entropy

MDPI

Article

# The Expected Missing Mass under an Entropy Constraint

**Daniel Berend [1,2], Aryeh Kontorovich [2,*] and Gil Zagdanski [1]**

[1] Department of Mathematics, Ben-Gurion University, Beer Sheva 84105, Israel; berend@cs.bgu.ac.il (D.B.); gilz1984@gmail.com (G.Z.)

[2] Department of Computer Science, Ben-Gurion University, Beer Sheva 84105, Israel

* Correspondence: karyeh@cs.bgu.ac.il; Tel.: +972-8-642-8048

Received: 7 June 2017 ; Accepted: 26 June 2017; Published: 29 June 2017

**Abstract:** In Berend and Kontorovich (2012), the following problem was studied: A random sample of size $t$ is taken from a world (i.e., probability space) of size $n$; bound the expected value of the probability of the set of elements not appearing in the sample (unseen mass) in terms of $t$ and $n$. Here we study the same problem, where the world may be countably infinite, and the probability measure on it is restricted to have an entropy of at most $h$. We provide tight bounds on the maximum of the expected unseen mass, along with a characterization of the measures attaining this maximum.

**Keywords:** missing mass; probability estimate; sampling; entropy

## 1. Introduction

Let $S$ be a finite probability space. Without loss of generality, suppose that $S = \{1, 2, ..., n\}$. Let $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ be a probability measure on $S$. Suppose that a random sample $X_1, X_2, \ldots, X_t$ is drawn from $S$ according to $\mathbf{p}$. The *missing mass* is the random variable $U_t$, defined by:

$$U_t = \sum_{i=1}^{n} p_i \mathbf{1}_{\{X_1 \neq i \wedge \ldots \wedge X_t \neq i\}}.$$

In words, $U_t$ is the total probability mass of the set of those elements of $S$ not observed at all in the sample. According to the definition of $U_t$, it is easy to verify that $\mathbb{E}U_t = \sum_{i=1}^{n} p_i(1-p_i)^t$. When we wish to make the dependence on the measure $\mathbf{p}$ explicit, we will write $\mathbb{E}_{\mathbf{p}}U_t$ instead of $\mathbb{E}U_t$.

One of the earliest mentions of the *missing mass* is in Good–Turing frequency estimation [1]. The latter is a statistical technique for estimating the probability of encountering an object of a hitherto unseen species, given a set of past observations of objects from different species. This estimator has been used extensively in many machine learning tasks. For example, in the field of natural language modeling, for any sample of words, there is a set of words not occurring in that sample. The total probability mass of the words not in the sample is the so-called *missing mass* [2]. Another example of using Good–Turing *missing mass* estimation is in [3], where the total summed probability of all patterns not observed in the training data is estimated. In [4], Berend and Kontorovich showed that the expectation of the *missing mass* is bounded above as follows:

$$\mathbb{E}U_t \leq \begin{cases} e^{-\frac{t}{n}}, & t \leq n, \\ \frac{n}{et}, & t > n. \end{cases}$$

(Additionally, deviation bounds were provided in [5].)

Moreover, they have shown that:

1. Every local maximum $\mathbf{p}$ of $\mathbb{E}U_t$ is of the form

$$p_1 = p_2 = \ldots = p_{n-1} \leq p_n,$$

(where without loss of generality we consider only vectors $\mathbf{p}$ with $p_1 \leq p_2 \leq ... \leq p_n$). That is, $\mathbf{p}$ consists of one "heavy" atom and $n-1$ "light" ones of identical size, where the possibility of "heavy" = "light" is not excluded.
2. There exists a threshold $\tau = \tau(n) > n$ such that:

   (a) For $t \leq \tau$, there is a unique global maximum:

$$p_1 = p_2 = \ldots = p_{n-1} = p_n = \frac{1}{n}.$$

   (b) For $t > \tau$, there is a unique global maximum, and it has the form:

$$p_1 = p_2 = \ldots = p_{n-1} < p_n.$$

For an infinitely countable set $S$, one cannot generally provide a non-trivial upper bound on $\mathbb{E}U_t$ in terms of $t$ only. Indeed, for each $n$, consider the probability measure on $\mathbb{N}$ supported on $\{1, 2, ..., n\}$, giving equal probabilities to these $n$ atoms. Clearly, $\mathbb{E}U_t \geq 1 - \frac{t}{n}$ in this case, and the right-hand side becomes arbitrarily close to 1 as $n$ grows. In [4] it was shown that

$$\mathbb{E}U_t \leq \frac{l(\mathbf{p})}{ct}, \tag{1}$$

where $l(\mathbf{p})$ roughly measures the size of sets of atoms of comparable mass and $c$ is a universal constant (for an exact definition, we refer to [4]). The bound given in (1) is non-trivial only if the sequence $(p_i)_{i=1}^{\infty}$ decreases "sufficiently fast". Such results may be useful, as shown in [6–9]. Another possible restriction that makes the problem interesting is that the entropy of $\mathbf{p}$ is bounded above by some given value. A similar restriction can be found in [10] in the context of discrete distribution estimation under $\ell_1$ loss. In this work, we study the possibility of providing tight bounds on $\mathbb{E}U_t$ under the restriction of some bound on the entropy. Thus, we can formulate our problem as follows:

$$\sup_{\mathbf{p}} \sum_{i \in S} p_i (1 - p_i)^t \tag{2}$$

subject to

$$\sum_{i \in S} p_i = 1, \tag{3}$$

$$\sum_{i \in S} p_i \ln \frac{1}{p_i} \leq h, \tag{4}$$

where $h \geq 0$ is the maximal allowed entropy.

