

# On the Convergence of the Empirical Distribution

Daniel Berend

Department of Mathematics and Department of Computer Science  
Ben-Gurion University  
Beer Sheva, Israel

Aryeh Kontorovich

Department of Computer Science  
Ben-Gurion University  
Beer Sheva, Israel

June 6, 2012

## Abstract

We develop a general technique for bounding the tail of the total variation distance between the empirical and the true distributions over countable sets. Our methods sharpen a deviation bound of Devroye (1983) for distributions over finite sets, and also hold for the broader class of distributions with countable support. We also provide some lower bounds of possible independent interest.

## 1 Introduction

Establishing conditions and rates for the convergence of empirical frequencies to their expected values is a central problem in statistics. For concreteness, let  $X$  be an  $\mathbb{N}$ -valued random variable distributed according to  $\mathbf{p} = (p_1, p_2, \dots)$  and let  $X_1, X_2, \dots, X_n$  be  $n$  independent copies of  $X$ . The canonical estimator for  $p_j$  is obtained via the maximum likelihood principle, which just amounts to a normalized frequency:

$$\hat{p}_j^{(n)} = \frac{1}{n} \sum_{i=1}^n \mathbb{1}_{\{X_i=j\}}, \quad j \in \mathbb{N}.$$

The weak law of large numbers guarantees that  $\hat{p}_j^{(n)} \xrightarrow[n \rightarrow \infty]{\text{probability}} p_j$  for all  $j \in \mathbb{N}$ . The Chernoff-Hoeffding bound  $P\left(\left|\hat{p}_j^{(n)} - p_j\right| > \varepsilon\right) \leq 2 \exp(-2n\varepsilon^2)$ , together with the Borel-Cantelli lemma, strengthens the convergence to be almost sure, thus establishing a strong law of large numbers. A uniform strong law of large numbers is provided by the Dvoretzky-Kiefer-Wolfowitz inequality [6, 10]

$$P\left(\sup_{i \in \mathbb{N}} \left|\hat{F}_n(i) - F(i)\right| > \varepsilon\right) \leq 2 \exp(-2n\varepsilon^2), \quad \varepsilon > 0, n \in \mathbb{N},$$

where  $\hat{F}_n(i) = \sum_{j \leq i} \hat{p}_j^{(n)}$  and  $F(i) = \sum_{j \leq i} p_j$ . Indeed, since  $\hat{p}_j^{(n)} = \hat{F}_n(j) - \hat{F}_n(j-1)$  and  $p_j = F(j) - F(j-1)$ , we have

$$\begin{aligned} \left| \hat{p}_j^{(n)} - p_j \right| &= \left| (\hat{F}_n(j) - \hat{F}_n(j-1)) - (F(j) - F(j-1)) \right| \\ &\leq \left| \hat{F}_n(j) - F(j) \right| + \left| \hat{F}_n(j-1) - F(j-1) \right| \end{aligned}$$

and therefore

$$P \left( \left\| \hat{\mathbf{p}}^{(n)} - \mathbf{p} \right\|_{\infty} > \varepsilon \right) \leq 4 \exp(-n\varepsilon^2/2), \quad \varepsilon > 0.$$

We conclude that  $\left\| \hat{\mathbf{p}}^{(n)} - \mathbf{p} \right\|_{\infty} \xrightarrow[n \rightarrow \infty]{} 0$  almost surely (again, Borel-Cantelli is invoked).

An even stronger observation is that  $\left\| \hat{\mathbf{p}}^{(n)} - \mathbf{p} \right\|_1 \xrightarrow[n \rightarrow \infty]{} 0$  almost surely. The  $\ell_1$  distance is in some sense the most natural one over distributions [7], since by Scheffé's identity [5],

$$2 \sup_{E \subseteq \mathbb{N}} |\mathbf{p}(E) - \mathbf{q}(E)| = \|\mathbf{p} - \mathbf{q}\|_1,$$

for any two distributions  $\mathbf{p}, \mathbf{q}$  over  $\mathbb{N}$  (for this reason,  $\ell_1$  is also referred to as the *total variation* distance). Almost-sure convergence in  $\ell_1$  may be surmised from Sanov's theorem [2, 3] — whose drawback, however, is that it does not readily yield explicit, analytically tractable estimates for  $P \left( \left\| \hat{\mathbf{p}}^{(n)} - \mathbf{p} \right\|_1 > \varepsilon \right)$ .

