Maximizing Surprise

Drive

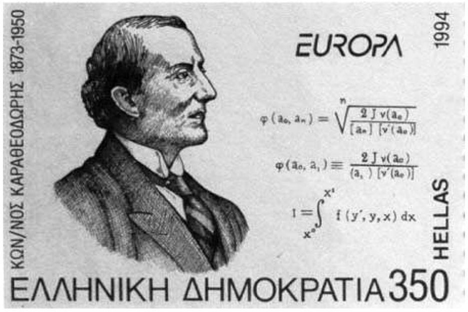
Jonathan M. Borwein, FRSC
Research Chair in IT
Dalhousie University

Halifax, Nova Scotia, Canada

Optimization Seminar

Portugal, November 2005

Constantin Carathéodory



MAA 1936

I'll be glad if I have succeeded in impressing the idea that it is not only pleasant to read at times the works of the old mathematical authors, but this may occasionally be of use for the actual advancement of science.

◄ Revised 27-10-05 ►

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, the reader is referred to [1].

Herein, the optimization problems arising from an information theoretic *avoidance* of the *Paradox* are examined and solved.

They provide a very satisfactory application of both the Kuhn-Tucker theory and of various classical inequalities and estimation techniques.

▷ Although the necessary convex analytic concepts are recalled in the course of the presentation, some elementary knowledge of optimization is assumed.

> Those without this background may simply skip a couple of proofs and few technical details.

INFORMATION MEASURE OF SURPRISE

Tim Chow's [3] version of the *Paradox*:

> *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?

In this work, we study two optimization problems arising from an entropic approach to maximizing surprise. 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; one may consult [1,3].

▷ Rather we ask the question:

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

▷ This requires us to find a *measure of surprise*.

$\triangleright$ 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 $i = 1, \ldots, m$.

We wish to find a probability distribution that maximizes the *average surprise* caused by the event when it occurs.

$\triangleright$ We consider a measure of surprise analogous to the one used in the celebrated definition of the *Shannon entropy* [2,4,6].

▷ The surprise on day i is the negative of the logarithm of the *probability the event occurs on day i given that it has not occurred so far*.

▷ As in the classical definition, $-\log p$ 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*: the information associated with independent events should sum up when they both occur.

▷ The use of conditional probabilities introduces some *causality*: it accounts for what is already known of the previous days.

The event *‘test occurs on day i’* is simply denoted by i, and its probability is denoted by $P(i)$ or p_i. The event *‘test does not occur on day i’* will be denoted by $\sim i$.

▷ Thus, we need to maximize:

$$-\sum_{i=1}^{m} P(i) \log P\Big(i \,|\, \sim 1, \ldots, \sim (i-1)\Big). \qquad (1)$$

Using *Bayes’ formula* for conditional probabilities, we obtain an explicit formula:

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

$\triangleright$ We are led to the next optimization problem:

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

Here, $\mathbf{u}$ is the $m-$vector of 1's and:

$\triangleright$ S_m is the (m-dimensional) *surprise function*

$$S_m(\mathbf{p}) := \sum_{j=1}^{m} p_j \log \frac{p_j}{\frac{1}{m} \sum_{i \geq j} p_i} - \sum_{j=1}^{m} p_j.$$

More precisely,

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

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

$$h(x,y) := \begin{cases} x \log \dfrac{x}{y} - x & \text{if } x > 0 \text{ and } y > 0, \\ 0 & \text{if } x = 0 \text{ and } y \geq 0, \\ +\infty & \text{otherwise.} \end{cases} \tag{3}$$

▷ 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} makes subsequent computations more aesthetic and the limit analysis more harmonious.

