A coincidence-based test for uniformity given very sparsely-sampled discrete data

Liam Paninski
Department of Statistics, Columbia University
liam@stat.columbia.edu
http://www.stat.columbia.edu/∼liam

***Abstract*— How many independent samples N do we need from a distribution p to decide that p is ϵ-distant from uniform in an L_1 sense, $\sum_{i=1}^m |p(i)-1/m| > \epsilon$? (Here m is the number of bins on which the distribution is supported, and is assumed known *a priori*.) Somewhat surprisingly, we only need $N\epsilon^2 \gg m^{1/2}$ to make this decision reliably (this condition is both sufficient and necessary). The test for uniformity introduced here is based on the number of observed "coincidences" (samples that fall into the same bin), the mean and variance of which may be computed explicitly for the uniform distribution and bounded non-parametrically for any distribution that is known to be ϵ-distant from uniform. Some connections to the classical birthday problem are noted.**

***Index Terms*— Hypothesis testing, minimax, convex bounds.**

Introduction

We look at a rather basic problem: how many i.i.d. samples N are required to decide that a discrete distribution p, supported on m points, is nonuniform in an L_1 sense? More precisely, how large must the sample size N be so that we may test between the null hypothesis

$$H_0 : p_i \equiv 1/m$$

and the nonparametric alternative

$$H_A : \sum_{i=1}^{m} |p(i) - 1/m| > \epsilon$$

with error approaching zero? We will be interested in the sparse case $m \gg N$, where the classical chi-square theory does not apply.

This question has seen a great deal of analysis in both the computer science (Batu, 2001; Bellare and Kohno, 2004) and statistics (Diaconis and Mosteller, 1989) literature; in particular, there are obvious connections to the "birthday problem" (Camarri and Pitman, 2000; Das Gupta, 2005) and related techniques for entropy estimation (Nemenman et al., 2004). In fact, our analysis makes essential use of a version of the so-called "birthday inequality," which states that coincident birthdays are least likely when birthdays are uniformly distributed (Bloom, 1973; Munford, 1977; Clevenson and Watkins, 1991). The symmetry of the uniform distribution plays a key role here.

LP was partially supported by an NSF CAREER award.

It turns out that the uniformity testing problem is easy, in the sense that we may reliably detect departures from uniformity with many fewer samples N than bins m. In fact, it turns out that the condition $N\epsilon^2 m^{-1/2} \to \infty$ guarantees the consistency of a fairly simple test based on the number of "coincidences," samples that fall into the same bin. Thus, for fixed ϵ, we only really need about $N \gg \sqrt{m}$ samples. This is similar in spirit to the recent observation that estimating the entropy of discrete distributions is easy (Paninski, 2004) (in that case, $N = cm$ for any $c > 0$ suffices, and hence by a subsequence argument in fact slightly fewer than $\sim m$ samples are required to estimate the entropy on m bins). Thus it is much easier to test whether a distribution is uniform than to actually estimate the full distribution (this requires $N \gg m$ samples, as is intuitively clear and as can be made rigorous by a variety of methods, e.g. (Braess and Dette, 2004; Paninski, 2005)).

In addition, we prove a lower bound implying that N must grow at least as quickly as $\epsilon^{-2}m^{1/2}$ to guarantee the consistency of *any* test (not just the coincidence-based test introduced here); with fewer samples, any test will fail to detect the nonuniformity of at least one distribution in the alternate class H_A.

Upper bound

Our uniformity test will be based on "coincidences," that is, bins i for which more than one sample is observed. Alternatively, we may look at K_1, the number of bins into which just one sample has fallen; for N fixed, K_1 is clearly directly related to the negative number of coincidences. The basic idea, as in the birthday inequality, is that deviations from uniformity necessarily lead to an increase in the expected number of coincidences, or equivalently a decrease in $E(K_1)$.