In the case of distributions over countably infinite spaces we set $S = \mathbb{N}$; otherwise, $S = \{1, 2, ..., n\}$, or in short, $S = [n]$. Note that we are looking for the supremum, since in the case $S = \mathbb{N}$ it is not a priori clear that the maximum exists (in fact, it turns out that the maximum does exist—see Theorem 2). Additionally, we will show that the maximum is obtained for a measure with finite support, which leads us to study the problem for the case of distributions over finite spaces. We also study the structure of local and global maxima and obtain some results analogous to [4].

## 2. Main Results

Our first result is that in the case of $S = \mathbb{N}$, an optimal solution exploits all the available entropy. Denote the entropy of a probability measure $\mathbf{p}$ by $H(\mathbf{p})$:

$$H(\mathbf{p}) = \sum_{i \in S} p_i \ln \frac{1}{p_i}.$$

**Proposition 1.** *Let $S = \mathbb{N}$, and let $\mathbf{p} = (p_1, p_2, ...)$ be a probability measure on $S$. If $H(\mathbf{p}) < h$, then there exists a probability measure $\mathbf{p}' = (p'_1, p'_2, ...)$ on $S$ for which $H(\mathbf{p}') = h$ and $\mathbb{E}_{\mathbf{p}'} U_t > \mathbb{E}_{\mathbf{p}} U_t$.*

**Corollary 1.** *In the problem given by (2)–(4), we may replace (4) by*

$$\sum_{i=1}^{\infty} p_i \ln \frac{1}{p_i} = h.$$

In Theorem 1, we refer to the case $S = [n]$ and show that an optimal solution of (2)–(4) cannot assume more than four distinct non-zero values.

**Theorem 1.** *Let $S = [n]$, and $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ be any locally optimal solution. Then, the $p_i$'s assume at most four non-zero values; i.e., if the $p_i$'s are sorted, then for some indices $j, k, l$ and $m$, we have $p_1 = \ldots = p_j = 0 < p_{j+1} = \ldots = p_k \leq p_{k+1} = \ldots = p_l \leq p_{l+1} = \ldots = p_m \leq p_{m+1} = \ldots = p_n$.*

We do not know whether in some cases there are indeed atoms of four distinct sizes in the optimal solution.

In the case $S = [n]$, it is easy to see that $\mathbb{E}_{\mathbf{p}} U_t$ is continuous with respect to $\mathbf{p}$, and thus $\mathbb{E}U_t$ attains its maximum. On the other hand, when $S = \mathbb{N}$, it is not a priori clear. Our next result shows that $\mathbb{E}U_t$ attains its maximum in this case as well.

**Theorem 2.** *Let $S = \mathbb{N}$.*

*(i) For each $h > 0$, the function $\mathbb{E}U_t$ attains its maximum.*
*(ii) If $\mathbf{p}$ is a global maximum point of $\mathbb{E}U_t$, then $\mathbf{p}$ has a finite support.*

Denote $H_t = 1 + \frac{1}{2} + \frac{1}{3} + \ldots + \frac{1}{t}$. Recall that $\ln t \leq H_t \leq \ln t + 1$ for each $t$.

**Theorem 3.** *For all $t \geq 1$, we have $\mathbb{E}U_t \leq \dfrac{h}{H_{2t-1}}$.*

In particular, $\mathbb{E}U_t \leq \dfrac{h}{\ln t}$.

In Theorem 4 we show that given a fixed $h$, we cannot significantly improve the upper bound from Theorem 3.

**Theorem 4.** *For fixed $h$ and every $\alpha > 1$, if $t$ is large enough then there exists a distribution $\mathbf{p}$ with $H(\mathbf{p}) \leq h$, for which $\mathbb{E}_{\mathbf{p}} U_t \geq \dfrac{h}{\alpha \ln t}$.*

As mentioned earlier, the parameter $t$ represents the size of the sample. It appears that the optimization problem cannot be solved analytically for any fixed arbitrary $t$. The following results relate to the case $t = 1$. Obviously, this case is not typical, as one would hardly try to learn much from a sample of size 1. Yet, it may be instructive, as in this case we obtain almost the best possible results.

**Proposition 2.** $\mathbb{E}U_1 \leq 1 - e^{-h}, \quad h > 0.$

Next, we describe the structure of an optimal solution for the case $S = [n]$ with $t = 1$.

**Proposition 3.** *Let $S = [n]$, and let $\mathbf{p} = (p_1, p_2, \ldots, p_n)$ be an optimal solution of the problem*

$$\max \sum_{i=1}^{n} p_i(1 - p_i) \tag{5}$$

*subject to*

$$\sum_{i=1}^{n} p_i = 1, \tag{6}$$

$$\sum_{i=1}^{n} p_i \ln \frac{1}{p_i} = h, \tag{7}$$

*where $\ln(k-1) < h \leq \ln k$, $k \leq n$. Then (after sorting), $\mathbf{p}$ is of the form*

$$p_1 = p_2 = \ldots = p_{n-k} = 0, \; p_{n-k+1} \leq p_{n-k+2} = \ldots = p_n.$$

*That is, the non-zero atoms of $\mathbf{p}$ consist of one "light" atom and $k-1$ "heavy" ones.*

