

# Surprise Maximization <sup>\*</sup>

D. Borwein<sup>†</sup>, J.M. Borwein<sup>‡</sup> and P. Maréchal<sup>‡</sup>

<sup>†</sup> Department of Mathematics  
University of Western Ontario  
London, Ontario, Canada N6A 5B7

<sup>‡</sup> Centre for Experimental and Constructive Mathematics  
Department of Mathematics and Statistics  
Simon Fraser University  
Burnaby, B.C., Canada V5A 1S6

December 7, 1998

## Abstract

The optimization problems arising from an information theoretic formulation of the *Surprise Examination* (or *Unexpected Hanging*) *Paradox* are examined and solved. They provide a nice application of both the Kuhn-Tucker Theorem and Jensen's inequality.

## Keywords

Surprise examination paradox, entropy optimization, Kuhn-Tucker Theorem, Jensen inequality.

## 1 Introduction

In this paper, we study the optimization problems arising from an entropic approach to the so-called *Surprise Examination* (or *Unexpected Hanging*) *Paradox*. The idea of such an approach was proposed by Karl Narveson and mentioned in a recent article by Timothy Y. Chow [1]. (We shall not discuss here the different approaches to the resolution of the paradox itself; the reader interested in these aspects is invited to consult [1].)

---

<sup>\*</sup>Research supported by NSERC (first two authors), the Shrum Endowment (second author) and the Pacific Institute for the Mathematical Sciences (third author).

An event (such as a test given by a teacher or a surprise tax audit) occurs once every  $m$  days, with probability  $p_i$  on day  $i$ ,  $i = 1, \dots, m$ . 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 definition of the Shannon entropy (see [4], for example). The surprise on day  $i$  will be the negative of the logarithm of the probability that the event occurs on day  $i$  given that it has not occurred so far. The event ‘*test occurs on day  $i$* ’ will be simply denoted by  $i$ , and its probability will be denoted by  $P(i)$  or  $p_i$ . The event ‘*test does not occur on day  $i$* ’ will be denoted by  $\sim i$ . The quantity to be maximized can therefore be written as

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

Using Bayes’ formula for conditional probabilities, we obtain

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

Consequently, we are led to consider the following optimization problem:

$$(P_m) \quad \inf \{ S_m(\mathbf{p}) \mid \mathbf{p} \in \mathbb{R}^m, \mathbf{1} = \langle \mathbf{u}, \mathbf{p} \rangle \}.$$

Here,  $\mathbf{u}$  is the vector of  $\mathbb{R}^m$  whose entries are all equal to 1 and  $S_m$  is the extended real-valued function defined by

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

More precisely,

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

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

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

For all  $\mathbf{p}$  satisfying the constraint in (P),  $S_m(\mathbf{p})$  differs from the negative of the quantity in (1) only by a constant. (The factor  $1/m$  in its definition was introduced so as to make subsequent computations more esthetic.) Note that  $S_m(\mathbf{p})$  can be regarded as Kullback-Leibler's information measure<sup>1</sup> of  $\mathbf{p}$  relative to its *tail*  $\mathbf{q}$ :

$$\mathbf{q} = (q_1, \dots, q_m) \quad \text{with} \quad q_i := \frac{1}{m} \sum_{j=1}^m p_j, \quad i = 1, \dots, m.$$

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

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

in which  $\mathcal{S}$  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 A preliminary result

In this paragraph, we establish the convexity of (the negative of) our measures of surprise.

---

<sup>1</sup>Kullback-Leibler's information measure is an extension of Boltzmann-Shannon's entropy. It is also referred to as *relative information measure*, *cross-entropy* or *I-divergence*. Given to probability measures  $P$  and a reference measure  $Q$  on a probability space, the information of  $P$  relative 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. The reader interested in the statistical meaning of this measure may refer to [3]. For a discussion on the Maximum Entropy Principle, see [2] and references therein.

Recall that an extended real-valued function on  $\mathbb{R}^n$  is said to be *closed* if its *epigraph* (i.e. the set of points which are above or on its graph) is closed (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. Given any function  $f$  on  $\mathbb{R}^n$  (convex or not), the *convex conjugate* of  $f$  is the function

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

It can be shown that  $f^*$  is always closed and convex. Furthermore, if  $f$  is closed, proper and convex, then so is  $f^*$  and the *bi-conjugate*  $f^{**} := (f^*)^*$  is  $f$  itself (cf. [5], Theorem 12.2 and corollaries).