Actually, Sanov's theorem guarantees that  $\hat{\mathbf{p}}^{(n)} \xrightarrow[n \rightarrow \infty]{} \mathbf{p}$  in yet a stronger sense, which may be called *complete convergence in  $\ell_1$* . Complete convergence was introduced in [8]. Applied to the random variable

$$J_n = \left\| \hat{\mathbf{p}}^{(n)} - \mathbf{p} \right\|_1, \tag{1}$$

it means that

$$\sum_{n=1}^{\infty} P(J_n > \varepsilon) < \infty$$

for all  $\varepsilon > 0$ . For  $\mathbf{p} \in \mathbb{R}^k$  (that is, distributions with support of size  $k$ ), one may combine the Chernoff-Hoeffding and the union bounds to obtain the following rough estimate:

$$P(J_n > \varepsilon) \leq 2k \exp(-2n\varepsilon^2/k^2). \tag{2}$$

Though crude, (2) suffices to establish the complete convergence in  $\ell_1$  of  $\hat{\mathbf{p}}^{(n)}$  to  $\mathbf{p}$  for distributions with finite support. A significant improvement is given by [4, Lemma 3], which may be stated as follows:

**Lemma 1** (Devroye). *For  $\mathbf{p} \in \mathbb{R}^k$ , we have*

$$P(J_n > \varepsilon) \leq 3 \exp(-n\varepsilon^2/25), \quad \varepsilon \geq \sqrt{20k/n}.$$

However, for  $\mathbf{p} \in \mathbb{R}^{\mathbb{N}}$  with infinite support, neither (2) nor Lemma 1 is applicable. Our goal in this paper is to establish analogues of Lemma 1 for distributions with countable support. As a by-product, we improve Devroye's Lemma, sharpening the constant in the exponent by an order of magnitude.

## 2 Main results

Our basic work-horse is McDiarmid’s inequality [11], which implies that whenever  $X_i$ ,  $i = 1, \dots, n$ , are independent  $\mathbb{N}$ -valued random variables and  $h : \mathbb{N} \rightarrow \mathbb{R}$  is 1-Lipschitz with respect to the Hamming metric<sup>1</sup>, we have

$$P(h(X_1, \dots, X_n) > \mathbf{E}h(X_1, \dots, X_n) + n\varepsilon) \leq \exp(-2n\varepsilon^2), \quad n \in \mathbb{N}, \varepsilon > 0. \quad (3)$$

We choose  $h$  to be the function mapping a sample  $(X_1, \dots, X_n)$  to the  $\ell_1$  deviation of the empirical frequencies from their expected values:

$$h(X_1, \dots, X_n) = \sum_{j \in \mathbb{N}} \left| np_j - \sum_{i=1}^n \mathbb{1}_{\{X_i=j\}} \right|.$$

In the notation above,  $h(X_1, \dots, X_n) = nJ_n$ . Since  $h$  is 2-Lipschitz under the Hamming metric (Lemma 7), it follows from (3) that

$$P(J_n > \mathbf{E}J_n + \varepsilon) \leq \exp(-n\varepsilon^2/2), \quad n \in \mathbb{N}, \varepsilon > 0. \quad (4)$$

(In fact, this estimate is near-optimal, as follows from an argument in the spirit of [1, Theorem 1].)

Hence, the crux of the matter is to bound  $\mathbf{E}J_n$ . For  $\mathbf{p} \in \mathbb{R}^k$ , it turns out that  $\mathbf{E}J_n \leq \sqrt{k/n}$ , which implies our first result:

**Theorem 2.** *For every  $k \in \mathbb{N}$ , distribution  $\mathbf{p} \in \mathbb{R}^k$ , and sample size  $n$ ,*

$$P(J_n > \varepsilon) \leq \exp\left(-\frac{n}{2} \left(\varepsilon - \sqrt{\frac{k}{n}}\right)^2\right), \quad \varepsilon \geq \sqrt{\frac{k}{n}}.$$

Observe that for  $\varepsilon \geq \sqrt{20k/n}$ , Theorem 2 yields  $P(J_n > \varepsilon) \leq \exp(-0.3n\varepsilon^2)$ , thus improving Lemma 1.