Denote the mass $p_{n-k+1}$ of the light atom in the proposition by $p$ and that of the heavy ones by $q$. In view of Proposition 3, it suffices to consider the case $k = n$, namely $\ln(n-1) < h \leq \ln n$. For $\ln(n-1) < h \leq \ln n$, the following proposition gives a tight upper bound on $\mathbb{E}U_1$:

**Proposition 4.** $\mathbb{E}U_1 \leq 1 - e^{-h} - \left| e^{-h}(2p - 1 + 2p \ln p + 2ph) + (n-1)q^2 - p^2 \right|$.

**Remark 1.** *When $h = \ln n$ (or $k = \ln(n-1)$), there is an equality without the last term (see the proof of Proposition 3). At these points, the last term indeed vanishes. Inside the interval $(\ln(n-1), \ln n)$, Proposition 4 provides an improvement over Proposition 2. In Figure 1, we plot the exact value of max $\mathbb{E}U_1$ (calculated numerically using MATLAB) against the bounds of Propositions 2 and 4 for $h \in [\ln 3, \ln 4]$. It appears that the additional term in Proposition 4 captures most of the error in Proposition 2.*

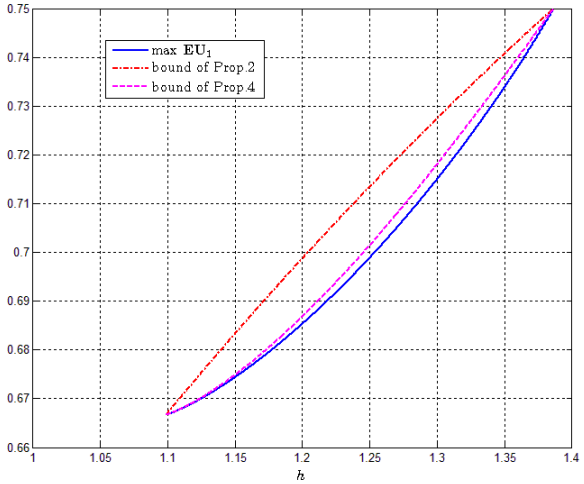


**Figure 1.** Max $\mathbb{E}U_1$ vs. the bounds provided by Propositions 2 and 4.

## 3. Proofs

**Proof of Proposition 1.** Change the measure $\mathbf{p}$ by splitting some atom $i$ into two atoms of sizes $p_i'$ and $p_i''$ where $0 < p_i' < p_i$ and $p_i'' = p_i - p_i'$. Let $\widetilde{\mathbf{p}} = (p_1, ..., p_{i-1}, p_i', p_i'', p_{i+1}, ...)$ be the new measure. The entropy of $\mathbf{p}$ is smaller than that of $\widetilde{\mathbf{p}}$ since

$$\begin{aligned} p_i \ln \frac{1}{p_i} &= [p_i' + p_i''] \ln \frac{1}{p_i} \\ &< p_i' \ln \frac{1}{p_i'} + p_i'' \ln \frac{1}{p_i''}. \end{aligned}$$

Now

$$\begin{aligned} p_i(1-p_i)^t &= [p_i' + p_i''](1-p_i)^t \\ &< p_i'(1-p_i')^t + p_i''(1-p_i'')^t, \end{aligned}$$

which implies that $\mathbb{E}_{\mathbf{p}} U_t < \mathbb{E}_{\widetilde{\mathbf{p}}} U_t$. Similarly, splitting any number of atoms, we increase both the entropy and $\mathbb{E}U_t$.

Now, take the first atom, for example, and split it into $k$ sub-atoms, the first $k-1$ of which are of size $p$ each and the $k$-th of size $p'$, where $0 \le p \le \frac{p_1}{k}$ and $p' = p_1 - (k-1)p$, and $k$ is still to be determined. The entropy of the new measure is

$$(k-1)p \ln \frac{1}{p} + (p_1 - (k-1)p) \ln \frac{1}{p_1 - (k-1)p} + \sum_{i=2}^{\infty} p_i \ln \frac{1}{p_i}.$$

For sufficiently large $k$ and $p = \frac{p_1}{k}$, this entropy becomes arbitrarily large, and in particular exceeds $h$. Take such a $k$, and consider the entropy of the obtained measure as $p$ grows continuously from 0 to $\frac{p_1}{k}$. For $p = 0$, we have basically the original measure (and thus an entropy less than $h$), while for $p = \frac{p_1}{k}$ the entropy is larger than $h$. Hence for an appropriate intermediate value of $p$, the entropy is exactly $h$. The measure obtained for this $p$ proves our claim. □

**Proof of Theorem 1.** Write down the Lagrangian:

$$L(\mathbf{p}, \lambda_1, \lambda_2) = \sum_{i=1}^{n} p_i(1-p_i)^t + \lambda_1 \left( \sum_{i=1}^{n} p_i - 1 \right) + \lambda_2 \left( \sum_{i=1}^{n} p_i \ln \frac{1}{p_i} - h \right).$$

The first-order conditions yield:

$$\frac{\partial L}{\partial p_i} = (1-p_i)^t - t p_i (1-p_i)^{t-1} + \lambda_1 - \lambda_2(\ln p_i + 1) = 0.$$

Denote:

$$f(x) = (1-x)^t - tx(1-x)^{t-1} + \lambda_1 - \lambda_2(\ln x + 1), \qquad 0 < x < 1.$$