To see this, we may directly write out the expectation of K_1 under a given p, using linearity of expectation:

$$E_p(K_1) = \sum_{i=1}^{m} \binom{N}{1} p_i (1-p_i)^{N-1}.$$

In the uniform case, $p_i \equiv 1/m$ and

$$E_u(K_1) = N \left(\frac{m-1}{m}\right)^{N-1}. \qquad (1)$$

Now we will compare these two expectations by computing the difference $E_u(K_1) - E_p(K_1) =$

$$N \left(\frac{m-1}{m}\right)^{N-1} \sum_{i=1}^{m} p_i \left[1 - \left(\frac{m}{m-1}(1-p_i)\right)^{N-1}\right].$$

After some approximations and an application of Jensen's inequality, we have the following key lower bound on $E(K_1)$ in terms of the distance from uniformity ϵ:

Lemma 1.

$$E_u(K_1) - E_p(K_1) \geq \frac{N^2\epsilon^2}{m}[1 + O(N/m)] \qquad \forall p \in H_A.$$

(A technical note: as noted above, we restrict our attention to the "sparse" regime $N = o(m)$, where direct estimation of the underlying distribution p is not feasible (Braess and Dette, 2004; Paninski, 2005).)

Proof:

Making the abbreviation

$$f(p_i) = p_i \left[1 - \left(\frac{m}{m-1}(1-p_i)\right)^{N-1}\right],$$

we have

$$E_u(K_1) - E_p(K_1) = N\left(\frac{m-1}{m}\right)^{N-1} \sum_{i=1}^{m} f(p_i). \tag{2}$$

The function $f(x)$ has a fairly simple form: $f(0) = 0$, $f(1/m) = 0$, $f(x) < 0$ for $0 < x < 1/m$, $f(x)$ is monotonically increasing for $x > 1/m$, and $f(x) \to x$ as x becomes large. However, $f(x)$ is not convex. To develop a lower bound on $E_u(K_1) - E_p(K_1)$, we develop a convex lower bound on $f(x)$, valid for all $x \in [0, 1]$ when $N \leq m$:

$$f(x) \geq g(|x - 1/m|) + f'(1/m)(x - 1/m),$$

with $g(z) =$

$$\begin{cases} f(z + 1/m) - f'(1/m)z, & z \in [0, 1/N - 1/m] \\ f(1/N) + (z + 1/m - 1/N) - f'(1/m)z, & o.w. \end{cases}$$

This lower bound on $f(.)$ looks more complicated than it is: for values of x close to $1/m$, where $f(x)$ is convex, we have simply reflected $f(x)$ about the point $1/m$ and added a line in order that the reflected function is smooth. For $x > 1/N$, we have replaced f with a linear lower bound of slope 1 (the limiting slope for $f(x)$ for large x). Here the point $x = 1/N$ is chosen as the solution of the equation

$$f'(x) = 1, \; 1/m < x < 1;$$

this solution exists uniquely when $m > N$. The derivative $f'(1/m)$ is easily computed as

$$f'(1/m) = (N-1)/(m-1),$$

and similarly we may directly compute $f'(1/N) = 1$. The key is that $g(|z|)$ is convex, symmetric, and strictly increasing in its argument z; see Fig. 1 for an illustration.

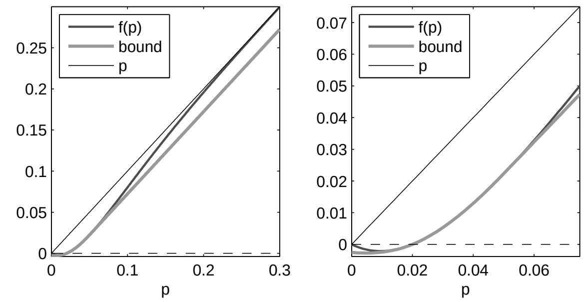

Fig. 1. Illustration of the convex lower bound on the function $f(p)$ in equation (2). Right panel is just a zoomed-in version of the left panel. $N = 20$; $m = 50$.

Now we subtract off the line and then apply Jensen. First, we have, for any constant c,

$$\begin{aligned} \sum_i [f(p_i) - c(p_i - 1/m)] &= \sum_i f(p_i) - \sum_i c(p_i - 1/m) \\ &= \sum_i f(p_i) - c\left[(\sum_i p_i) - 1\right] \\ &= \sum_i f(p_i); \end{aligned}$$

in particular, we have that

$$\begin{aligned} \sum_i f(p_i) &\geq \sum_i [g(|p_i - 1/m|) + f'(1/m)(p_i - 1/m)] \\ &= \sum_i g(|p_i - 1/m|). \end{aligned}$$