Our technique works just as well for  $\mathbf{p} \in \mathbb{R}^{\mathbb{N}}$  with infinite support. Indeed, as we show in Lemma 8,

$$\sqrt{n}\mathbf{E}J_n \leq \sum_{j \in \mathbb{N}} \sqrt{p_j} =: \nu(\mathbf{p}), \quad n \in \mathbb{N}. \quad (5)$$

When the right-hand side of (5) is finite (as is the case for “most” common distributions), the following result provides a simple and informative bound:

**Theorem 3.** *When  $\nu(\mathbf{p})$  is finite,*

$$P(J_n > n^{-1/2}\nu(\mathbf{p}) + \varepsilon) \leq \exp(-n\varepsilon^2/2), \quad n \in \mathbb{N}, \varepsilon > 0.$$

---

<sup>1</sup> The Hamming metric is defined by  $d(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^n \mathbb{1}_{\{x_i \neq y_i\}}$  for  $\mathbf{x}, \mathbf{y} \in \mathbb{N}^n$ .

When  $\nu(\mathbf{p})$  is infinite, we can still extract meaningful bounds, albeit with a bit more effort. As we show in Lemma 9,

$$\mathbf{E}J_n \leq \alpha_n(\mathbf{p}) + \beta_n(\mathbf{p}), \quad (6)$$

where

$$\alpha_n(\mathbf{p}) = 2 \sum_{p_j < 1/n} p_j, \quad \beta_n(\mathbf{p}) = \frac{1}{\sqrt{n}} \sum_{p_j \geq 1/n} \sqrt{p_j}. \quad (7)$$

At its most general, our result has the following form:

**Theorem 4.** *For all distributions  $\mathbf{p} \in \mathbb{R}^{\mathbb{N}}$ ,*

$$(i) \ P(J_n > \alpha_n + \beta_n + \varepsilon) \leq \exp(-n\varepsilon^2/2), \quad n \in \mathbb{N}, \ \varepsilon > 0.$$

$$(ii) \ \alpha_n + \beta_n \xrightarrow{n \rightarrow \infty} 0$$

(iii) *the rate of decay in (ii) may be arbitrarily slow.*

The bound in Theorem 4(i) may be rendered effective by our control over  $\alpha_n$  and  $\beta_n$  for specific distribution families. Moreover, our estimate in (6) for  $\mathbf{E}J_n$  in terms of  $\alpha_n$  and  $\beta_n$  is nearly tight, in the following sense:

**Proposition 5.** *For all  $n \geq 2$  and all distributions  $\mathbf{p} \in \mathbb{R}^{\mathbb{N}}$ ,*

$$\mathbf{E}J_n \geq \frac{\alpha_n + \beta_n}{4} - \frac{1}{\sqrt{n}}.$$

*Remark.* To keep the expressions simple, we have chosen  $1/n$  as the break-point in defining  $\alpha_n$  and  $\beta_n$ . We note in passing that a minor improvement in the constants is achieved by the (optimal) break-point  $1/4n$ .

The lower bound on  $\mathbf{E}J_n$  follows directly from the lemma below, in which the first inequality may be of independent interest:

**Lemma 6.** *If  $Y \sim \text{Bin}(n, p)$ , then*

$$\sqrt{np(1-p)/2} \leq \mathbf{E}|Y - np| \leq \sqrt{np(1-p)}, \quad n \geq 2, \ p \in [1/n, 1 - 1/n].$$

### 3 Proofs

We state the following elementary fact without proof:

**Lemma 7.** *Suppose  $n \in \mathbb{N}$  and  $\mathbf{p} \in \mathbb{R}^{\mathbb{N}}$  is a distribution. Define  $h : \mathbb{N}^n \rightarrow \mathbb{R}$  by*

$$h(\mathbf{x}) = \sum_{j \in \mathbb{N}} \left| np_j - \sum_{i=1}^n \mathbb{1}_{\{x_i=j\}} \right|, \quad \mathbf{x} \in \mathbb{N}^n.$$