**Lemma 1** *The function  $h$  defined in Section 1 is closed and convex.*

**Proof** One can easily show that  $h$  is the convex conjugate of the *indicator function*

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

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

Indicator functions play the same role in convex analysis as characteristic in measure theory. Figure 1 displays the graph of the function  $h$ . Notice that convexity of  $h$  can also be derived from the easily proved fact that a function  $(x, y) \mapsto y f(x y^{-1})$  on  $I \times (0, \infty)$  is convex if and only if  $f$  is convex on  $I$ , where  $I$  is any interval. (Take  $f(s) = s \ln s - s$ .)

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

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

in which  $J$  is the  $(m \times m)$ -matrix whose entries are  $m^{-1}$  in the upper triangle (including the diagonal) and 0 elsewhere, and  $\mathcal{J}$  is the linear mapping defined by

$$\begin{aligned} \mathcal{J}: \quad L_1([0, T]) &\longrightarrow \mathcal{C}([0, T]) \\ p &\longmapsto [\mathcal{J}p](t) := \frac{1}{T} \int_t^T p(s) ds. \end{aligned}$$

(The composition of a convex function with a linear mapping is convex.)

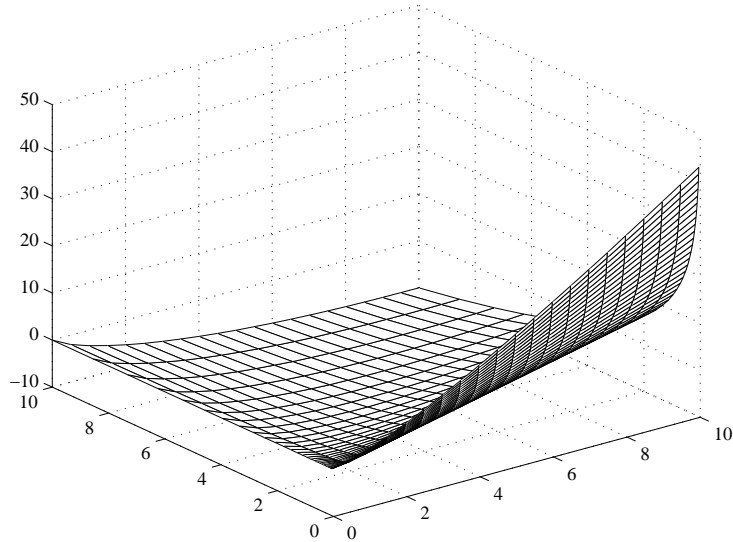


Figure 1: Graph of  $(x, y) \mapsto x \ln \frac{x}{y} - x$ .

### 3 Discrete time analysis

Constrained optimization problems such as  $(P_m)$  are traditionally approached using concepts from *duality theory*. Such concepts can be traced back to the theory of *Lagrange multipliers*. More recent highlights are the Fenchel duality theorem and the Kuhn-Tucker Theorem (cf. [5]). Roughly speaking, they provide techniques making it possible to solve constrained optimization problems via unconstrained ones. We now recall some useful results.

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

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

We shall denote the optimal value of  $(P)$  by  $V(P)$ , the *feasible set* by  $F(P)$  and the *solution set* by  $S(P)$ <sup>2</sup>. The *Lagrangian* of Problem  $(P)$  is the function

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

For a given  $\boldsymbol{\lambda}$ ,  $\mathcal{L}(\boldsymbol{\lambda}, \mathbf{x})$  can be regarded as a *penalized* version of  $f$ . Each component of  $\boldsymbol{\lambda}$  fixes the *price to be paid* if the corresponding constraint is violated. Under favorable circumstances, it is possible to find a particular

<sup>2</sup>Recall that  $F(P) := \{\mathbf{x} \mid \mathbf{y} - A\mathbf{x} = \mathbf{0}\}$  and that  $S(P) := \{\mathbf{x} \in F(P) \mid f(\mathbf{x}) = V(P)\}$ .

value  $\bar{\lambda}$  of  $\lambda$  such that minimizers of  $\mathcal{L}(\bar{\lambda}, \cdot)$  also solve  $(P)$ . Such a  $\bar{\lambda}$  is then called a *Lagrange Multiplier* or a *shadow price*. (Note that minimizing  $\mathcal{L}(\bar{\lambda}, \cdot)$  is an unconstrained problem.) The Kuhn-Tucker Theorem will provide necessary and sufficient conditions (on  $\lambda$  and  $\mathbf{x}$ ) for  $\mathbf{x}$  to be a solution of  $(P)$ . Finally, recall that the *domain* of a convex function  $f$  is the set of points where it is less than  $+\infty$ . It is denoted by  $\text{dom } f$ .

