Electron. Commun. Probab. **18** (2013), no. 3, 1–7.
DOI: 10.1214/ECP.v18-2359
ISSN: 1083-589X

ELECTRONIC COMMUNICATIONS in PROBABILITY

On the concentration of the missing mass*

Daniel Berend† Aryeh Kontorovich ‡

Abstract

A random variable is sampled from a discrete distribution. The missing mass is the probability of the set of points not observed in the sample. We sharpen and simplify McAllester and Ortiz's results (JMLR, 2003) bounding the probability of large deviations of the missing mass. Along the way, we refine and rigorously prove a fundamental inequality of Kearns and Saul (UAI, 1998).

Keywords: concentration; missing mass; Hoeffding inequality.
AMS MSC 2010: 60F10;39B72.
Submitted to ECP on October 9, 2012, final version accepted on January 6, 2013.

1 Introduction

Hoeffding's classic inequality [3] states that if X is a $[a,b]$-valued random variable with $\mathbb{E}X = 0$ then

$$\mathbb{E}e^{tX} \leq e^{(b-a)^2 t^2/8}, \qquad t \geq 0. \tag{1.1}$$

A standard proof of (1.1) proceeds by writing $x \in [a,b]$ as $x = pb + (1-p)a$, for $p = (x-a)/(b-a)$, and using convexity to obtain

$$\mathbb{E}e^{tX/(b-a)} \leq (1-p)e^{-tp} + pe^{t(1-p)} =: f(t) \leq e^{t^2/8}, \tag{1.2}$$

where the last inequality follows by noticing that $\log f(0) = [\log f(t)]'|_{t=0} = 0$ and that $[\log f(t)]'' \leq 1/4$.

Although (1.1) is tight, it is a "worst-case" bound over all distributions with the given support. Refinements of (1.1) include the Bernstein and Bennett inequalities [5], which take the variance into account — but these are also too crude for some purposes.

In 1998, Kearns and Saul [4] put forth an exquisitely delicate inequality for (generalized) Bernoulli random variables, which is sensitive to the underlying distribution:

$$(1-p)e^{-tp} + pe^{t(1-p)} \leq \exp\left(\frac{1-2p}{4\log((1-p)/p)}t^2\right), \qquad p \in [0,1],\ t \in \mathbb{R}. \tag{1.3}$$

One easily verifies that (1.3) is superior to (1.2) — except for $p = 1/2$, where the two coincide. In fact, (1.3) is optimal in the sense that, for every p, there is a t for which equality is achieved. The Kearns-Saul inequality

*Supported in part by the Israel Science Foundation (grant #1141/12).
†Ben-Gurion University, Israel. E-mail: berend@cs.bgu.ac.il
‡Ben-Gurion University, Israel. E-mail: karyeh@cs.bgu.ac.il

allows one to analyze various inference algorithms in neural networks, and the influential paper [4] has inspired a fruitful line of research [1, 7, 9, 10].

One specific application of the Kearns-Saul inequality involves the concentration of the missing mass. Let $\mathbf{p} = (p_1, p_2, \ldots)$ be a distribution over $\mathbb{N}$ and suppose that $X_1, X_2, \ldots, X_n$ are sampled iid according to $\mathbf{p}$. Define the indicator variable ξ_j to be 0 if j occurs in the sample and 1 otherwise:

$$\xi_j = \mathbb{1}_{\{j \notin \{X_1, \ldots, X_n\}\}}, \qquad j \in \mathbb{N}.$$

The *missing mass* is the random variable

$$U_n = \sum_{j \in \mathbb{N}} p_j \xi_j. \tag{1.4}$$

McAllester and Schapire [8] first established subgaussian concentration for the missing mass via a somewhat intricate argument. Later, McAllester and Ortiz [7] showed how the standard inequalities of Hoeffding, Angluin-Valiant, Bernstein and Bennett are inadequate for obtaining exponential bounds of the correct order in n, and developed a thermodynamic approach for systematically handling this problem[1].