*Then  $h$  is 2-Lipschitz with respect to the Hamming metric.*

**Lemma 8.** *Suppose  $n \in \mathbb{N}$  and  $\mathbf{p} \in \mathbb{R}^{\mathbb{N}}$  is a distribution. Then*

$$\sqrt{n}\mathbf{E}J_n \leq \sum_{j \in \mathbb{N}} \sqrt{p_j}.$$

*Proof.* Let  $Y_j \sim \text{Bin}(n, p_j)$ . Then

$$(\mathbf{E}|Y_j - np_j|)^2 \leq \mathbf{E}(Y_j - np_j)^2 = np_j(1 - p_j) \leq np_j,$$

whence

$$\mathbf{E}|Y_j - np_j| \leq \sqrt{np_j(1 - p_j)} \leq \sqrt{np_j}. \quad (8)$$

Since

$$n\mathbf{E}J_n = \sum_{j \in \mathbb{N}} \mathbf{E}|Y_j - np_j|, \quad (9)$$

the claim follows.  $\square$

**Lemma 9.** *Suppose  $n \in \mathbb{N}$  and  $\mathbf{p} \in \mathbb{R}^{\mathbb{N}}$  is a distribution. Then*

$$\mathbf{E}J_n \leq \alpha_n + \beta_n.$$

*Proof.* As in the proof of Lemma 8, let  $Y_j \sim \text{Bin}(n, p_j)$  and use (9) to obtain

$$n\mathbf{E}J_n = \sum_{p_j < 1/n} \mathbf{E}|Y_j - np_j| + \sum_{p_j \geq 1/n} \mathbf{E}|Y_j - np_j|. \quad (10)$$

By (8), the second term on the right-hand side of (10) is clearly upper-bounded by  $n\beta_n(\mathbf{p})$ . To bound the first term, we appeal to the *mean absolute deviation* formula for the binomial distribution [9]

$$\mathbf{E}|Y_j - np_j| = 2(1 - p_j)^{n - \lfloor np_j \rfloor} p_j^{\lfloor np_j \rfloor + 1} (\lfloor np_j \rfloor + 1) \binom{n}{\lfloor np_j \rfloor + 1}, \quad (11)$$

which simplifies to

$$\mathbf{E}|Y_j - np_j| = 2n(1 - p_j)^n p_j \leq 2np_j, \quad p_j < 1/n. \quad (12)$$

This shows that the first term on the right-hand side of (10) is upper-bounded by  $n\alpha_n(\mathbf{p})$  and proves the claim.  $\square$

*Proof of Theorem 2.* We claim that

$$\mathbf{E}J_n \leq \sqrt{\frac{k}{n}}. \quad (13)$$

Indeed, by Lemma 8,

$$\sqrt{n}\mathbf{E}J_n \leq \sum_{j=1}^k \sqrt{p_j}. \quad (14)$$

Define  $\mathbf{x} \in \mathbb{R}^k$  by  $x_j = \sqrt{p_j}$  and recall that

$$\sum_{j=1}^k \sqrt{p_j} = \|\mathbf{x}\|_1 \leq \sqrt{k} \|\mathbf{x}\|_2 = \sqrt{k}, \quad (15)$$

which yields (13). In view of (4), this implies the theorem.  $\square$

*Proof of Theorem 3.* Immediate from (4) and Lemma 8.  $\square$

**Lemma 10.** *Let  $\mathbf{p} \in \mathbb{R}^{\mathbb{N}}$  be a distribution. Then*

$$\alpha_n(\mathbf{p}) + \beta_n(\mathbf{p}) \xrightarrow{n \rightarrow \infty} 0.$$