We have:

$$f'(x) = t(1-x)^{t-2}[x(t+1) - 2] - \frac{\lambda_2}{x}.$$

We claim that $f'$ vanishes at most three times in $(0,1)$. Indeed, $f'(x) = 0$ when

$$xt(1-x)^{t-2}[x(t+1) - 2] = \lambda_2. \tag{8}$$

Denote the left-hand side of (8) by $g(x)$. Then:

$$g'(x) = -t(1-x)^{t-3}[t^2x^2 + t(x-4)x + 2].$$

For every two points $x_1$,$x_2$ for which (8) holds, there is an intermediate point $x_1 < \xi < x_2$ such that $g'(\xi) = 0$. Now, $g'$ clearly vanishes at no more than two points, so that (8) holds for at most three values of $x$. It follows that each $p_i$ assumes one of up to (the same) four values. □

Before we prove Theorem 2 we need two auxiliary lemmas. For $h \geq 0$, let $X_h$ be the subset of $\ell_1$ consisting of all non-increasing sequences $(p_1, p_2, \ldots)$ satisfying the following properties:

1. $p_i \geq 0$ for each $i$ and $\sum_{i=1}^{\infty} p_i = 1$.
2. $H(\mathbf{p}) \leq h$.

**Lemma 1.** *$X_h$ is compact under the $\ell_1$ metric.*

**Proof of Lemma 1.** Let $(\mathbf{p}_n)_{n=1}^{\infty}$ be a sequence in $X_h$, say $\mathbf{p}_n = (p_{n1}, p_{n2}, \ldots)$ for $n \geq 1$. We want to show that it has a convergent subsequence in $X_h$. Employing the diagonal method, we may assume that $\mathbf{p}_n$ converges component-wise. Let $\mathbf{p} = (p_1, p_2, \ldots)$ be the limit. It is clear that $\mathbf{p}$ has non-negative and non-increasing entries, so we only need to show that $\sum_{i=1}^{\infty} p_i = 1$, that $H(\mathbf{p}) \leq h$, and that $\mathbf{p}_n \underset{n\to\infty}{\longrightarrow} \mathbf{p}$ in $\ell_1$.

Assume first that $\sum_{i=1}^{\infty} p_i > 1$. Then, there exists an index $i_0$ such that $\sum_{i=1}^{i_0} p_i > 1$. Hence for sufficiently large $n$, we have $\sum_{i=1}^{i_0} p_{ni} > 1$, which is a contradiction. Hence $\sum_{i=1}^{\infty} p_i \leq 1$. Now assume that $\sum_{i=1}^{\infty} p_i < 1$. Put $\varepsilon = 1 - \sum_{i=1}^{\infty} p_i$. Let $i_0$ be an integer, to be determined later. We have $\sum_{i=1}^{i_0} p_{ni} < 1 - \frac{\varepsilon}{2}$ for all sufficiently large $n$. Note that for every $\mathbf{q} = (q_1, q_2, \ldots) \in X_h$ we have $q_i \leq \frac{1}{i}(q_1 + q_2 + \ldots + q_i) \leq \frac{1}{i}$. Now we can bound from below the tail entropy of $\mathbf{p}_n$:

$$\sum_{i=i_0+1}^{\infty} p_{ni} \ln \frac{1}{p_{ni}} > \sum_{i=i_0+1}^{\infty} p_{ni} \ln i_0 > \frac{\varepsilon}{2} \ln i_0.$$

Taking $i_0$ large enough, we can make the right-hand side larger than $h$, which is impossible. Hence $\sum_{i=1}^{\infty} p_i = 1$.

We now show similarly that $H(\mathbf{p}) \leq h$. Assume that $\sum_{i=1}^{\infty} p_i \ln \frac{1}{p_i} > h$. Then there exists an $i_0$ such that $\sum_{i=1}^{i_0} p_i \ln \frac{1}{p_i} > h$. Then, however, $\sum_{i=1}^{i_0} p_{ni} \ln \frac{1}{p_{ni}} > h$ for sufficiently large $n$, which yields a contradiction.

To prove convergence in $\ell_1$, we estimate $\|\mathbf{p}_n - \mathbf{p}\|_1 = \sum_{i=1}^{\infty} |p_{ni} - p_i|$. Let $\varepsilon > 0$. Since $\sum_{i=1}^{\infty} p_i = 1$, we can find an $i_0$ such that $\sum_{i=i_0+1}^{\infty} p_i < \frac{\varepsilon}{6}$. Due to the component-wise convergence, for sufficiently large $n$ we have $\sum_{i=1}^{i_0} |p_{ni} - p_i| < \frac{\varepsilon}{6}$. For such $n$ we also have $\sum_{i=i_0+1}^{\infty} p_{ni} < \frac{\varepsilon}{3}$ since

$$\begin{aligned}\sum_{i=1}^{i_0}|p_{ni}-p_i|<\frac{\varepsilon}{6} &\Rightarrow \left|\sum_{i=1}^{i_0}(p_{ni}-p_i)\right|<\frac{\varepsilon}{6}\\ &\Rightarrow \sum_{i=1}^{i_0}p_{ni}>\sum_{i=1}^{i_0}p_i-\frac{\varepsilon}{6}>1-\frac{\varepsilon}{3}\\ &\Rightarrow \sum_{i=i_0+1}^{\infty}p_{ni}<\frac{\varepsilon}{3}.\end{aligned}$$