▷ Note that $S_m(\mathbf{p})$ can be viewed as the *Kullback-Leibler information measure* of $\mathbf{p}$ relative to its (normalized) *tail* $\mathbf{q}$:

$$\begin{aligned} \mathbf{q} &:= (q_1, \ldots, q_m) \quad \text{with} \\ q_j &:= \tfrac{1}{m} \textstyle\sum_{i=j}^{m} p_i, \quad j = 1, \ldots, m. \end{aligned} \tag{4}$$

The *Kullback-Leibler information measure* [2, 5] is an extension of Boltzmann-Shannon entropy. It is also called the *relative information measure*, *cross-entropy* or *I-divergence*.

Given two probability measures P and Q, the relative information of P with respect to Q is

$$\begin{aligned}\mathcal{K}(P\|Q) &:= \int \left(\frac{dP}{dQ}\log\frac{dP}{dQ} - \frac{dP}{dQ}\right) dQ \\ &= \int \left(\log\frac{dP}{dQ} - 1\right) dP\end{aligned}$$

if P is *absolutely continuous* with respect to Q, and $\mathcal{K}(P\|Q) := +\infty$ otherwise, [5].

▷ For an extended discussion on the *Maximum Entropy Principle*, one may consult [4] and references therein.

▷ Also of interest is the following *continuous time* formulation of the above problem.

We 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, we consider the following optimization problem:

$$(\mathcal{P}) \quad \inf \left\{ \mathcal{S}(p) \,\middle|\, p \in L_1\Big([0,T]\Big),\ 1 = \langle u\,,\,p \rangle \right\} \tag{5}$$

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

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

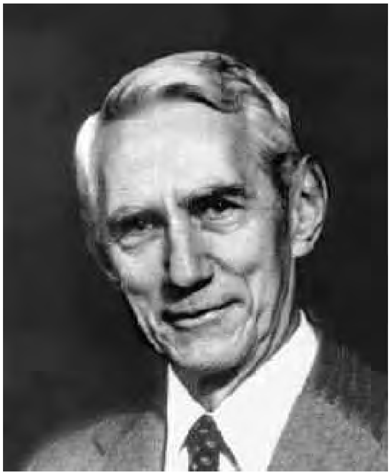
and $u \equiv 1\ [0,T]$.

As above h is defined by

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

WHAT is

Boltzmann (1844-1906)


Shannon (1916-2001)

WHAT is ENTROPY?

Despite the narrative force that the concept of entropy appears to evoke in everyday writing, in scientific writing entropy remains a thermodynamic quantity and a mathematical formula that numerically quantifies disorder. When the American scientist Claude Shannon found that the mathematical formula of Boltzmann defined a useful quantity in information theory, he hesitated to name this newly discovered quantity entropy because of its philosophical baggage. The mathematician John Von Neumann encouraged Shannon to go ahead with the name entropy, however, since "no one knows what entropy is, so in a debate you will always have the advantage."

- **19C**: Boltzmann—thermodynamic disorder
- **20C**: Shannon—information uncertainty
- **21C**: JMB—potentials with superlinear growth

SURPRISINGLY, SURPRISE IS CONCAVE

▷ We now establish the *convexity* of (the negative of) our measure of surprise. An extended real-valued function on $\mathbb{R}^n$ is *closed (convex)* if its *epigraph* (the set of points which are above or on its graph) is closed (convex) in $\mathbb{R}^{n+1}$.

▷ The *domain* of a convex function f is the set of points where it is less than $+\infty$, denoted by dom f.

▷ If a convex function is not identically $+\infty$ and is nowhere $-\infty$ (such functions are *proper*), then being closed is the same as being *lower semi-continuous*.

$\triangleright$ Given any function f on $\mathbb{R}^n$ (convex or not), the *convex conjugate* of f is the function

$$\boxed{\ f^\star(\boldsymbol{\xi}) := \sup\{\langle \mathbf{x}, \boldsymbol{\xi}\rangle - f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n\}}$$

for $\boldsymbol{\xi} \in \mathbb{R}^n$.

It is easily shown that $f^\star$ is always closed and convex [2, 7]. Furthermore, if f is closed, proper, and convex, then so is $f^\star$ and the *bi-conjugate* $f^{\star\star} := (f^\star)^\star$ is f itself [2, 7].