*Proof.* The decay of  $\alpha_n(\mathbf{p})$  to zero is obvious, since it is the tail of a convergent series. To prove that

$$\lim_{n \rightarrow \infty} \frac{1}{\sqrt{n}} \sum_{p_j \geq 1/n} \sqrt{p_j} = 0, \quad (16)$$

we define the function  $\sigma : \mathbb{N} \rightarrow 2^{\mathbb{N}}$  by

$$\sigma(n) = \{j \in \mathbb{N} : p_j \geq 1/n\}.$$

Since as in (15),

$$\sum_{p_j \geq 1/n} \sqrt{p_j} \leq \sqrt{|\sigma(n)|},$$

it suffices to show that

$$|\sigma(n)| = o(n).$$

Suppose, to the contrary, that there exist a  $c > 0$  and an increasing sequence  $(n_k)_{k=1}^{\infty}$  such that

$$|\sigma(n_k)| \geq cn_k, \quad k \geq 1.$$

Put  $n_0 = 1$ . Passing to a subsequence, we may assume that  $n_k \geq 2n_{k-1}/c$  for every  $k \geq 1$ . Now

$$\begin{aligned} 1 &= \sum_{j=1}^{\infty} p_j \\ &\geq \sum_{k=1}^{\infty} \sum_{\frac{1}{n_k} \leq p_j < \frac{1}{n_{k-1}}} p_j \\ &\geq \sum_{k=1}^{\infty} (|\sigma(n_k)| - |\sigma(n_{k-1})|) \cdot \frac{1}{n_k} \\ &\geq \sum_{k=1}^{\infty} (cn_k - n_{k-1}) \cdot \frac{1}{n_k} \\ &\geq \sum_{k=0}^{\infty} (cn_k - cn_k/2) \cdot \frac{1}{n_k} = \sum_{k=0}^{\infty} \frac{c}{2} = \infty. \end{aligned}$$

The contradiction completes the proof.  $\square$

**Lemma 11.** For any rate sequence  $1 > r_1 > r_2 > \dots \searrow 0$ , there is a distribution  $\mathbf{p} \in \mathbb{R}^{\mathbb{N}}$  such that

$$\alpha_n(\mathbf{p}) + \beta_n(\mathbf{p}) > r_n, \quad n \in \mathbb{N}.$$

*Proof.* It suffices to show that there is no rate sequence bounding  $\alpha_n$ . But this is obvious, since  $\alpha_n$  may be expressed as the tail of a series converging to 2 — and although any such tail must decay to zero, the rate may be arbitrarily slow. In particular, given some rate sequence  $(r_n)$ , to ensure that  $\sum_{p_j \geq 1/n} p_j \leq 1 - r_n$  for each  $n \in \mathbb{N}$ , we may choose the appropriate  $p_j$  in an iterative greedy fashion, for  $n = 1, 2, \dots$   $\square$

*Proof of Theorem 4.* Item (i) is an immediate consequence of (4) and (6). Items (ii) and (iii) are the contents of Lemmas 10 and 11, respectively.  $\square$

*Proof of Lemma 6.* The upper bound is contained in (8) — and in fact, holds for all  $p$ . To establish the lower bound, let us rewrite the mean absolute deviation formula (11) as

$$\mathbf{E} |Y - np| = 2k \binom{n}{k} p^k (1-p)^{n-k+1}, \quad (k = \lfloor np \rfloor + 1).$$

Denote the right-hand side by  $E(n, k, p)$ , and put  $G(n, k, p) = 2E(n, k, p)^2 / (p(1-p))$ . The left-hand inequality in the lemma is equivalent to the claim

$$G(n, k, p) \geq n, \quad p \in [1/n, 1 - 1/n], \quad k = \lfloor np \rfloor + 1. \quad (17)$$

The domain where (17) is to be proved may be reparametrized by the inequalities

$$2 \leq k \leq n - 1, \quad \frac{k-1}{n} \leq p < \frac{k}{n}.$$

Now the function  $G(n, k, \cdot)$  is increasing on  $[(k-1)/n, (2k-1)/2n]$  and decreasing on  $[(2k-1)/2n, k/n]$  — and hence we need only consider the endpoints  $p = (k-1)/n$  and  $p = k/n$ .