Thus we have

$$\begin{aligned}\sum_{i=1}^{\infty}|p_{ni}-p_i| &= \sum_{i=1}^{i_0}|p_{ni}-p_i|+\sum_{i=i_0+1}^{\infty}|p_{ni}-p_i|\\ &< \frac{\varepsilon}{6}+\sum_{i=i_0+1}^{\infty}|p_{ni}|+\sum_{i=i_0+1}^{\infty}|p_i|\\ &< \frac{\varepsilon}{6}+\frac{\varepsilon}{3}+\frac{\varepsilon}{6}<\varepsilon.\end{aligned}$$

Hence we have convergence in $\ell_1$. This proves the lemma. □

**Example 1.** *Note that the subset of $X_h$ consisting of all those vectors whose entropy is exactly $h$ is not compact. Let us demonstrate this fact, say, for $h=\ln 2$. We choose $\mathbf{p}_n=(x_n,\underbrace{\frac{1-x_n}{n},\frac{1-x_n}{n},\frac{1-x_n}{n},\ldots,\frac{1-x_n}{n}}_{n\text{ times}})$, $n\geq 3$, where $x_n$ will be defined momentarily. For arbitrary fixed $n\geq 3$, put:*

$$f_n(x)=-x\ln x-(1-x)\ln\frac{1-x}{n},\ \ 0\leq x\leq 1.$$

*We claim that there exists a unique solution $x_n$ to the equation $f_n(x)=\ln 2$. Indeed, this follows readily from the fact that $f_n(x)$ is concave and $f_n(1)=0<\ln 2<\ln n=f_n(0)$. Denoting $t_n=1-x_n$, we have*

$$-(1-t_n)\ln(1-t_n)-t_n\ln t_n+t_n\ln n=\ln 2.$$

*Hence*

$$t_n = \frac{1}{\ln n}\left(\ln 2+t_n\ln t_n+(1-t_n)\ln(1-t_n)\right)\leq\frac{\ln 2}{\ln n},$$

*so that*

$$x_n\geq 1-\frac{\ln 2}{\ln n},\qquad n\geq 3,$$

*and in particular $x_n\underset{n\to\infty}{\longrightarrow}1$. Thus, $\mathbf{p}_n\underset{n\to\infty}{\longrightarrow}(1,0,0,0,\ldots)$ while $H(\mathbf{p}_n)=\ln 2$ and $H((1,0,0,0,\ldots))=0$, which completes the example.*

For arbitrary fixed $t$, the quantity $\mathbb{E}U_t$ assigns to each point in $X_h$ a real number. We will denote this function by $\mathbb{E}U_t$.

**Lemma 2.** *The mapping $\mathbb{E}U_t$: $X_h\longrightarrow\mathbb{R}$ is Lipschitz with constant 1 with respect to the $\ell_1$ metric.*

**Proof of Lemma 2.** Consider the function $f:[0,1]\longrightarrow\mathbb{R}$ given by $f(x)=x(1-x)^t$, $0\leq x\leq 1$. Let $M$ be the Lipschitz constant for $f(x)$. According to Lemma 7 from [4], the candidates for assuming

the maximum of $|f'(x)|$ are the points $x_1 = 0$ and $x_2 = \frac{2}{1+t}$. Now $|f'(x_1)| = 1$ and $|f'(x_2)| = (1 - \frac{2}{1+t})^{t-1} \leq 1$. Hence the Lipschitz constant for $f(x)$ is 1. It follows that if $\mathbf{p}, \mathbf{p}' \in X_h$, then:

$$\begin{aligned} \left| \sum_{i=1}^{\infty} p_i(1-p_i)^t - \sum_{i=1}^{\infty} p_i'(1-p_i')^t \right| &= \left| \sum_{i=1}^{\infty} \left( p_i(1-p_i)^t - p_i'(1-p_i')^t \right) \right| \\ &\leq \sum_{i=1}^{\infty} |p_i - p_i'| . \end{aligned}$$

□

**Proof of Theorem 2.**

(i) Follows from Lemma 1 and Lemma 2.

(ii) Suppose that $\mathbf{p} = (p_1, p_2, \ldots)$ does not have a finite support. Then, we can find an $n_0$ such that the first $n_0$ entries $p_1, p_2, \ldots, p_{n_0}$ of $\mathbf{p}$ assume more than four different values. Put $\widetilde{\mathbf{p}} = (p_1, p_2, \ldots, p_{n_0})$ and let $c = p_1 + p_2 + \ldots + p_{n_0}$ and $\widetilde{h} = H(\widetilde{\mathbf{p}})$. Consider the optimization problem

$$\max_{\mathbf{p}} \sum_{i=1}^{n_0} p_i(1-p_i)^t \tag{9}$$

subject to

$$\sum_{i=1}^{n_0} p_i \ln \frac{1}{p_i} \leq \widetilde{h}, \tag{10}$$

$$\sum_{i=1}^{n_0} p_i = c. \tag{11}$$

Theorem 1 is still applicable to (9)–(11) with a minor variation. In the beginning of the proof, replace the Lagrangian by

$$L(\mathbf{p}, \lambda_1, \lambda_2) = \sum_{i=1}^{n} p_i(1-p_i)^t + \lambda_1 \left( \sum_{i=1}^{n} p_i - c \right) + \lambda_2 \left( \sum_{i=1}^{n} p_i \ln \frac{1}{p_i} - \widetilde{h} \right)$$

and proceed as previously. Since $\mathbf{p} \in X_h$ maximizes $\mathbb{E}U_t$, the vector $\widetilde{\mathbf{p}}$ is a global optimum of this finite-dimensional problem. By Theorem 1, $\widetilde{\mathbf{p}}$ cannot assume more than four distinct values.