> 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 (3) is closed and convex.*

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

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

where C is the convex set

$$\left\{(\xi,\eta) \in \mathbb{R}^2 | \eta \leq -\exp\xi\right\}.$$

This proves that h is closed and convex. ■

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

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

is convex on $I \times (0,\infty)$ if and only if f is convex on I. [A ‘bad’ way is to check the *Hessian* matrix is positive semi-definite.]

$\triangleright$ Figure 1 displays h.

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

$$\begin{aligned} S_m(\mathbf{p}) &= \textstyle\sum_{i=1}^{m} h(p_i, [J\mathbf{p}]_i) \qquad \text{and} \\ \mathcal{S}(p) &= \textstyle\int_0^T h\Big(p(t), [\mathcal{J}p](t)\Big)\, dt, \end{aligned}$$

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

$$[\mathcal{J}p](t) := \frac{1}{T}\int_t^T p(s)\, ds. \tag{6}$$

In passing, we recall that the composition of a convex function with an arbitrary linear mapping is convex.

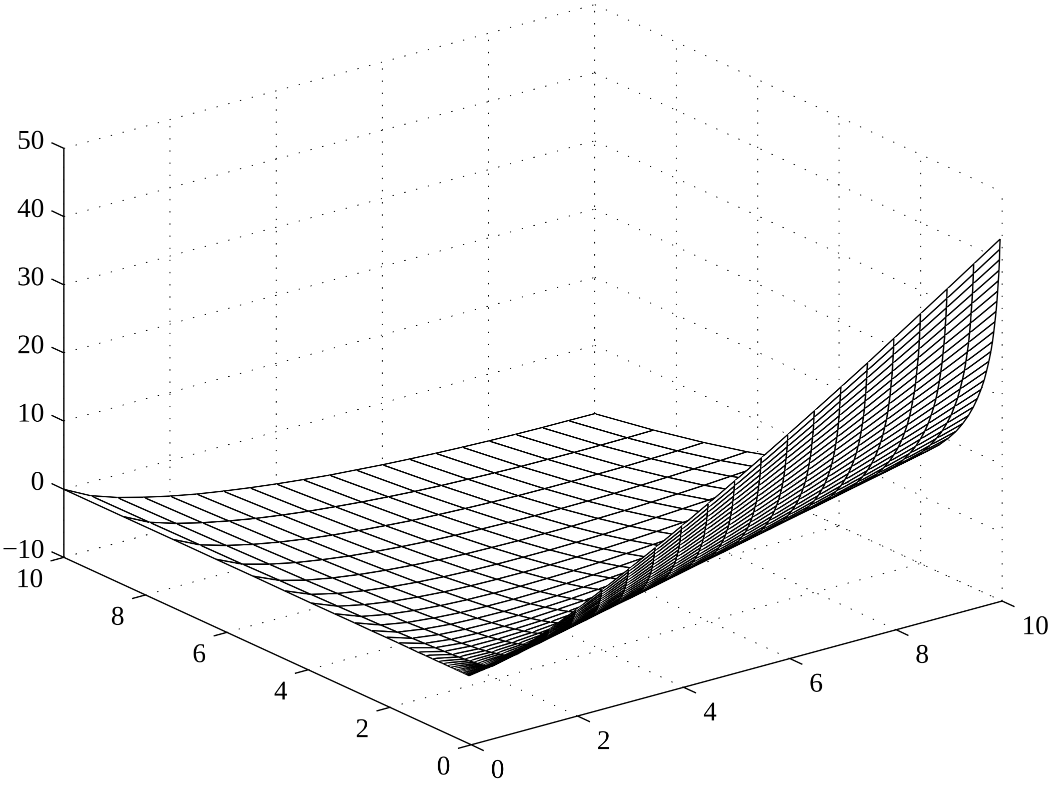


Figure 1. Graph of $(x,y) \mapsto x \log \dfrac{x}{y} - x.$

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 simpler or unconstrained ones.