To examine the first possibility, we take  $p = (k-1)/n$  and seek a  $k$  that minimizes  $G(n, k, (k-1)/n)$ . To this end, we consider the inequality  $G(n, k+1, k/n) \geq G(n, k, (k-1)/n)$ , which is equivalent (after a routine calculation) to

$$\left( \frac{k}{k-1} \right)^{2k-1} \geq \left( \frac{n-k+1}{n-k} \right)^{2n-2k+1}. \quad (18)$$

Since the function  $f(x) = (1 + 1/x)^{2x+1}$  is monotonically decreasing on  $[1, \infty)$ , the inequality (18) holds whenever  $k \leq (n+1)/2$ . We conclude that  $G(n, k, (k-1)/n)$  is minimized at the smallest allowed value of  $k$ , which is  $k = 2$ . We easily verify that the inequality  $G(n, 2, 1/2) \geq n$  is equivalent to  $8(n-1)^{2n-1} \geq n^{2n-1}$  for all  $n \geq 2$ , which again follows from the monotonicity of  $(1 + 1/x)^{2x+1}$ .

The second case,  $p = k/n$ , is analyzed in an exactly analogous manner.  $\square$

*Proof of Proposition 5.* Let  $n \geq 2$  and  $Y_j \sim \text{Bin}(n, p_j)$ . We group the probabilities as follows:  $S_1 = \{j : p_j < 1/n\}$ ,  $S_2 = \{j : 1/n \leq p_j \leq 1/2\}$  and  $S_3 = \{j : p_j > 1/2\}$ . By (12) and Lemma 6,

$$\mathbf{E} |Y_j - np_j| \geq \frac{1}{2} \begin{cases} np_j, & j \in S_1 \\ \sqrt{np_j}, & j \in S_2 \end{cases}.$$

Now

$$n\alpha_n(\mathbf{p}) = \sum_{j \in S_1} 2np_j \leq 4 \sum_{j \in S_1} \mathbf{E} |Y_j - np_j|$$

and

$$\begin{aligned} n\beta_n(\mathbf{p}) &= \sum_{j:p_j \geq 1/n} \sqrt{np_j} \\ &\leq \sum_{j \in S_2} \sqrt{np_j} + \sqrt{n} \\ &\leq 2 \sum_{j \in S_2} \mathbf{E} |Y_j - np_j| + \sqrt{n} \end{aligned}$$

and thus

$$4 \sum_{j \in S_1} \mathbf{E} |Y_j - np_j| + 2 \sum_{j \in S_2} \mathbf{E} |Y_j - np_j| + \sqrt{n} \geq n\alpha_n + n\beta_n,$$

which proves the claim. □

## Acknowledgements

We thank to Larry Wasserman for referring us to Devroye's Lemma, and David McAllester for reminding us about Sanov's Theorem. We are grateful to the Stone family for providing a venue for this work.

## References

- [1] Jan Beirlant, Luc Devroye, László Györfi, and Igor Vajda. Large deviations of divergence measures on partitions. *J. Statist. Plann. Inference*, 93(1-2):1–16, 2001.
- [2] Thomas M. Cover and Joy A. Thomas. *Elements of information theory*. Wiley-Interscience, Hoboken, NJ, second edition, 2006.
- [3] Frank den Hollander. *Large deviations*, volume 14 of *Fields Institute Monographs*. American Mathematical Society, Providence, RI, 2000.
- [4] Luc Devroye. The equivalence of weak, strong and complete convergence in  $L_1$  for kernel density estimates. *Ann. Statist.*, 11(3):896–904, 1983.
- [5] Luc Devroye and Gábor Lugosi. *Combinatorial methods in density estimation*. Springer Series in Statistics. Springer-Verlag, New York, 2001.
- [6] Aryeh Dvoretzky, Jack Kiefer, and Jacob Wolfowitz. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Ann. Math. Statist.*, 27:642–669, 1956.



- [7] Alison L. Gibbs and Francis E. Su. On choosing and bounding probability metrics. *International Statistical Review*, 70(3):419–435, 2002.
- [8] Pao-Lu Hsu and Herbert Robbins. Complete convergence and the law of large numbers. *Proc. Nat. Acad. Sci. U. S. A.*, 33:25–31, 1947.
- [9] John F. Kenney and Ernest S. Keeping. *Mathematics of Statistics, 3rd ed.* Princeton, NJ: Van Nostrand, 1962.
- [10] Pascal Massart. The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Ann. Probab.*, 18(3):1269–1283, 1990.
- [11] Colin McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, volume 141 of LMS Lecture Notes Series*, pages 148–188. Morgan Kaufmann Publishers, San Mateo, CA, 1989.