□

**Proof of Theorem 3.** For $0 < x < 1$ :

$$\begin{aligned} \ln(1-x) &= -x - \frac{x^2}{2} - \ldots - \frac{x^{2t-1}}{2t-1} - \ldots \\ &< -x - \frac{x^2}{2} - \ldots - \frac{x^{2t-1}}{2t-1} \\ &= \left( -x - \frac{x^{2t-1}}{2t-1} \right) + \ldots + \left( -\frac{x^{t-k}}{t-k} - \frac{x^{t+k}}{t+k} \right) \\ &\quad + \ldots + \left( -\frac{x^{t-1}}{t-1} - \frac{x^{t+1}}{t+1} \right) - \frac{x^t}{t}. \end{aligned} \tag{12}$$

For each term on the right-hand side of (12) and for $1 \leq k \leq t-1$, we have

$$-\frac{x^{t-k}}{t-k} - \frac{x^{t+k}}{t+k} < -\frac{x^t}{t-k} - \frac{x^t}{t+k}.$$

Indeed,

$$-\frac{x^{t-k}}{t-k}-\frac{x^{t+k}}{t+k}+\frac{x^t}{t-k}+\frac{x^t}{t+k}=-x^{t-k}(1-x^k)(\frac{1}{t-k}-\frac{x^k}{t+k})<0.$$

This gives us

$$-x-\frac{x^2}{2}-\ldots-\frac{x^{2t-1}}{2t-1}-\ldots<-x^t-\frac{x^t}{2}-\ldots-\frac{x^t}{t}-\ldots-\frac{x^t}{2t-1}<-x^tH_{2t-1}.$$

Hence

$$\ln p<-(1-p)^tH_{2t-1},$$

and therefore

$$(1-p)^t<-\frac{\ln p}{H_{2t-1}}.$$

Finally,

$$\mathbb{E}U_t=\sum_{i=1}^{\infty}p_i(1-p_i)^t\leq\frac{h}{H_{2t-1}}.$$

□

**Proof of Theorem 4.** Let $t>1$ be an integer and $\alpha>1$. Define $\mathbf{p}$ by:

$$p_1=p_2=\ldots=p_t=\frac{h}{\sqrt{\alpha}t\ln t},\quad p_{t+1}=1-\frac{h}{\sqrt{\alpha}\ln t}.$$

For such $t$:

$$\begin{aligned}\mathbb{E}_{\mathbf{p}}U_t &= \left(1-\frac{h}{\sqrt{\alpha}\ln t}\right)\left(\frac{h}{\sqrt{\alpha}\ln t}\right)^t+\frac{h}{\sqrt{\alpha}\ln t}\left(1-\frac{h}{\sqrt{\alpha}t\ln t}\right)^t\\ &\geq \frac{h}{\sqrt{\alpha}\ln t}\left(1-\frac{h}{\sqrt{\alpha}t\ln t}\right)^t.\end{aligned}$$

Now:

$$\left(1-\frac{h}{\sqrt{\alpha}t\ln t}\right)^t=\left(\left(1-\frac{h}{\sqrt{\alpha}t\ln t}\right)^{t\ln t}\right)^{1/\ln t}\underset{t\to\infty}{\longrightarrow}(e^{-h/\sqrt{\alpha}})^0=1.$$

□

**Proof of Proposition 2.** Let $P$ be the random variable assigning to each atom $i\in S$ its probability:

$$P(i)=p_i,\qquad i\in S.$$

Denote $I=-\ln P$. Then:

$$\mathbb{E}U_1=\sum_{i\in S}p_i(1-p_i)=\mathbb{E}[1-P]=\mathbb{E}[1-e^{-I}].$$

The function $f(x)=1-e^{-x}$ is concave, so by Jensen's inequality:

$$\begin{aligned}\mathbb{E}U_1 &= \mathbb{E}[1-e^{-I}]\\ &\leq 1-e^{-\mathbb{E}[I]}\\ &= 1-e^{-\mathbb{E}[-\ln P]}\\ &= 1-e^{\sum_{i\in S} p_i \ln p_i}\\ &= 1-e^{-h}.\end{aligned}$$

□

**Remark 2.** *If $h = \ln k$ for some positive integer $k$, then the bound is attained for the uniform distribution on a space of $k$ points.*

**Proof of Proposition 3.** First, in the case where $h = \ln k$ we have a unique optimal solution, which is $p_1^* = p_2^* = \ldots = p_{n-k}^* = 0 < p_{n-k+1}^* = p_{n-k+2}^* = \ldots = p_n^* = \frac{1}{k}$. It is straightforward to check that $\mathbf{p}^*$ is feasible and attains the upper bound in Proposition 2, and is optimal. Moreover, $\mathbf{p}^*$ is unique because any feasible non-uniform choice of $\mathbf{p}^*$ leads to a strict inequality in Jensen's inequality that was used in Proposition 2.