▷ A modern version of duality theory is posed in the language of Fenchel conjugation [2, 7].

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*

$$(\mathcal{P}) \quad \text{inf}\,\{\,f(\mathbf{x}) \mid \mathbf{x} \in \mathbb{R}^n,\, \mathbf{y} - A\mathbf{x} = \mathbf{0}\}\,. \qquad (7)$$

▷ We denote the *optimal value* of $(\mathcal{P})$ by $V(\mathcal{P})$, the *feasible set* by $F(\mathcal{P})$ and the *solution set* by $S(\mathcal{P})$. Thus,

$$F(\mathcal{P}) := \{\mathbf{x} | \mathbf{y} - A\mathbf{x} = \mathbf{0}\}$$

and

$$S(\mathcal{P}) := \{\mathbf{x} \in F(\mathcal{P}) \,|\, f(\mathbf{x}) = V(\mathcal{P})\}\,.$$

$\triangleright$ The *Lagrangian* of (7) is the function

$$\boxed{\ \mathcal{L}(\boldsymbol{\lambda}, \mathbf{x}) := f(\mathbf{x}) + \langle \boldsymbol{\lambda}, \mathbf{y} - A\mathbf{x} \rangle\,,}$$

for $\boldsymbol{\lambda} \in \mathbb{R}^m$, $\quad \mathbf{x} \in \mathbb{R}^n$. For a given $\boldsymbol{\lambda}$, $\mathcal{L}(\boldsymbol{\lambda}, \mathbf{x})$ 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.

$\triangleright$ Under favourable circumstances, it is possible to find a particular value $\bar{\boldsymbol{\lambda}}$ of $\boldsymbol{\lambda}$ such that minimizers of $\mathcal{L}(\bar{\boldsymbol{\lambda}}, \cdot\,)$ also solve (7). Such a $\bar{\boldsymbol{\lambda}}$ is then called a *Lagrange Multiplier* or a *shadow price*.

$\triangleright$ Now minimizing $\mathcal{L}(\bar{\boldsymbol{\lambda}}, \cdot)$ 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 (7), [7] or [2].

Theorem 1 (Kuhn-Tucker) *Suppose* $V(\mathcal{P}) \neq -\infty$ *and that*

$$\text{(CQ)} \quad \boxed{F(\mathcal{P}) \cap \operatorname{int} \operatorname{dom} f \neq \emptyset.}$$

Then, the following are equivalent:

(i) $\mathbf{x} \in S(\mathcal{P})$;

(ii) $\sup \mathcal{L}(\cdot, \mathbf{x}) = \mathcal{L}(\bar{\boldsymbol{\lambda}}, \mathbf{x}) = \inf \mathcal{L}(\bar{\boldsymbol{\lambda}}, \cdot)$ *for some* $\bar{\boldsymbol{\lambda}}$;

(iii) $\mathbf{x} \in F(\mathcal{P})$ *and* $A^{\star}\bar{\boldsymbol{\lambda}} \in \partial f(\mathbf{x})$ *for some* $\bar{\boldsymbol{\lambda}}$.

$\triangleright$ In condition (iii), $A^\star$ 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 $\mathbf{x}$.

$\triangleright$ Precisely, a vector $\boldsymbol{\xi} \in \mathbb{R}^n$ is a *subgradient* of f at $\mathbf{x}$ if the *subgradient inequality*

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

holds for all $\mathbf{z} \in \mathbb{R}^n$.

If f is convex and differentiable at $\mathbf{x}$, $\nabla f(\mathbf{x})$ is the unique subgradient of f at $\mathbf{x}$, and conversely.

- 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].

$\triangleright$ Points $(\bar{\boldsymbol{\lambda}}, \mathbf{x})$ satisfying condition (ii) are said to be *saddle points* of $\mathcal{L}$.

