:

$$p_1 = \mu_1, \quad p_k = \mu_k \times \left(1 - \sum_{j=1}^{k-1} p_j\right), \quad k = 2, \ldots, m. \tag{13}$$

The vector $\mathbf{p}$ defined in (13) satisfies conditions (α) and (β), and therefore *uniquely* solves Problem $(\mathscr{P}_m)$ in (2). Indeed, if $\mathbf{p}.$ were a nonnegative solution of (2), then $(\mathbf{p} + \mathbf{p}.)/2$ would be a positive solution, which must equal $\mathbf{p}$ since a positive solution is uniquely determined.

Most pleasingly, the iteration is easy to handle both numerically and theoretically. For example, the components of $\mathbf{p}$ form an increasing sequence. Indeed,

$$p_k = \mu_k(p_k + \cdots + p_m) \quad \text{and} \quad p_{k-1} = \mu_{k-1}(p_{k-1} + \cdots + p_m),$$

from which we deduce, using (12), that

$$\begin{aligned} \frac{p_k}{p_{k-1}} &= \frac{\mu_k(1 - \mu_{k-1})}{\mu_{k-1}} = \exp \mu_k \times (1 - \mu_k \exp(-\mu_k)) \\ &= \exp \mu_k - \mu_k > 1, \end{aligned} \tag{14}$$

since $\mu_k > 0$. We recapitulate the prior discussion as:

Algorithm 1. *The unique probability distribution* $\mathbf{p}^m$ *that maximizes surprise in Problem* $(\mathscr{P}_m)$ *(given in* (2)*) is strictly increasing and is determined as follows. Compute*

$$\mu_m = 1, \quad \mu_{j-1} = \mu_j \exp(-\mu_j), \quad j = m, \ldots, 2, \tag{15}$$

and then compute

$$p_1 = \mu_1, \quad p_k = \mu_k \times \left(1 - \sum_{i=1}^{k-1} p_i\right), \quad k = 2, \ldots, m. \tag{16}$$

Remark 1. As observed in [**3**, p. 50], the (optimal) conditional probability that the event occurs on the ith-to-the-last day, given that it has not occurred thus far, is *independent* of m. This is immediate from (12) and the equality

$$P(m - i | \sim 1, \ldots, \sim (m - i - 1)) = p_{m-i}\left(\sum_{j=m-i}^{m} p_j\right)^{-1} = \mu_{m-i}.$$

Furthermore, as the μ_k's are defined via a backward recursion, p_{m-i}/p_{m-i-1} is also independent of m.

Remark 2. We may also obtain the solution of Problem $(\mathscr{P}_m)$ of (2) via the optimization problem

$$\inf\{S'_m(\mathbf{p}, \mathbf{q}) | 1 = \langle \mathbf{1}, \mathbf{p} \rangle, \mathbf{q} = J\mathbf{p}\},$$

where $S'_m(\mathbf{p}, \mathbf{q}) := \Sigma h(p_j, q_j)$. The corresponding Kuhn-Tucker conditions are

(α') $0 = 1 - \langle \mathbf{u}, \mathbf{p} \rangle$ and $\mathbf{0} = \mathbf{q} - J\mathbf{p}$;

(β') there exist $\lambda \in \mathbb{R}$ and $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_m) \in \mathbb{R}^m$ such that

$$\mathbf{0} \in \partial S'_m(\mathbf{p}, \mathbf{q}) + \lambda \partial f(\mathbf{p}, \mathbf{q}) + \lambda_1 \partial f_1(\mathbf{p}, \mathbf{q}) + \cdots + \lambda_m \partial f_m(\mathbf{p}, \mathbf{q})$$

with f and $\mathbf{f} = (f_1, \ldots, f_m)$ defined by

$$f(\mathbf{p}, \mathbf{q}) := 1 - \langle \mathbf{u}, \mathbf{p} \rangle \quad \text{and} \quad \mathbf{f}(\mathbf{p}, \mathbf{q}) := \mathbf{q} - J\mathbf{p}.$$

It is then easy to check that the λ_j's derived from (α') and (β') coincide with the μ_j's of the previous discussion multiplied by m.