Now Jensen implies

$$\frac{1}{m}\sum_i g(|p_i - 1/m|) \geq g\left(\frac{1}{m}\sum_i |p_i - 1/m|\right) \geq g(\epsilon/m),$$

where the last inequality is by the fact that g is increasing and $p \in H_A$. Thus we find that

$$\sum_i f(p_i) \geq mg(\epsilon/m),$$

and therefore

$$E_u(K_1) - E_p(K_1) \geq Nm\left(\frac{m-1}{m}\right)^{N-1} g(\epsilon/m).$$

Now we need to look at $g(.)$. Near 0, $g(.)$ behaves like a quadratic matched to f at the point $1/m$:

$$g(z) = \frac{A}{2}z^2 + o(z^2), \; z \to 0,$$

with

$$\begin{aligned} A &= \left.\frac{\partial^2 f(x)}{\partial x^2}\right|_{x=1/m} \\ &= \left(\frac{m}{m-1}\right)^{N-1}\Big[2(N-1)(1-x)^{N-2} \\ &\quad -(N-1)(N-2)x(1-x)^{N-3}\Big]_{x=1/m} \\ &= 2N + O(N^2/m). \end{aligned}$$

Thus, we have $E_u(K_1) - E_p(K_1)$

$$\begin{aligned} &\geq Nm\left(\frac{m-1}{m}\right)^{N-1}\left([N+O(N^2/m)]\frac{\epsilon^2}{m^2} + o(\frac{\epsilon^2}{m^2})\right) \\ &= \frac{N^2\epsilon^2}{m}[1+O(N/m)], \end{aligned}$$

which completes the proof. □

On the other hand, we may bound the variance of K_1 under p as follows:

Lemma 2.

$$Var_p(K_1) \leq E_u(K_1) - E_p(K_1) + O(N^2/m).$$

Proof: It is not difficult to compute $Var_p(K_1)$ exactly: $Var_p(K_1) =$

$$E_p(K_1) - E_p(K_1)^2 + N(N-1)\sum_{i\neq j} p_i p_j (1-p_i-p_j)^{N-2}.$$

However, we found it inconvenient to bound this formula directly. Instead, we use the Efron-Stein inequality (Steele, 1986)

$$Var(S) \leq \frac{1}{2}E\sum_{j=1}^{N}(S-S^{(i)})^2,$$

where S is an arbitrary function of N independent r.v.'s x_i and

$$S^{(i)} = S(x_1, x_2, \ldots, x_i', \ldots, x_N)$$

denotes S computed with x_i' substituted for x_i, where x_i' is an i.i.d. copy of x_i. We will apply this inequality to $S = K_1$, with x_i the independent samples from p.

Since we are dealing with i.i.d. samples here, by symmetry we may write $\frac{1}{2}E\sum_{j=1}^{N}(S-S^{(i)})^2 =$

$$\begin{aligned} &\frac{N}{2}E_{\{x_i\}_{1\leq i\leq N-1}\sim p}\Big[\sum_{i\leq i,j\leq m} p_i p_j \big(1(n_i = 0 \cap n_j > 0) \\ &\qquad +1(n_j = 0 \cap n_i > 0)\big)\Big] \\ &= N\sum_{i,j} p_i p_j P_{\{x_i\}_{1\leq i\leq N-1}\sim p}(n_i = 0 \cap n_j > 0) \\ &= N\sum_{i,j} p_i p_j (1-p_i)^{N-1}\left(1-\left(1-\frac{p_j}{1-p_i}\right)^{N-1}\right) \\ &= N\sum_{i,j} p_i p_j \left((1-p_i)^{N-1} - (1-p_i-p_j)^{N-1}\right) \\ &\leq N\sum_{j=1}^{m} p_j\left(1-(1-p_j)^{N-1}\right) \\ &= E_u(K_1) - E_p(K_1) + N\left(1-\left(\frac{m-1}{m}\right)^{N-1}\right) \\ &= E_u(K_1) - E_p(K_1) + O(N^2/m). \end{aligned}$$

(Here n_i denotes the number of samples observed to have fallen in bin i after $N-1$ samples have been drawn, the second-to-last equality follows from equation (2), and the inequality uses the fact that $(1-y)^n - (1-y-x)^n$ is a decreasing function of y for $n > 1$, $x \in [0,1]$, and $0 < y < 1-x$.) □

Now we may construct our test for H_0 versus H_A: we reject H_0 if

$$T \equiv E_u(K_1) - K_1 = N\left(\frac{m-1}{m}\right)^{N-1} - K_1 > T_\alpha,$$

for some threshold T_α.