We were led to the Kearns-Saul inequality (1.3) in an attempt to understand and simplify the missing mass concentration results of McAllester and Ortiz [7], some of which rely on (1.3). However, we were unable to complete the proof of (1.3) sketched in [4], and a literature search likewise came up empty. The proof we give here follows an alternate path, and may be of independent interest. As an application, we simplify and sharpen some of the missing mass concentration results given in [8, 7].

2 Main results

In [4, Lemma 1], Kearns and Saul define the function

$$g(t) = \frac{1}{t^2} \log \left[(1-p)e^{-tp} + pe^{t(1-p)} \right], \qquad t \in \mathbb{R}. \tag{2.1}$$

A natural attempt to find the maximum of g leads one to the transcendental equation

$$g'(t) = \frac{(e^t - 1)(1-p)pt - 2(1 + (e^t - 1)p) \log[1 + (e^t - 1)pe^{-pt}]}{(1 + (e^t - 1)p)t^3} = 0.$$

In an inspired tour de force, Kearns and Saul were able to find that $g'(t^*) = 0$ for

$$t^* = 2 \log \frac{1-p}{p}.$$

This observation naturally suggests (i) arguing that t^* is the unique zero of g' and (ii) supplying (perhaps via second-order information) an argument for t^* being a local maximum. In fact, all evidence points to $g'(t)$ having the following properties:

(*) $g' > 0$ on $(-\infty, t^*)$,
(**) $g' = 0$ at $t = t^*$,
(***) $g' < 0$ on (t^*, ∞).

Unfortunately, besides straightforwardly verifying (**), we were not able to formally establish (*) or (***) — and we leave this as an intriguing open problem. Instead, in Theorem 3.2 we prove the Kearns-Saul inequality (1.3) via a rather different approach. Moreover, for $p \geq 1/2$ and $t \geq 0$, the right-hand side of (1.3) may be improved to $\exp[p(1-p)t^2/2]$. This refinement, proved in Lemma 3.3, may be of independent interest.

As an application, we recover the upper tail estimate on the missing mass in [7, Theorem 16]:

[1] The latter has, in turn, inspired a general thermodynamic approach to concentration [6].

Theorem 2.1.

$$\mathbb{P}(U_n > \mathbb{E}U_n + \varepsilon) \le e^{-n\varepsilon^2}.$$

We also obtain the following lower tail estimate:

Theorem 2.2.

$$\mathbb{P}(U_n < \mathbb{E}U_n - \varepsilon) \le e^{-C_0 n\varepsilon^2/4},$$

where

$$C_0 = \inf_{0<x<1/2} \frac{2}{x(1-x)\log(1/x)} \approx 7.6821.$$

Since $C_0/4 \approx 1.92$, Theorem 2.2 sharpens the estimate in [7, Theorem 10], where the constant in the exponent was $e/2 \approx 1.36$. Our bounds are arguably simpler than those in [7] as they bypass the thermodynamic approach.

3 Proofs

The following well-known estimate is an immediate consequence of (1.2):

Lemma 3.1.

$$\frac{1}{2}e^{-t} + \frac{1}{2}e^{t} = \cosh t \le e^{t^2/2}, \qquad t \in \mathbb{R}.$$

We proceed with a proof of the Kearns-Saul inequality.[2]

Theorem 3.2. *For all $p \in [0,1]$ and $t \in \mathbb{R}$,*

$$(1-p)e^{-tp} + pe^{t(1-p)} \le \exp\left(\frac{1-2p}{4\log((1-p)/p)}t^2\right). \tag{3.1}$$

Proof. The cases $p = 0, 1$ are trivial. Since

$$\lim_{p\to 1/2} \frac{1-2p}{\log((1-p)/p)} = 1/2,$$

for $p = 1/2$ the claim follows from Lemma 3.1.

For $p \neq 1/2$, we multiply both sides of (3.1) by e^{tp}, take logarithms, and put $t = 2s\log((1-p)/p)$ to obtain the equivalent inequality

$$s\left(s + 2p(1-s)\right)\log((1-p)/p) - \log\left(1 - p + p((1-p)/p)^{2s}\right) \ge 0. \tag{3.2}$$

For $s \in \mathbb{R}$, denote the left-hand side of (3.2) by $h_s(p)$. A routine calculation yields

$$h_s(1/2) = h_s'(1/2) = 0 \tag{3.3}$$

and

$$h_s''(p) = \left(\frac{(\mu-1)p^2 - s + p(1-\mu+s+\mu s)}{p(1-p)(1+(\mu-1)p)}\right)^2,$$

where $\mu = ((1-p)/p)^{2s}$.