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*

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

$$\cdots$$

$\triangleright$ We now return to the study of Problem (2).

The *Lagrangian* of (2) is

$$\mathcal{L}(\mathbf{p}, \lambda) := S_m(\mathbf{p}) + \lambda(1 - \langle\, \mathbf{u} \,,\, \mathbf{p} \,\rangle),$$

for $\mathbf{p} \in \mathbb{R}^m, \quad \lambda \in \mathbb{R}$.

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) for some $\bar{\lambda} \in \mathbb{R}$ $\mathbf{0} \in \partial S_m(\mathbf{p}) + \bar{\lambda}\,\partial\Big[\mathbf{1} - \langle \mathbf{u}, \cdot \rangle\Big](\mathbf{p})$.

Indeed, one can check that $V(\mathcal{P}_m) \neq -\infty$ and that $(\mathcal{P}_m)$ has a feasible solution in

$$\operatorname{int}\operatorname{dom} S_m = \{\mathbf{p} \in \mathbb{R}^m \mid \mathbf{p} > \mathbf{0}\}\,.$$

$\triangleright$ Furthermore, S_m is differentiable in the interior of its domain, and we have

$$\frac{\partial S_m}{\partial p_k}(\mathbf{p}) = \log m\mu_k - \sum_{i \leq k} \mu_i,$$

where

$$\mu_k := p_k / \sum_{j \geq k} p_j. \tag{8}$$

▷ Consequently, condition (β) becomes

$$0 = \log m\mu_k - \sum_{i \le k} \mu_i - \lambda, \qquad k = 1, \ldots, m. \tag{9}$$

Now, by definition, $\mu_m = 1$, so setting $k = m$ in (9) gives

$$\lambda = \log m - \sum \mu_i,$$

from which we obtain the recursion

$$\mu_m = 1, \quad \mu_k = \exp\left(-\sum_{j=k+1}^m \mu_j\right), \tag{10}$$

for $k = m-1, \ldots, 1$. Also

$$\begin{aligned} \mu_{k-1} &= \exp\left(-\sum_{j=k}^m \mu_j\right) \\ &= \exp(-\mu_k)\exp\left(-\sum_{j=k+1}^m \mu_j\right). \end{aligned}$$

Thus, the *backward recursion* (10) can be rewritten as

$$\mu_m = 1, \quad \mu_{k-1} = \mu_k \exp(-\mu_k), \tag{11}$$

for $k = m, \ldots, 2$.

▷ Values of μ_k are shown in Figure 2, while Figure 3 shows optimal probability distributions.