Theorem 3. *The size of this test is*

$$P_u(T \geq T_\alpha) = O\left(\frac{N^2}{mT_\alpha^2}\right).$$

The power is greater than

$$P_p(T \geq T_\alpha) \geq 1 - \frac{E_u(K_1) - E_p(K_1) + O(N^2/m)}{(E_u(K_1) - E_p(K_1) - T_\alpha)^2},$$

uniformly over all alternatives $p \in H_A$. If

$$N^2\epsilon^4/m \to \infty,$$

then the threshold T_α may be chosen so that the size tends to zero and the power to one, uniformly over all $p \in H_A$ (i.e., this condition is sufficient for the test to be uniformly consistent). For example,

$$T_\alpha = N^2\epsilon^2/2m$$

suffices.

We should note that the above bounds are based on a simple application of Chebysheff's inequality and therefore are by no means guaranteed to be tight.

Proof: We have that $E_u(T) = 0$ and $V_u(T) = O(N^2/m)$ (by lemma 2), and therefore by Chebysheff the size is bounded by

$$P_u(T \geq T_\alpha) = O\left(\frac{N^2}{mT_\alpha^2}\right).$$

For the power, we have that

$$\begin{aligned} P_p(T < T_\alpha) &= P_p\left(T - E_p(T) < T_\alpha - E_p(T)\right) \\ &\leq \frac{E_p(T) + O(N^2/m)}{(E_p(T) - T_\alpha)^2}, \end{aligned}$$

again by lemma 2.

Now, by lemma 1, it is clear that for the size to tend to zero and the power to tend to one, it is sufficient that the "z-score"

$$\frac{N^2\epsilon^2/m}{(N^2/m + N^2\epsilon^2/m)^{1/2}} = \left(\frac{N^2\epsilon^4/m}{1+\epsilon^2}\right)^{1/2}$$

tends to infinity. Since $0 < \epsilon \leq 2$, and $\epsilon^4/(1+\epsilon^2) \sim \epsilon^4$ for $\epsilon \in [0, 2]$, the proof is complete. □

Lower bound

The theorem above states that $N^2\epsilon^4/m \to \infty$ is a sufficient condition for the existence of a uniformly consistent test of H_0 vs. H_A. The following result is a converse:

Theorem 4. *If $N^2\epsilon^4 < m \log 5$, then no test reliably distinguishes H_0 from H_A; more precisely, for any test with critical region B and size bounded away from one, the minimum power*

$$\inf_{p \in H_A} \int_B p(x)dx$$

remains bounded away from one.

Proof: It is a well-known (LeCam, 1986; Ritov and Bickel, 1990; Donoho and Liu, 1991) consequence of the classical Neyman-Pearson theory that no uniformly consistent test exists if the L_1 distance

$$\left|\left|u(\vec{x}) - \int_{q \in H_A} q(\vec{x})d\mu(q)\right|\right|_1,$$

is bounded away from 2 for any $\mu \in \mathcal{P}(H_A)$, with $\mathcal{P}(H_A)$ the class of all probability measures on H_A, and $||.||_1$ denoting the L_1 norm on the sample space $\vec{x} \in \{1, \ldots, m\}^N$ equipped with the counting measure.

We develop this bound for one particular tractable mixing measure μ. (We make no claims that this measure will lead to optimal bounds.) Assume that m is even. (An obvious modification applies if m is odd.) We choose q randomly according to the following distribution $\mu(q)$: choose $m/2$ independent Bernoulli r.v.'s $z_j \in \{-1, 1\}$ (i.e., z samples uniformly from the corners of the $m/2$-dimensional hypercube). Given $\{z_j\}$, set

$$q(i) = \begin{cases} (1 + \epsilon z_{i/2})/m & i \text{ even}, \\ (1 - \epsilon z_{(i+1)/2})/m & i \text{ odd}. \end{cases}$$

Such a q will be a probability measure satisfying the equality $||u - q||_1 = \epsilon$ (and therefore lie on the boundary of the alternate hypothesis class H_A) with probability one, assuming $\epsilon \leq 1$. We let $Q(\vec{x}) = \int q(\vec{x})d\mu(q)$ denote the resulting probability measure. Similar mixing measures have appeared, e.g., in (Ritov and Bickel, 1990); this mixture of indistinguishable distributions technique is a fundamental idea in the minimax density estimation literature.