As $h_s'' \ge 0$, we have that h_s is convex, and from (3.3) it follows that $h_s(p) \ge 0$ for all s, p. □

[2] Rising to our challenge, Maxim Raginsky has found an elegant proof of Theorem 3.2 based on transportation and information-theoretic techniques [11, Theorem 37].

We will also need a refinement of (1.3):

Lemma 3.3. *For $p \in [1/2, 1]$ and $t \geq 0$,*

$$\frac{1}{t^2} \log \left[(1-p)e^{-tp} + pe^{t(1-p)} \right] \leq \frac{p(1-p)}{2}. \tag{3.4}$$

Remark: Since the right-hand side of (3.1) majorizes the right-hand side of (3.4) uniformly over $[1/2, 1]$, the latter estimate is tighter.

Proof. The claim is equivalent to

$$L(p) := (1-p) + pe^t \leq \exp(pt + p(1-p)t^2/2) =: R(p), \qquad t \geq 0.$$

For $t \geq 4$, we have

$$\begin{aligned} R(p) = \exp(pt + p(1-p)t^2/2) &\geq \exp(pt + p(1-p)(4t)/2) \\ &= \exp(pt + 2p(1-p)t) \\ &\geq \exp(pt + (1-p)t) = e^t \geq L(p). \end{aligned}$$

For $0 \leq t < 4$,

$$R''(p) - L''(p) = \frac{1}{4} \exp(pt(2+t-pt)/2)(2p-1)t^3((2p-1)t - 4), \qquad p \in [1/2, 1],$$

which is obviously non-positive. Now the inequality clearly holds at $p = 1$ (as equality), and the $p = 1/2$ case is implied by Lemma 3.1. The claim now follows by convexity.

□

Our numerical constants are defined in the following lemma, whose elementary proof is omitted:

Lemma 3.4. *Define the function*

$$f(x) = x(1-x)\log(1/x), \qquad x \in (0, 1/2).$$

Then $x_0 \approx 0.2356$ is the unique solution of $f'(x) = 0$ on $(0, 1/2)$. Furthermore,

$$C_0 := \inf_{0<x<1/2} 2/f(x) = 2/f(x_0) \approx 7.6821. \tag{3.5}$$

The main technical step towards obtaining our missing mass deviation estimates is the following lemma.

Lemma 3.5. *Let $n \geq 1$, $\lambda \geq 0$, $p \in [0, 1]$, and put $q = (1-p)^n$. Then:*

(a)

$$qe^{\lambda(p-pq)} + (1-q)e^{-\lambda pq} \leq \exp(p\lambda^2/4n),$$

(b)

$$qe^{\lambda(pq-p)} + (1-q)e^{\lambda pq} \leq \exp(p\lambda^2/C_0 n).$$

Proof. (a) We invoke Theorem 3.2 with $p = q$ and $t = \lambda p$ to obtain

$$qe^{\lambda(p-pq)} + (1-q)e^{-\lambda pq} \leq \exp[(1-2q)\lambda^2 p^2/4 \log[(1-q)/q]].$$

Thus it suffices to show that

$$(1-2q)\lambda^2 p^2/4 \log[(1-q)/q] \leq p\lambda^2/4n,$$

or equivalently,

$$(1-2q)p/\log[(1-q)/q] \le \log(1-p)/\log q, \qquad p,q \in [0,1].$$

Collecting the p and q terms on opposite sides, it remains to prove that

$$L(q) := \frac{(1-2q)\log(1/q)}{\log[(1-q)/q]} \le \frac{\log(1/(1-p))}{p} =: R(p), \qquad 0 < p,q < 1.$$