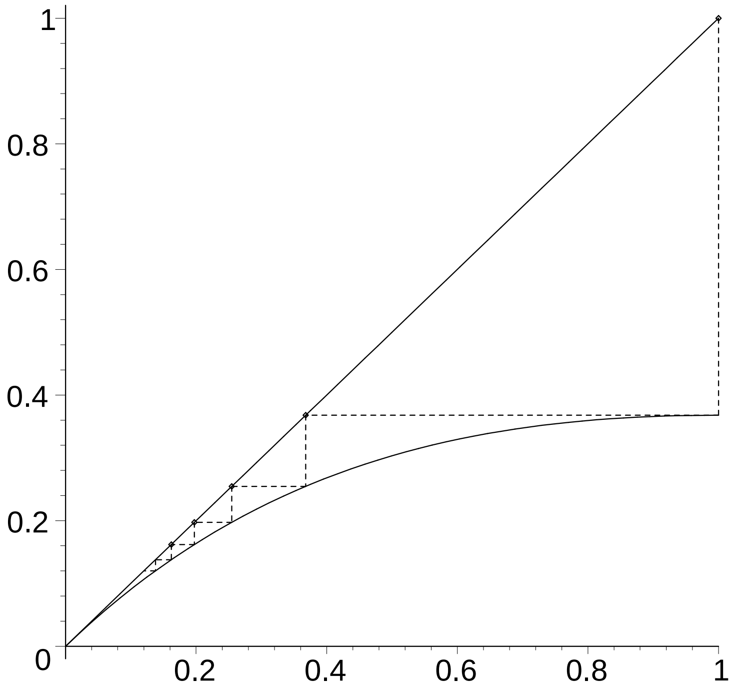


Fig. 2. **Recursion for the μ_k's.**

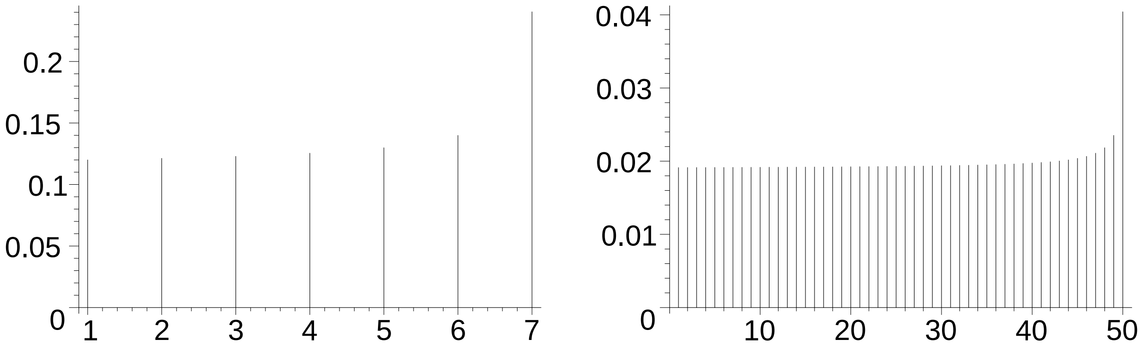


Fig. 3. **Optimal distributions**
$m = 7$ (left) and $m = 50$ (right).

▷ 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*:

$$\begin{aligned} p_1 &= \mu_1, \;\; p_k = \mu_k \times \left(1 - \textstyle\sum_{j=1}^{k-1} p_j\right), \\ k &= 2, \ldots, m. \end{aligned} \tag{12}$$

The vector $\mathbf{p}$ defined in (12) satisfies conditions (α) and (β), and therefore *uniquely* solves Problem ($\mathcal{P}_m$) in (2).

Most pleasingly, the iteration is easy to handle both numerically and theoretically. For example, its components form an increasing sequence. Indeed,

$$p_k = \mu_k \, (p_k + \cdots + p_m)$$

and

$$p_{k-1} = \mu_{k-1} \, (p_{k-1} + \cdots + p_m).$$

$\triangleright$ From whence we deduce, using (11), that

$$\begin{aligned} \frac{p_k}{p_{k-1}} &= \frac{\mu_k \left(1 - \mu_{k-1}\right)}{\mu_{k-1}} \\ &= \exp \mu_k \times \Big(1 - \mu_k \exp(-\mu_k)\Big) \\ &= \exp \mu_k - \mu_k > 1, \end{aligned} \tag{13}$$

since $\mu_k > 0$.

$\triangleright$ We recapitulate the prior discussion as:

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

a. Compute for $j = m, \ldots, 2$

$$\mu_m = 1, \qquad \mu_{j-1} = \mu_j \exp\Big(-\mu_j\Big), \tag{14}$$

and then

b. compute for $k = 2, \ldots, m$

$$p_1 = \mu_1, \;\; p_k = \mu_k \times \left(1 - \sum_{i=1}^{k-1} p_i\right). \tag{15}$$

Remark 1 As 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.

$\triangleright$ This is immediate from (11) 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. $\blacksquare$

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

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

where

$$S'_m(\mathbf{p},\mathbf{q}) := \sum h(p_j, q_j).$$

▷ The needed 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

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

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

$$f(\mathbf{p}, \mathbf{q}) := 1 - \langle \mathbf{u}, \mathbf{p} \rangle$$

and

$$\mathbf{f}(\mathbf{p}, \mathbf{q}) := \mathbf{q} - J\mathbf{p}.$$

▷ 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. ■

HOW THE DISTRIBUTION BEHAVES?