4. HOW DOES THE DISTRIBUTION BEHAVE?. Some striking characteristics of the optimal distribution are mentioned in Remark 1. It is also natural to consider the asymptotic behaviour of Problem $(\mathscr{P}_m)$ as m tends to infinity. We are now ready to establish three key properties. First, we show that asymptotically the least probability $p_1^{(m)}$ behaves like m^{-1}. The nub is an analysis of the rate of convergence of the *Picard-Banach iteration*

$$t_{n+1} = g(t_n)$$

to the unique fixed point of a *contractive* (but not *strictly contractive*) self-map g on $[0, 1]$ when the fixed point occurs at a point where $|g'(t)| = 1$. Recall that g is contractive if

$$|g(t) - g(s)| < |t - s|$$

for all $t \neq s$ in $[0, 1]$. In our case we use the map $x \mapsto x \exp(-x)$.

Proposition 1. *The quantity $mp_1^{(m)}$ tends to one as m tends to ∞.*

Proof: We define a sequence $\{t_n\}$ by setting

$$t_i := \mu_{m+1-i}^{(m)}, \qquad i = 1, \ldots, m, \quad m = 1, 2, \ldots . \tag{17}$$

Observe that t_i is independent of m, that $t_m = p_1^{(m)}$, and that the sequence satisfies the recursion

$$t_1 = 1, \quad t_{k+1} = t_k \exp(-t_k), \qquad k = 1, 2, \ldots .$$

We note that t_k tends monotonically to a limit l, which must necessarily be zero. Hence $t_{k+1}^{-1} - t_k^{-1} = t_k^{-1}(\exp t_k - 1)$, which tends to $\exp'(0) = 1$ as k tends to infinity. Whence, since Cesàro averaging preserves limits,

$$\frac{1}{mt_m} = \frac{1}{m} \sum_{k=1}^{m-1} \frac{e^{t_k} - 1}{t_k} + \frac{1}{mt_1}$$

also tends to 1. ∎

Next, we show that the ratio between the last (biggest) and first (smallest) components converges.

Proposition 2.

$$\lim_{m \to \infty} \frac{p_m^{(m)}}{p_1^{(m)}} \text{ exists and is finite.}$$

Proof: We have from (14) and (17) that

$$\lim_{m \to \infty} \frac{p_m^{(m)}}{p_1^{(m)}} = \lim_{m \to \infty} \prod_{j=2}^{m} \left(e^{\mu_j^{(m)}} - \mu_j^{(m)} \right) = \lim_{m \to \infty} \prod_{j=1}^{m-1} \left(e^{t_j} - t_j \right) \simeq 2.132979 \ldots .$$

The limit exists because of the inequality $1 \leq \exp t_j - t_j \leq 1 + t_j^2$, while $\Sigma_j t_j^2 < \infty$ by Proposition 1. We now appeal to the standard fact that $\prod_n (1 + |a_n|)$ and $\Sigma_n |a_n|$ converge or diverge together. ∎

Finally, we show that in the limit our solution value approaches zero.

Proposition 3. *The optimal value $V(\mathscr{P}_m)$ of $(\mathscr{P}_m)$ tends to 0 as m tends to infinity.*

Proof: To establish this, we show that $\limsup V(\mathscr{P}_m) \le 0 \le \liminf V(\mathscr{P}_m)$. The first inequality is easily obtained from identifying a Riemann sum:

$$V(\mathscr{P}_m) \le S_m\left(\frac{1}{m}, \ldots, \frac{1}{m}\right) = \ln m - \frac{\ln m!}{m} - 1$$
$$= -\frac{1}{m}\sum_{k=1}^{m} \ln\frac{k}{m} - 1 \to -\int_0^1 \ln t\, dt - 1 = 0.$$

To obtain the other inequality, consider

$$\tau_m := \sum_{i=1}^{m-1}\left(p_i^{(m)} \ln\frac{p_i^{(m)}}{q_{i+1}^{(m)}} - p_i^{(m)}\right) \quad \text{and} \quad \sigma_m := \sum_{i=1}^{m-1}\left(p_i^{(m)} \ln\frac{p_i^{(m)}}{q_i^{(m)}} - p_i^{(m)}\right).$$

We make two claims:

(i) $\tau_m - \sigma_m$ tends to 0 as m tends to infinity;
(ii) $\tau_m \ge -p_m^{(m)} \ln m$.