Thus, we deal with the case of strict inequalities, $\ln(k-1) < h < \ln k$. We start by showing that any optimal solution $(p_1^*, p_2^*, \ldots, p_n^*)$ assumes at most two non-zero distinct values. Write down the Lagrangian:

$$L(p_1, p_2, \ldots, p_n, \lambda_1, \lambda_2) = \sum_{i=1}^{n} p_i(1-p_i) + \lambda_1\left(\sum_{i=1}^{n} p_i - 1\right) + \lambda_2\left(\sum_{i=1}^{n} p_i \ln\frac{1}{p_i} - h\right).$$

The first-order conditions yield, at any optimal point,

$$\frac{\partial L}{\partial p_i} = 1 - 2p_i + \lambda_1 - \lambda_2(\ln p_i + 1) = 0,$$

for every $i$ with $p_i^* > 0$. Define the function $f$ by $f(x) = 1 - 2x + \lambda_1 - \lambda_2 - \lambda_2 \ln x$. The function vanishes at most twice in $(0,1)$ because its derivative $f'(x) = -2 - \frac{\lambda_2}{x}$ vanishes at most once. Thus, the non-zero $p_i^*$s assume at most two distinct values. In fact, if all were equal, we would have $k = \ln k$, where $k$ is the number of non-zero $p_i^*$s, so that we would have exactly two distinct values for the $p_i^*$s. Disposing of the points of mass 0, we may assume that all $n$ points of $S$ have positive mass. Denote the number of "light" atoms by $\ell$. We will show that $\mathbb{E}U_1$ decreases as we increase $\ell$. Denote the mass of a "light" atom by $p$ and write down the entropy constraint with $\ell$ "light" atoms and $n-\ell$ "heavy" ones:

$$-\ell p \ln p - (1-\ell p)\ln\left(\frac{1-\ell p}{n-\ell}\right) = h.$$

Now, define the function $F(\ell, p)$ by:

$$F(\ell, p) = -\ell p \ln p - (1-\ell p)\ln\left(\frac{1-\ell p}{n-\ell}\right) - h.$$

Note that we treat $\ell$ as a continuous variable. The equation $F(\ell, p) = 0$ implicitly defines the function $p(\ell)$. Using the implicit function theorem, we can write an analytic expression for $\frac{dp}{d\ell}$:

$$\frac{dp}{d\ell} = -\frac{p(\ell)}{\ell} + \frac{\frac{1-\ell p(\ell)}{n-\ell} - p(\ell)}{\ell\left[\ln\left(\frac{1-\ell p(\ell)}{n-\ell}\right) - \ln p(\ell)\right]}.$$

Now write down $\mathbb{E}U_1$ as a function of $\ell$ and take the derivative with respect to $\ell$:

$$\mathbb{E}U_1 = \ell p(\ell)(1-p(\ell)) + (1-\ell p(\ell))\left(1 - \frac{1-\ell p(\ell)}{n-\ell}\right).$$

$$\begin{aligned}\frac{d\mathbb{E}U_1}{d\ell} &= \left(p(\ell)+\ell p'\right)(\ell)\left(1-p(\ell)\right) - \ell p(\ell)p'(\ell) \\ &\quad - \left(p(\ell)+\ell p'(\ell)\right)\left(1 - \frac{1-\ell p(\ell)}{n-\ell}\right) \\ &\quad - \frac{(1-\ell p(\ell))}{(n-\ell)^2}\left(-np(\ell) - n\ell p'(\ell) + \ell^2 p'(\ell) + 1\right). \end{aligned} \tag{13}$$

Notice that the term $\frac{1-\ell p(\ell)}{n-\ell}$ is actually the mass of the "heavy" atom, so to simplify notation we put $q(\ell) = \frac{1-\ell p(\ell)}{n-\ell}$. Substituting the expression for $\frac{dp}{d\ell}$, we obtain:

$$\begin{aligned}\frac{\partial \mathbb{E}U_1}{\partial \ell} &= (q(\ell)-p(\ell))^2\left[\frac{2}{\ln\frac{q(\ell)}{p(\ell)}} - \frac{2p(\ell)}{q(\ell)-p(\ell)} - 1\right] \\ &= (q(\ell)-p(\ell))^2\left[\frac{2}{\ln\frac{q(\ell)}{p(\ell)}} - \frac{2}{\frac{q(\ell)}{p(\ell)}-1} - 1\right]. \end{aligned} \tag{14}$$

To show that $\mathbb{E}U_1$ decreases as we increase $\ell$, it is enough to check that $\frac{\partial \mathbb{E}U_1}{\partial \ell} < 0$. It suffices to work out the second term in the product of (14). Using the change of variables $y = \frac{q(\ell)}{p(\ell)} - 1$, we may write:

$$\frac{2}{\ln\frac{q(\ell)}{p(\ell)}} - \frac{2}{\frac{q(\ell)}{p(\ell)}-1} - 1 = \frac{2}{\ln(1+y)} - \frac{2}{y} - 1.$$

It is straightforward to check that $\frac{2}{\ln(1+y)} - \frac{2}{y} - 1 \leq 0$:

$$\frac{2}{\ln(1+y)} - \frac{2}{y} - 1 = \frac{2y - 2\ln(1+y) - y\ln(1+y)}{y\ln(1+y)}.$$

Notice that $y\ln(1+y) > 0$, and hence it is enough to check that the numerator is negative. Indeed,

$$\left[2y - 2\ln(1+y) - y\ln(1+y)\right]_{y=0} = 0$$

and

$$\frac{d}{dy}\left[2y - 2\ln(1+y) - y\ln(1+y)\right] = \frac{y}{1+y} - \ln(1+y).$$

Now $\left[\frac{y}{1+y} - \ln(1+y)\right]_{y=0} = 0$ and $\frac{d}{dy}\left[\frac{y}{1+y} - \ln(1+y)\right] = \frac{1}{(1+y)^2} - \frac{1}{1+y} < 0$. Thus, $\frac{\partial \mathbb{E}U_1}{\partial \ell} < 0$.