Striking characteristics of the optimal distribution were already shown in Remark 1. We will study asymptotic behaviour of Problem $(\mathcal{P}_m)$ as m tends to infinity.

We now establish three key properties.

$\triangleright$ 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* self-map, g, on $[0, 1]$.

▷ But, when the fixed point, t, has $|g'(t)| = 1$, and so is not *strictly contractive*. Recall that g is contractive if

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

for all $t \neq s$ in $[0, 1]$. We use $x \mapsto x\,\mathrm{exp}(-x)$.

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

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

$$t_i := \mu_{m+1-i}^{(m)}$$

for $i = 1, \ldots, m, \quad m = 1, 2, \ldots$. Observe that t_i is independent of m, that $t_m = p_1^{(m)}$, and satisfies the recursion

$$t_1 = 1, \quad t_{k+1} = t_k\,\mathrm{exp}(-t_k),$$

for $k \geq 1$.

$\triangleright$ 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. ■

$\triangleright$ It is fun to perform a similar analysis for a general $g : [0,1] \mapsto [0,1]$.

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 (13) and the above definition of $\{t_n\}$, that

$$\begin{aligned}
\lim \tfrac{p_m^{(m)}}{p_1^{(m)}} &= \lim_{m\to\infty} \textstyle\prod_{j=2}^{m} (e^{\mu_j^{(m)}} - \mu_j^{(m)}) \\
&= \lim_{m\to\infty} \textstyle\prod_{j=1}^{m-1} (e^{t_j} - t_j) \\
&\simeq 2.132979\ldots.
\end{aligned}$$

The limit exists since

$$1 \le \exp t_j - t_j \le 1 + t_j^2,$$

while $\sum_j t_j^2 < \infty$ by Proposition 1.

Finally recall that $\prod_n (1 + |a_n|)$ and $\sum_n |a_n|$ converge together. ∎

Third – and more subtly - we establish that in the limit our solution value approaches that of the *uniform solution* of the next section.

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

Proof. To establish this, we show that

$$\limsup V(\mathcal{P}_m) \le 0,$$

and that

$$0 \le \liminf V(\mathcal{P}_m).$$

a. The first inequality is easily obtained from identifying a Riemann sum:

$$\begin{aligned} V(\mathcal{P}_m) &\leq S_m\left(\frac{1}{m},\ldots,\frac{1}{m}\right) \\ &= \log m - \frac{\log m!}{m} - 1 \\ &= -\frac{1}{m}\sum_{k=1}^{m} \log\frac{k}{m} - 1 \\ &\to -\int_0^1 \log t\,dt - 1 = 0. \end{aligned}$$

b. obtain the other inequality, consider

$$\tau_m := \sum_{i=1}^{m-1}\left(p_i^{(m)} \log\frac{p_i^{(m)}}{q_{i+1}^{(m)}} - p_i^{(m)}\right)$$

and

$$\sigma_m := \sum_{i=1}^{m-1}\left(p_i^{(m)} \log\frac{p_i^{(m)}}{q_i^{(m)}} - p_i^{(m)}\right).$$

▷ We make two claims:

(i) $\tau_m - \sigma_m$ tends to 0 as m tends to infinity;

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

Proof of (i). We recall from (4) and (8) that $\mu_i^{(m)} = p_i^{(m)}/(mq_i^{(m)})$ and so

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

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

$$\begin{aligned} 0 &\leq \tau_m - \sigma_m = -\textstyle\sum_{i=1}^{m-1} p_{m-i}^{(m)} \log(1 - t_{i+1}) \\ &\leq -p_m^{(m)} \textstyle\sum_{i=1}^{m-1} \log(1 - t_{i+1}) \to 0, \end{aligned}$$

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

The proof of (ii) is deferred to the next section where it is a consequence of a general integral inequality.