To compute the corresponding bound, we use the elegant (albeit somewhat involved) method outlined in Pollard's "Asymptopia" minimax notes (Pollard, 2003).

1) First, we substitute a more manageable L_2 bound for the L_1 norm:

$$||Q - u||_1 \leq ||Q - u||_2.$$

2) Next, we write out the likelihood ratio:

$$Q = 2^{-m/2} \sum_{z \in \{-1,1\}^{m/2}} Q_z,$$

with

$$\frac{dQ_z}{du}(\vec{x}) = \prod_{j=1}^{N}(1 + G(x_j, z)),$$

where $G(x_j, z) = \epsilon z_{j/2}$ or $-\epsilon z_{(j+1)/2}$, depending as j is even or odd, respectively. Note that

$$E_u G(x_j, z) = 0$$

for all j, z. Define

$$\begin{aligned} \Delta(\vec{x}) &\equiv \frac{dQ}{du}(\vec{x}) \\ &= 2^{-m/2} \sum_z \left(1 + \sum_{j=1}^{N} G(x_j, z) + \sum_{j > j'} G(x_j, z)G(x_{j'}, z) + \ldots\right), \end{aligned}$$

the sums ending with the N-fold product.

3) Now we expand the L_2 norm:

$$||Q - u||_2^2 = E_u(\Delta - 1)^2$$

$$(\Delta - 1)^2 = 2^{-m} \sum_{z,z'} \left(1 + \sum_j G(x_j, z) + \sum_{j > j'} G(x_j, z)G(x_{j'}, z') + \ldots\right).$$

Because of the independence of x_j and and z, and the fact that G has zero mean, we may cancel all of the terms that are not products of the form

$$H_j(z, z') \equiv E_u G(x_j, z)G(x_j, z') = \frac{2\epsilon^2}{m} \sum_{i=1}^{m/2} v(z_i, z_i'),$$

with $v(z_i, z_i') = 1$ if $z_i = z_i'$ and $v(z_i, z_i') = -1$ otherwise. So we have

$$\begin{aligned} E_u(\Delta - 1)^2 &= 2^{-m} \sum_{z,z'} \Big(\sum_j H_j(z, z') \\ &\quad + \sum_{j>j'} H_j(z, z') H_{j'}(z, z') + \dots \Big) \\ &= 2^{-m} \sum_{z,z'} \prod_j (1 + H_j(z, z')) - 1. \end{aligned}$$

4) The above term may be regarded as an average over two i.i.d. r.v.'s z and z':

$$E_u(\Delta - 1)^2 = E_{z,z'} \prod_j (1 + H_j(z, z')) - 1.$$

5) Now we use $\log(1 + t) \leq t$:

$$\prod_j (1 + H_j(z, z')) \leq \exp\left(\sum_j H_j(z, z') \right).$$

6) Finally, we compute

$$E_{z,z'} \exp\left(\sum_j H_j(z, z') \right) = \left(\frac{1}{2} \exp\left(\frac{2N\epsilon^2}{m} \right) + \frac{1}{2} \exp\left(-\frac{2N\epsilon^2}{m} \right) \right)^{m/2}$$

and use the bound

$$\frac{1}{2} \left(\exp(u) + \exp(-u) \right) \leq \exp\left(\frac{u^2}{2} \right),$$

to obtain

$$E_{z,z'} \exp\left(\sum_j H_j(z, z') \right) \leq \exp\left(\frac{N^2 \epsilon^4}{m} \right).$$

Putting everything together,

$$||Q - u||_1 \leq \left(\exp\left(\frac{N^2 \epsilon^4}{m} \right) - 1 \right)^{1/2}$$

Thus if $N^2 m^{-1} \epsilon^4$ is not sufficiently large, then $||Q-u||_1$ is bounded away from 2, and no uniformly consistent test exists. □

An alternate lower bound

The above result provides a quantitative, nonasymptotic lower bound on the error probability, but the bound is loose and the result becomes useless for a fixed ϵ if N^2/m becomes too large. It is worth deriving a simpler, asymptotic result to handle this case of large but bounded N^2/m:

Theorem 5. *If N^2/m remains bounded, then no test reliably distinguishes H_0 from H_A.*