We can now state the Kuhn-Tucker Theorem, a proof of which can be found in [5], Section 28.

**Theorem 1 (Kuhn-Tucker)** *Suppose that  $V(P) \neq -\infty$  and that  $F(P)$  meets the interior of  $\text{dom } f$ . Then, the following are equivalent:*

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

In the last condition,  $A^*$  is the transpose of  $A$  and  $\partial f(\mathbf{x})$  denotes the *subdifferential* of  $f$  at  $\mathbf{x}$ , i.e. the set of *subgradients*<sup>3</sup> of  $f$  at  $\mathbf{x}$ .

Points  $(\bar{\lambda}, \mathbf{x})$  satisfying Condition (ii) are said to be *saddle points* of  $\mathcal{L}$ . The conditions in (iii) are called the Kuhn-Tucker conditions. Notice that, in Condition (ii),  $\bar{\lambda}$  appears as the maximizer of the (concave) *dual function*  $D(\lambda) := \inf \mathcal{L}(\lambda, \cdot)$ .

Note finally that  $\text{dom } f$  may have an empty interior. Theorem 1 will still apply, however, with the weaker assumption that  $F(P)$  intersects the *relative interior* of  $\text{dom } f$  [5], that is the interior relative to the smallest affine manifold containing  $\text{dom } f$ .

We now return to the study of  $(P_m)$ . The *Lagrangian* of Problem  $(P_m)$  is the function

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

Theorem 1 tells us that, for  $\mathbf{p}$  to be a solution for  $(P_m)$  it is necessary and sufficient that

---

<sup>3</sup>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  $(\mathbf{x}, f(\mathbf{x}))$ ” (cf. [5], Section 23). If  $f$  is convex and differentiable at  $\mathbf{x}$ ,  $\nabla f(\mathbf{x})$  is the unique subgradient of  $f$  at  $\mathbf{x}$ , and conversely.

( $\alpha$ )  $0 = 1 - \langle \mathbf{u}, \mathbf{p} \rangle$ ;

( $\beta$ ) 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 easily check that  $V(P_m) \neq -\infty$  and that  $(P_m)$  has a feasible solution in

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

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

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

Consequently, Condition ( $\beta$ ) becomes

$$0 = \ln m \mu_k - \sum_{i \leq k} \mu_i - \lambda, \quad k = 1, \dots, m. \quad (2)$$

Now, by definition,  $\mu_m = 1$ , so that the last of Equations (2) gives  $\lambda = \ln m - \sum \mu_i$ , from which we obtain the recursion

$$\mu_m = 1, \quad \mu_k = \exp\left(-\sum_{j=k+1}^m \mu_j\right), \quad k = m-1, \dots, 1. \quad (3)$$

Note that, 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 above recursion can be rewritten as

$$\mu_m = 1, \quad \mu_{k-1} = \mu_k \exp(-\mu_k), \quad k = m, \dots, 2. \quad (4)$$

The values of the  $\mu_k$ 's can be obtain as shown in Figure 2 below. Figure 3 shows examples of optimal probability distributions, for  $m = 7$  and  $m = 50$ .

Finally, from Condition ( $\alpha$ ) above and the values of the  $\mu_k$ 's, we see that the components of  $\mathbf{p}$  must obey the following recursion :

$$p_1 = \mu_1, \quad p_k = \mu_k \times \left(1 - \sum_{j=1}^{k-1} p_j\right), \quad k = 2, \dots, m. \quad (5)$$

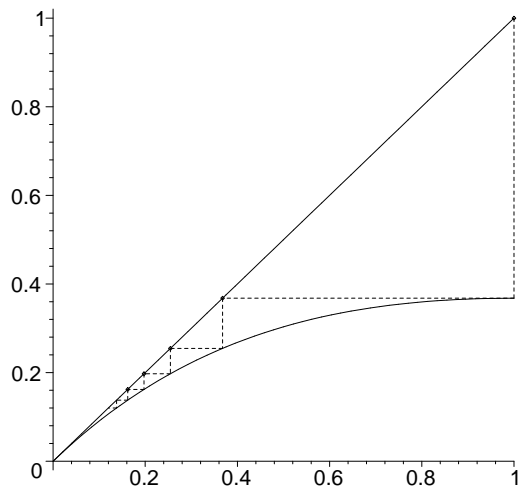


Figure 2: Recursion for the  $\mu_k$ 's.