For (i), we recall from (5) and (9) that $\mu_i^{(m)} = p_i^{(m)}/(mq_i^{(m)})$ and so

$$\tau_m - \sigma_m = -\sum_{i=1}^{m-1} p_i^{(m)} \ln\left(1 - \mu_i^{(m)}\right),$$

whence, as $p_i^{(m)}$ increases with i,

$$0 \le \tau_m - \sigma_m = -\sum_{i=1}^{m-1} p_{m-i}^{(m)} \ln(1 - t_{i+1}) \le -p_m^{(m)} \sum_{i=1}^{m-1} \ln(1 - t_{i+1}) \to 0,$$

since $t_i \to 0$ and $mp_m^{(m)} = O(1)$.

A proof of (ii) is deferred to Section 5 (Corollary 1), where it is a consequence of a general integral inequality.

By design,

$$V(\mathscr{P}_m) = \sigma_m + p_m^{(m)} \ln m - p_m^{(m)}.$$

It follows from (ii) that $V(\mathscr{P}_m) \ge \sigma_m - \tau_m - p_m^{(m)}$ and so, since $p_m^{(m)} \to 0$, (i) shows $\liminf V(\mathscr{P}_m) \ge 0$ as needed. ∎

The techniques of these three propositions allow us to make considerably more precise assertions about the asymptotics of $\mathbf{p}^{(m)}$.

5. CONTINUOUS TIME ANALYSIS. In the discrete case, the distribution is strictly increasing, with a sharp increase at the tip of the tail (see Figure 3). In measure, this is washed out in the limit. Indeed, the optimal continuous distribution is flat, as the following theorem shows.

Theorem 2. *For all $p \in L_1([0, T])$, we have*

$$\int_0^T p(t) \ln\frac{p(t)}{\frac{1}{T}\int_t^T p(s)\, ds} dt \ge \int_0^T p(t)\, dt,$$

or, equivalently, $\mathscr{S}(p) \ge 0$, with equality if and only if p is constant on $[0, T]$.

Proof: We can assume that p is (almost everywhere) nonnegative, for otherwise $\mathscr{S}(p) = \infty$. Let us put $q(t) := [\mathscr{J}p](t) = 1/T\int_t^T p(s)\,ds$, as in (7). Observe that, on integrating by parts,

$$\begin{aligned}\mathscr{S}(p) &= \int_0^T \left(p(t)\ln\frac{p(t)}{q(t)} - p(t)\right) dt \\ &= \int_0^T \big(p(t)\ln p(t) - p(t)\big)\, dt + T\int_0^T q'(t)\ln q(t)\, dt \\ &= \int_0^T p(t)\ln p(t)\, dt - Tq(0)\ln q(0),\end{aligned}$$

The theorem will therefore be proved if we can show that

$$\int_0^T p(t)\ln p(t)\, dt \geq Tq(0)\ln q(0), \tag{18}$$

with equality if and only if p is constant. Now, applying the integral version of Jensen's inequality to the strictly convex function $g := x \mapsto x\ln x - x$ yields

$$\frac{1}{T}\int_0^T \left(\frac{p(t)}{q(0)}\ln\frac{p(t)}{q(0)} - \frac{p(t)}{q(0)}\right) dt \geq g(1) = -1,$$

from which (18) follows immediately. ∎

Theorem 2 shows that the (unique) solution of Problem $(\mathscr{P})$ given in (6) is the uniform probability density on $[0, T]$. A consequence of Theorem 2, which completes to the considerations of Section 4, is the following:

Corollary 1. *With the notation of Section* 4, *we have*

$$\tau_m \geq -p_m^{(m)}\ln m.$$

Proof: Apply Theorem 2 with

$$T := 1 \quad \text{and} \quad p(t) := p_n^{(m)} \text{ if } t \in \left(\frac{n-1}{m}, \frac{n}{m}\right] (n = 1, \ldots, m).$$

Observe that, for $(n-1)/m < t \leq n/m$, $n = 1, 2, \ldots, m-1$,

$$q(t) \geq \sum_{k=n+1}^{m} \int_{\frac{k-1}{m}}^{\frac{k}{m}} p(t)\, dt = \frac{1}{m}\sum_{k=n+1}^{m} p_k^{(m)} = q_{n+1}^{(m)},$$