Proof: The proof here is much more direct. We write out the ratio of marginal likelihoods, using the same uniform-hypercube mixture prior on H_A as above. Letting n_i denote the number of samples observed to have fallen into the i-th bin, we have

$$\begin{aligned} \frac{L(\vec{n}|H_A)}{L(\vec{n}|H_0)} &= E_{\vec{z}} \prod_{i=2,4,\dots,m} (1 - z_{i/2}\epsilon)^{n_{i-1}} (1 + z_{i/2}\epsilon)^{n_i} \\ &= \prod_{i=2,4,\dots m} E (1 - z_{i/2}\epsilon)^{n_{i-1}} (1 + z_{i/2}\epsilon)^{n_i} \\ &= \prod_{i=2,4,\dots m} \Big((1+\epsilon)^{n_{i-1}} (1-\epsilon)^{n_i} \\ &\qquad + (1+\epsilon)^{n_i} (1-\epsilon)^{n_{i-1}} \Big) / 2 \\ &= \prod_{i=2,4,\dots m} (1 - \epsilon^2)^{m_i} \left((1-\epsilon)^{d_i} + (1+\epsilon)^{d_i} \right) / 2 \\ &= \prod_{i=2,4,\dots m} (1 - \epsilon^2)^{m_i} \left(1 + \binom{d_i}{2} \epsilon^2 + \binom{d_i}{4} \epsilon^4 + \dots \right), \end{aligned}$$

where we have abbreviated $m_i = \min(n_i, n_{i-1})$ and $d_i = |n_i - n_{i-1}|$, and used the independence of z_j. (We interpret $\binom{d_i}{k}$ as 0 whenever $d_i < k$.)

Now note that the above multiplicands are greater than one only if $d_i \geq 2$, and less than one only if $m_i \geq 1$. And since the number of "two-bin coincidences" — pairs of bins into which two or more samples have fallen — is bounded in probability if $N = O(\sqrt{m})$, the likelihood ratio is bounded in probability as well, implying that the error probability of any test is bounded away from zero, and the proof is complete.

Finally, it is worth noting that the expected numbers of the events $(m_i = 1, d_i = 0)$ and $(m_i = 0, d_i = 2)$ scale together, leading (after an expansion of the logarithm and a cancellation of the ϵ^2 terms) to exactly the $N^2\epsilon^4/m$ scaling we observed previously. □

Batu, T. (2001). *Testing Properties of Distributions*. PhD thesis, Cornell.

Bellare, M. and Kohno, T. (2004). Hash function balance and its impact on birthday attacks. *EUROCRYPT*, pages 401–418.

Bloom, D. (1973). A birthday problem. *The American Mathematical Monthly*, 80:1141–1142.

Braess, D. and Dette, H. (2004). The asymptotic minimax risk for the estimation of constrained binomial and multinomial probabilities. *Sankhya*, 66:707–732.

Camarri, M. and Pitman, J. (2000). Limit distributions and random trees derived from the birthday problem with unequal probabilities. *Electronic Journal of Probability*, 5:1–18.

Clevenson, M. and Watkins, W. (1991). Majorization and the birthday inequality. *Mathematics Magazine*, 64:183–188.

Das Gupta, A. (2005). The matching, birthday and the strong birthday problem: a contemporary review. *J. Statist. Plann. Inference*, 130:377–389.

Diaconis, P. and Mosteller, F. (1989). Methods for studying coincidences. *Journal of the American Statistical Association*, 84:853–861.

Donoho, D. and Liu, R. (1991). Geometrizing rates of convergence. *Annals of Statistics*, 19:633–701.

LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, New York.

Munford, A. (1977). A note on the uniformity assumption in the birthday problem. *The American Statistician*, 31:119.

Nemenman, I., Bialek, W., and de Ruyter van Steveninck, R. (2004). Entropy and information in neural spike trains: Progress on the sampling problem. *Physical Review E*, 69:056111.

Paninski, L. (2004). Estimating entropy on m bins given fewer than m samples. *IEEE Transactions on Information Theory*, 50:2200–2203.

Paninski, L. (2005). Variational minimax estimation of discrete distributions under KL loss. *Advances in Neural Information Processing Systems*, 17.

Pollard, D. (2003). *Asymptopia*. www.stat.yale.edu/∼pollard.

Ritov, Y. and Bickel, P. (1990). Achieving information bounds in non- and semi-parametric models. *Annals of Statistics*, 18:925–938.

Steele, J. (1986). An Efron-Stein inequality for nonsymmetric statistics. *Annals of Statistics*, 14:753–758.