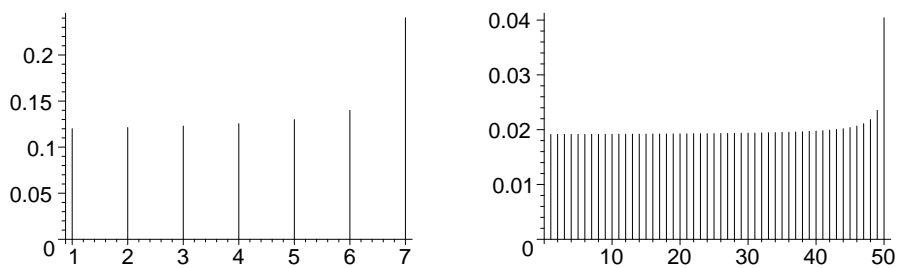


Figure 3: Optimal distributions for  $m = 7$  (left) and  $m = 50$  (right).

The vector  $\mathbf{p}$  defined above satisfies Conditions  $(\alpha)$  and  $(\beta)$ , and therefore solves Problem  $(P)$ . Moreover, its components form an increasing sequence. As a matter of fact, we have

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

from which we deduce, using (4), that

$$\begin{aligned} \frac{p_k}{p_{k-1}} &= \frac{\mu_k (1 - \mu_{k-1})}{\mu_{k-1}} \\ &= \exp \mu_k \times (1 - \mu_k \exp(-\mu_k)) \\ &= \exp \mu_k - \mu_k \\ &> 1. \end{aligned} \tag{6}$$

**Remark** As pointed out in [1], the (optimal) probability that the event occurs on the  $i$ th-to-the-last day, given that it has not occurred so far does not depend on  $m$ . This is immediate from Recursion (4) and from the equality

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

Furthermore, the fact that the  $\mu_k$ 's are defined via a backward recursion implies that  $p_{m-i}/p_{m-i-1}$  does not depend on  $m$  either (cf. Eq. (6)). ■

## 4 Transitional comments

Note first that we could have obtained the solution by considering the optimization problem

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

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

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

$$(\beta') \quad \text{there exist } \lambda \in \mathbb{R} \text{ and } \boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m) \in \mathbb{R}^m \text{ such that}$$

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

in which we have defined  $f$  and  $\mathbf{f} = (f_1, \dots, f_m)$  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  $\lambda_j$ 's derived from  $(\alpha')$  and  $(\beta')$  coincide with the  $\mu_j$ 's of the previous paragraph multiplied by  $m$ .

Some immutable characteristics of the optimal probability distribution were pointed out in the remark. It is also interesting to consider asymptotic properties of Problem  $(P_m)$  when  $m$  tends to infinity. We shall simply mention here three facts.

1) The ratio between the last and the first components of  $\mathbf{p}^{(m)}$  tends to a finite value. Indeed, from Eq. (6), we get

$$\lim_{m \rightarrow \infty} \frac{p_m^{(m)}}{p_1^{(m)}} = \lim_{m \rightarrow \infty} \prod_{j=2}^m (\exp \mu_j^{(m)} - \mu_j^{(m)}) \simeq 2.13.$$

The limit exists because of the inequality  $1 \leq \exp \mu_j^{(m)} - \mu_j^{(m)} \leq 1 + (\mu_j^{(m)})^2$  for  $\mu_j^{(m)} \in [0, 1]$ , and because  $\sum_j (\mu_j^{(m)})^2 < \infty$ .

2)  $mp_1^{(m)}$  tends to 1 as  $m$  tends to  $\infty$ . As a matter of fact, let us denote  $t_m = p_1^{(m)}$ . Then we see that the sequence  $\{t_m\}$  is defined by the recurrence

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

We deduce that  $t_{k+1}^{-1} - t_k^{-1} = t_k^{-1}(\exp(t_k) - 1)$ , which tends to  $\exp'(0) = 1$  as  $k$  tends to infinity (since  $t_k$  tends to 0). Consequently,

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

also tends to 1.