It follows that $\ell$ should be as small as possible, which means (since there is at least one light atom) that $\ell = 1$. Finally, as there is one light atom and $n-1$ heavy ones, the entropy $h$ lies in the interval $(\ln(n-1), \ln n)$. Reverting to the original notations, we have $\ln(k-1) < h < \ln k$. □

**Proof of Proposition 4.** We use the following refinement of Jensen's inequality [11]: For any random variable $X$ and concave function $\phi$,

$$\phi(\mathbb{E}(X)) - \mathbb{E}(\phi(X)) \geq \left|\mathbb{E}\left(\left|\phi(X) - \phi(\mathbb{E}(X))\right|\right) - \left|\phi'_+(\mathbb{E}(X))\right| \cdot \mathbb{E}\left(\left|X - \mathbb{E}(X)\right|\right)\right|, \tag{15}$$

where $\phi'_+$ denotes the right-hand derivative of $\phi$. For $I = -\ln P$ and $\phi(x) = 1 - e^{-x}$, the left-hand side of (15) is

$$\phi(\mathbb{E}[I]) - \mathbb{E}[\phi(I)] = 1 - e^{-h} - \mathbb{E}U_1.$$

The right-hand side of (15) gives:

$$\begin{aligned}
&\left|\mathbb{E}\left(\left|\phi(I) - \phi(\mathbb{E}[I])\right|\right) - \left|\phi'_+(\mathbb{E}[I])\right| \cdot \mathbb{E}\left(\left|I - E[I]\right|\right)\right| \\
&= \left|\mathbb{E}\left(\left|1 - e^{-I} - \left(1 - e^{-h}\right)\right|\right) - e^{-h}\mathbb{E}\left(\left|-\ln P - h\right|\right)\right| \\
&= \Big|p \cdot \left|e^{-h} - p\right| + (n-1)\, q \cdot \left|e^{-h} - q\right| \\
&\quad -e^{-h}\left(p \cdot \left|-\ln p - h\right| + (n-1)\, q \cdot \left|-\ln q - h\right|\right)\Big| \\
&= \Big|p\left(e^{-h} - p\right) - (n-1)\, q\left(e^{-h} - q\right) \\
&\quad -e^{-h}\Big(p\left(-\ln p - h\right) - (n-1)\, q\left(-\ln q - h\right)\Big)\Big| \\
&= \left|pe^{-h} - p^2 - (n-1)\, qe^{-h} + (n-1)\, q^2 - e^{-h}\Big(-2p\ln p - h - 2ph + h\Big)\right| \\
&= \left|e^{-h}\Big(2p - 1 + 2p\ln p + 2ph\Big) + (n-1)\, q^2 - p^2\right|.
\end{aligned}$$

□

**Acknowledgments:** Daniel Berend is supported by the Milken Families Foundation Chair in Mathematics. Aryeh Kontorovich is supported in part by the Israel Science Foundation (grant No. 755/15), Paypal and IBM.

**Author Contributions:** The authors contributed equally to this work.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

1. Good, I.J. The Population Frequencies of Species and the Estimation of Population Parameters. *Biometrika* **1953**, *40*, 237–264.
2. McAllester, D.; McAllester, R.E. On The Convergence Rate of Good-Turing Estimators. In Proceedings of the Thirteenth Annual Conference on Computational Learning Theory, Stanford, CA, USA, 28 June–1 July 2000; pp. 1–6.
3. Haslinger, R.; Pipa, G.; Lewis, L.D.; Nikolić, D.; Williams, Z.; Brown, E. Encoding through Patterns: Regression Tree-Based Neuronal Population Models. *Neural Comput.* **2013**, *25*, 1953–1993.
4. Berend, D.; Kontorovich, A. The Missing Mass Problem. *Stat. Probab. Lett.* **2012**, *82*, 1102–1110.
5. Berend, D.; Kontorovich, A. On The Concentration of the Missing Mass. *Electron. Commun. Probab.* **2013**, *18*, 1–7.
6. Kontorovich, A.; Hendler, D.; Menahem, E. Metric Anomaly Detection via Asymmetric Risk Minimization. In *SIMBAD 2011. Lecture Notes in Computer Science*; Springer: Berlin/Heidelberg, Germany, 2011; Volume 7005, pp. 17–30.
7. Luo, H.P.; Schapire, R. Towards Minimax Online Learning with Unknown Time Horizon. In Proceedings of the 31st International Conference on International Conference on Machine Learning, Beijing, China, 21–26 June 2014.

8. Sadeqi, M.; Holte, R.C.; Zilles, S. Detecting Mutex Pairs in State Spaces by Sampling. In *Australasian Conference on Artificial Intelligence (2013)*; Lecture Notes in Computer Science; Springer: Cham, Switzerland; Volume 8272, pp. 490–501.
9. Ben-Hamou, A.; Boucheron, S.; Ohannessian, M.I. Concentration Inequalities in the Infinite Urn Scheme for Occupancy Counts and the Missing Mass, with Applications. *Bernoulli* **2017**, *23*, 249–287.
10. Han, Y.J.; Jiao, J.T.; Weissman, T. Minimax Estimation of Discrete Distributions under $\ell_1$ Loss. *IEEE Trans. Inf. Theory* **2015**, *61*, 6343–6354.
11. Hussain, S.; Pečarić, J. An Improvement of Jensen's Inequality with Some Applications. *Asian Eur. J. Math.* **2009**, *2*, 85–94.

© 2017 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).