We claim that $L \le 1 \le R$. The second inequality is obvious from the Taylor expansion, since

$$\frac{\log(1/(1-p))}{p} = 1 + p/2 + p^2/3 + p^3/4 + \dots. \tag{3.6}$$

To prove that $L \le 1$, we note first that $L(q) \ge L(1-q)$ for $q \in (0,1/2)$. Hence, it suffices to consider $q \in (0,1/2)$. To this end, it suffices to show that the function

$$f(q) = \log[(1-q)/q] - (1-2q)\log(1/q)$$

is positive on $(0,1/2)$. Since $\lim f(q) \underset{q\to 0}{\longrightarrow} 0 = f(1/2)$ and

$$f''(q) = \frac{-2+3q-2q^2}{(1-q)^2 q} \le 0,$$

it follows that $f \ge 0$ on $[0,1/2]$.

(b) The inequality is equivalent to

$$L(\lambda) := \frac{1}{\lambda^2 p^2}\log\left[qe^{-\lambda p(1-q)} + (1-q)e^{\lambda pq}\right] \le \frac{1}{\lambda^2 p^2}\frac{\lambda^2 p}{C_0 \log q/\log(1-p)} =: R,$$

where L is obtained from the left-hand side of (3.1) after replacing p by $1-q$ and t by λp. We analyze the cases $q < 1/2$ and $q > 1/2$ separately (as above, the case where $q = 1/2$ is trivial). For $q > 1/2$, put $\lambda^* = \frac{2}{p}\log\frac{q}{1-q} > 0$ and invoke Theorem 3.2 to conclude that $\sup_{\lambda\ge 0} L(\lambda) \le L(\lambda^*)$. Hence, it remains to prove that $L(\lambda^*) \le R$, or equivalently,

$$\frac{(2q-1)}{4\log(q/(1-q))} \le \frac{1}{\lambda^2 p^2}\frac{\lambda^2 p}{C_0 \log q/\log(1-p)}.$$

After simplifying, this amounts to showing that

$$4\frac{\log(1/(1-p))}{p}\frac{\log(q/(1-q))}{(2q-1)\log(1/q)} \ge C_0.$$

As in (3.6), the factor $\log[1/(1-p)]/p$ is bounded below by 1. We claim that the factor $\frac{\log(q/(1-q))}{(2q-1)\log(1/q)}$, increases for $q \in [1/2,1]$. Indeed, this is obvious for $1/\log(1/q)$, and the expansion about $q = 1/2$

$$\frac{\log(q/(1-q))}{(2q-1)} = \sum_{n=0}^{\infty}\frac{2^{2n+1}}{2n+1}\left(q-\frac{1}{2}\right)^{2n}$$

shows that the same holds for $\frac{\log(q/(1-q))}{2q-1}$. In particular,

$$\begin{aligned} 4\frac{\log(1/(1-p))}{p}\frac{\log(q/(1-q)}{(2q-1)\log(1/q)} &\ge 4\cdot 1\cdot \lim_{q\to 1/2}\frac{\log(q/(1-q)}{(2q-1)\log(1/q)} \\ &= 8/\log 2 \approx 11.542 > C_0. \end{aligned}$$

When $q < 1/2$, we invoke Lemma 3.3 together with the observation that

$$\lim_{\lambda \to 0_+} L(\lambda) = \frac{q(1-q)}{2}$$

to conclude that $\sup_{\lambda \geq 0} L(\lambda) \leq L(0)$. Hence, it remains to show that

$$\frac{\log(1/(1-p))}{p} \frac{2}{q(1-q)\log(1/q)} \geq C_0.$$

As in (3.6), $\log[1/(1-p)]/p \geq 1$ and the claim follows by Lemma 3.4.