3) The optimal value  $V(P_m)$  of  $(P_m)$  tends to 0 as  $m$  tends to infinity. To prove this claim, we show that  $\liminf V(P_m) = 0 = \limsup V(P_m)$ . The first equality is easily obtained from

$$\begin{aligned} V(P_m) &\leq S_m \left( \frac{1}{m}, \dots, \frac{1}{m} \right) = \ln m - \frac{\ln m!}{m} - 1 \\ &= -\frac{1}{m} \sum_{k=1}^m \ln \frac{k}{m} - 1 \rightarrow -\int_0^1 \ln t \, dt - 1 = 0. \end{aligned}$$

The second equality results from the following three facts :

- (i)  $V(P_m) = \sigma_m + p_m^{(m)} \ln m - p_m^{(m)}$ ;
- (ii)  $\tau_m - \sigma_m$  tends to 0 as  $m$  tends to infinity;

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

Here, we have defined

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

Fact (i) is an immediate consequence of the definitions. As for (ii), we have

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

since  $t_i \rightarrow 0$  and  $mp_m^{(m)} = O(1)$ . Finally the proof of (iii) is deferred to the end of Section 5 (cf. Corollary 1) since it relies on Theorem 2 below.

## 5 Continuous time analysis

In the discrete case, the distribution is strictly increasing, with a sharper increase at the tip of the *tail* (see Figure 3). By contrast, the optimal continuous distribution is flat, as the following theorem shows.

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

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

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

**Proof** We can assume that  $p$  is (almost everywhere) nonnegative, for otherwise  $\mathcal{S}(p) = \infty$ . Observe that

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

in which we have put  $q(t) := [\mathcal{J}p](t)$ . The theorem will therefore be proved if we can show that

$$\int_0^T p(t) \ln p(t) dt \geq Tq(0) \ln q(0), \quad (7)$$

with equality if and only if  $p$  is constant. Now, applying Jensen's integral inequality<sup>4</sup> to the strictly convex function  $g: x \mapsto x \ln x - x$  yields

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

from which (7) follows immediately. ■

Theorem 2 shows that the (unique) solution of Problem ( $\mathcal{P}$ ) is the uniform probability density on  $[0, T]$ . Another immediate consequence of Theorem 2, which brings us back to the considerations of Section 4 is the following result:

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

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

**Proof** Apply Theorem 2 with

$$T := 1 \quad \text{and} \quad p(t) := p_i^{(m)} \quad \text{if } t \in \left( \frac{i-1}{m}, \frac{i}{m} \right] \quad (i = 1, \dots, m). \quad \blacksquare$$

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

## 6 Conclusion

The entropic formulation of the Surprise Examination problem provides a beautiful example of the application of concepts from the elementary theory of convex constrained optimization. Its attractiveness comes in part from the explicit recursive nature of the (discrete time) solution which follows from the Kuhn-Tucker Theorem.

Greg Chaitin recently communicated to us his opinion that the entropy analysis is :

---

<sup>4</sup>In our case Jensen's integral inequality reads :

$$g \left( \frac{1}{T} \int_0^T p(t) dt \right) \leq \frac{1}{T} \int_0^T g(p(t)) dt$$

where  $g$  is convex on the positive half line and  $p$  is nonnegative and integrable on  $[0, T]$ .

*surprising* because everyone (including me) usually just throws up their hands at the paradox. Gödel didn't and got his incompleteness result. You don't and you determine the (surprisingly flat) distribution that maximizes the surprise.

*Acknowledgements* : We thank the anonymous referees for their valuable suggestions which improved the presentation of this work.

## References

- [1] T. Y. CHOW, *The surprise examination or unexpected hanging paradox*, The American Mathematical Monthly, 105 (1998), pp. 41–51.
- [2] H. GZYL, *The Method of Maximum Entropy*, World Scientific, (1995).
- [3] S. KULLBACK, *Information Theory and Statistics*, Dover, New York, (1968).
- [4] 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.
- [5] R. T. ROCKAFELLAR, *Convex Analysis*, Princeton University Press, Princeton, N.J., 1970. Princeton Mathematical Series, No. 28.