and, for $(m-1)/m < t \leq 1$, $q(t) = p_m^{(m)}(1-t)$. Hence τ_m majorizes

$$\begin{aligned}&m\sum_{n=1}^{m-1} \int_{\frac{n-1}{m}}^{\frac{n}{m}} p(t)\left\{ \ln\left(\frac{p(t)}{q(t)}\right) - 1\right\} dt \\ &\quad = m\int_0^{1-\frac{1}{m}} p(t)\left\{ \ln\left(\frac{p(t)}{q(t)}\right) - 1\right\} dt \\ &\quad = \int_0^1 p(t)\left\{ \ln\left(\frac{p(t)}{q(t)}\right) - 1\right\} dt + m\int_{1-\frac{1}{m}}^1 p_m^{(m)}\{\ln(1-t) + 1\}\, dt \\ &\quad \geq 0 - p_m^{(m)}\ln m,\end{aligned}$$

on evaluating the second integral and applying Theorem 2. ∎

This finishes the proof that the optimal value of $(\mathscr{P}_m)$ tends to 0 (which is also the optimal value of $(\mathscr{P})$), as claimed in Section 4.

6. CONCLUSION. The entropic formulation of the Surprise Examination Problem provides a beautiful case study of the application of concepts from the elementary theory of convex constrained optimization, probability, and classical inequality theory. Its attractiveness comes in part from the very explicit recursive nature of the (discrete time) solution, which derives from the Kuhn-Tucker Theorem.

REFERENCES

1. `http://front.math.ucdavis.edu/search/author:Chow+and+Timothy`
2. J. M. Borwein and A. S. Lewis, *Convex Analysis and Nonlinear Optimization*, CMS-Springer Advanced Texts, Springer-Verlag, New York, to appear.
3. T. Y. Chow, *The surprise examination or unexpected hanging paradox*, Amer. Math. Monthly 105 (1998) 45–51.
4. H. Gzyl, *The Method of Maximum Entropy*, World Scientific, Singapore, 1995.
5. S. Kullback, *Information Theory and Statistics*, Dover, New York, 1968.
6. A. Rényi, *Calcul des Probabilités. Avec un appendice sur la théorie de l'information*, Dunod, Paris, 1966. Traduit de l'allemand par C. Bloch. Collection Universitaire de Mathématiques, No. 21.
7. R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970. Princeton Mathematical Series, No. 28.

DAVID BORWEIN was born in Lithuania and raised in South Africa. He obtained a Ph.D. and D.Sc. from the University of London. His Ph.D. supervisor was L. S. Bosanquet whose own supervisor was G. H. Hardy. In 1959 he was elected to a Fellowship of the Royal Society of Edinburgh. He taught in St. Andrews, Scotland until 1963, and was head of the mathematics department at the University of Western Ontario from 1967 until he retired in 1989. He remains actively engaged in research and has written extensively on summability theory and other areas of analysis. He is a former Advanced Problems Editor of this MONTHLY, and served as President of the Canadian Mathematical Society from 1987 until 1989.
University of Western Ontario, London, Ontario, Canada N6A 5B7
dborwein@uwo.ca

JONATHAN M. BORWEIN is Shrum Professor of science and Director of the Centre for Experimental and Constructive Mathematics at Simon Fraser University. He received his Ph.D. from Oxford in 1974, as a Rhodes Scholar. Prior to joining SFU in 1993, he worked at Dalhousie University, Carnegie-Mellon, and Waterloo. He has received a Fellowship in the Royal Society of Canada (1994), and an Honorary Degree from Limoges (1999). His research interests span pure (analysis), applied (optimization), and computational (complexity and numerical analysis) mathematics. Following in his father's footsteps, he is President Elect of the Canadian Mathematical Society.
Centre for Experimental and Constructive Mathematics, Department of Mathematics and Statistics, Simon Fraser University, Burnaby, B.C., Canada V5A 1S6
jborwein@cecm.sfu.ca

PIERRE MARÉCHAL was a postdoctoral fellow at the Centre for Experimental and Constructive Mathematics of Simon Fraser University and Vancouver General Hospital, and is now Maître de Conférence at the University of Montpellier, France. He obtained his Ph.D. from the University of Toulouse, France on the regularization of Fourier-type inverse problems. His interest in convex analysis was generated by the study of entropy optimization. His research is concerned with convex analysis, optimization, and inverse problems of medical tomography.
Laboratoire ACSIOM, Département de Mathématiques, Université de Montpellier 2, Case Courrier 051, Place Eugène Bataillon, 34 095 Montpellier Cedex 5, France
marechal@math.univ-montp2.fr