□

Our proof of Theorems 2.1 and 2.2 is facilitated by the following observation, also made in [7]. Although the random variables ξ_j whose weighted sum comprises the missing mass (1.4) are not independent, they are *negatively associated* [2]. A basic fact about negative association is that it is "at least as good as independence" as far as concentration about the mean is concerned [7, Lemmas 5-8]:

Lemma 3.6. *Let ξ'_j be independent random variables, where ξ'_j is distributed identically to ξ_j for all $j \in \mathbb{N}$. Define also the "independent analogue" of U_n:*

$$U'_n = \sum_{j \in \mathbb{N}} p_j \xi'_j.$$

Then for all $n \in \mathbb{N}$ and $\varepsilon > 0$,

(a)

$$\mathbb{P}(U_n \geq \mathbb{E}U_n + \varepsilon) \quad \leq \quad \mathbb{P}(U'_n \geq \mathbb{E}U'_n + \varepsilon),$$

(b)

$$\mathbb{P}(U_n \leq \mathbb{E}U_n - \varepsilon) \quad \leq \quad \mathbb{P}(U'_n \leq \mathbb{E}U'_n - \varepsilon).$$

Proof of Theorems 2.1 and 2.2. Observe that the random variables ξ'_j defined in Lemma 3.6 have a Bernoulli distribution with $\mathbb{P}(\xi'_j = 1) = q_j = (1-p_j)^n$ and put $X_j = \xi'_j - \mathbb{E}\xi'_j$. Using standard exponential bounding with Markov's inequality,

$$\begin{aligned}
\mathbb{P}(U_n \geq \mathbb{E}U_n + \varepsilon) &\leq \mathbb{P}(U'_n \geq \mathbb{E}U'_n + \varepsilon) \\
&= \mathbb{P}\left[\exp\left(\lambda \sum_{j \in \mathbb{N}} X_j\right) \geq e^{\lambda\varepsilon}\right], \qquad \lambda \geq 0 \\
&\leq e^{-\lambda\varepsilon} \prod_{j \in \mathbb{N}} \mathbb{E}e^{\lambda X_j} \\
&= e^{-\lambda\varepsilon} \prod_{j \in \mathbb{N}} \left(q_j e^{\lambda(p_j - p_j q_j)} + (1-q_j)e^{-\lambda p_j q_j}\right) \\
&\leq e^{-\lambda\varepsilon} \prod_{j \in \mathbb{N}} \exp(p_j \lambda^2 / 4n) \\
&= \exp(\lambda^2/4n - \lambda\varepsilon),
\end{aligned}$$

where the last inequality invoked Lemma 3.5(a). Choosing $\lambda = 2n\varepsilon$ yields Theorem 2.1.

The proof of the Theorem 2.2 is almost identical, except that X_j is replaced by $-X_j$ and Lemma 3.5(b) is invoked instead of Lemma 3.5(a). □

References

[1] Chiranjib Bhattacharyya and S. Sathiya Keerthi. Mean field methods for a special class of belief networks. *J. Artif. Intell. Res. (JAIR)*, 15:91–114, 2001. MR-1884078

[2] Devdatt Dubhashi and Desh Ranjan. Balls and bins: a study in negative dependence. *Random Struct. Algorithms*, 13(2):99–124, September 1998. MR-1642566

[3] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963. MR-0144363

[4] Michael J. Kearns and Lawrence K. Saul. Large deviation methods for approximate probabilistic inference. In *UAI*, 1998.

[5] Gábor Lugosi. Concentration-of-measure inequalities, `http://www.econ.upf.es/~lugosi/anu.ps`. 2003.

[6] Andreas Maurer. Thermodynamics and concentration. *Bernoulli*, 18(2):434–454, 2012. MR-2922456

[7] David A. McAllester and Luis E. Ortiz. Concentration inequalities for the missing mass and for histogram rule error. *Journal of Machine Learning Research*, 4:895–911, 2003. MR-2076001

[8] David A. McAllester and Robert E. Schapire. On the convergence rate of good-turing estimators. In *COLT*, 2000.

[9] Andrew Y. Ng and Michael I. Jordan. Approximate inference algorithms for two-layer bayesian networks. In *NIPS*, 1999.

[10] XuanLong Nguyen and Michael I. Jordan. On the concentration of expectation and approximate inference in layered networks. In *NIPS*, 2003.

[11] Maxim Raginsky and Igal Sason. Concentration of Measure Inequalities in Information Theory, Communications and Coding. arXiv:1212.4663, 2012.

Acknowledgments. We thank Maxim Raginsky for enlightening conversations and the anonymous referee for corrections to the manuscript.