▷ Now, by design,

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

It follows from (ii) that

$$V(\mathcal{P}_m) \geq \sigma_m - \tau_m - p_m^{(m)}.$$

And so, since

$$p_m^{(m)} \to 0,$$

(i) shows

$$\liminf V(\mathcal{P}_m) \geq 0$$

as needed. ■

▷ These techniques allow much more precise assertions about the asymptotics of $\mathbf{p}^m$.

CONTINUOUS TIME ANALYSIS

In the discrete case, the distribution is strictly increasing, with a sharp increase at the tip of the tail (see Figure). 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*

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

– equivalently $\mathcal{S}(p) \geq 0$ – with equality if and only if p is constant on $[0,T]$.

Proof. Without loss p is (a.e.) nonnegative, else $\mathcal{S}(p) = \infty$.

As in (6), set

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

On integrating by parts,

$$\begin{aligned}
\mathcal{S}(p) &= \int_0^T \left(p(t)\,\mathsf{log}\,\frac{p(t)}{q(t)} - p(t)\right)\,dt \\
&= \int_0^T \Big(p(t)\,\mathsf{log}\,p(t) - p(t)\Big)\,dt \\
&\qquad\qquad\qquad + T\textstyle\int_0^T q'(t)\,\mathsf{log}\,q(t)\,dt \\
&= \int_0^T p(t)\,\mathsf{log}\,p(t)\,dt - Tq(0)\,\mathsf{log}\,q(0),
\end{aligned}$$

▷ We shall be done once we show

$$\int_0^T p(t)\,\mathsf{log}\,p(t)\,dt \geq Tq(0)\,\mathsf{log}\,q(0). \qquad (16)$$

with equality if and only if p is constant.

But, applying the integral version of *Jensen's inequality* to the strictly convex function $g := x \mapsto x \log x - x$ yields

$$\frac{1}{T}\int_0^T \left(\frac{p(t)}{q(0)} \log \frac{p(t)}{q(0)} - \frac{p(t)}{q(0)}\right) dt$$

$$\geq g(1) = -1,$$

from which (16) follows immediately. ■

▷ Theorem 2 shows that the (unique) solution of Problem ($\mathcal{P}$) given in (5) is the uniform probability density on $[0, T]$.

▷ A consequence of Theorem 2, which completes the considerations of the last Section, follows:

Corollary 1 *As claimed in Section ,*

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

Proof. Apply Theorem 2 with

$$T := 1 \quad \text{and} \quad p(t) := p_n^{(m)}$$

if

$$t \in \left(\frac{n-1}{m}, \frac{n}{m}\right] \quad (n = 1, \ldots, m).$$

For $\frac{n-1}{m} < t \leq \frac{n}{m}$ and $n \leq m-1$,

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

and, for $\frac{m-1}{m} < t \leq 1$,

$$q(t) = p_m^{(m)}(1-t).$$

Hence τ_m majorizes

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

on evaluating the second integral and applying Theorem 2. ■

▷ This finishes the proof that the optimal value of $(\mathcal{P}_m)$ tends to 0 ($= V(\mathcal{P})$), as claimed above.

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.

REFERENCE. D. Borwein, J. M. Borwein and P. Marchal, “Surprise maximization,” *The American Mathematical Monthly*, **107** June-July 2000, 527–537. [CECM Preprint 98:116].

`www.cecm.sfu.ca/preprints/1998pp.html`

Open Question. How does one quantify *average multiple surprise*?

REFERENCES

- [1] **URL** 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 Books, Vol. 3, 2000 and 2005.

- [3] T. Y. Chow, *The surprise examination or unexpected hanging paradox*, MAA Monthly, 105 (1998), pp. 41–51.

- [4] H. Gzyl, *The Method of Maximum Entropy*, World Scientific, 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.

- [7] R. T. Rockafellar, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970.