Chapter 9

Distance between multinomial and multivariate normal models

SECTION 1 introduces Andrew Carter's recursive procedure for bounding the Le Cam distance between a multinomial model and its approximating multivariate normal model.

SECTION 2 develops notation to describe the recursive construction of randomizations via conditioning arguments, then proves a simple Lemma that serves to combine the conditional bounds into a single recursive inequality.

SECTION 3 applies the results from Section 2 to establish the bound involving randomization of the multinomial distribution in Carter's inequality.

SECTION 4 sketches the argument for the bound involving randomization of the multivariate normal in Carter's inequality.

SECTION 5 outlines the calculation for bounding the Hellinger distance between a smoothed Binomial and its approximating normal distribution.

1. Introduction

The multinomial distribution $\mathcal{M}(n,\theta)$, where $\theta := (\theta_1,\ldots,\theta_m)$, is the probability measure on $\mathbb{Z}_+^m$ defined by the joint distribution of the vector of counts in m cells obtained by distributing n balls independently amongst the cells, with each ball assigned to a cell chosen from the distribution θ. The variance matrix nV_θ corresponding to $\mathcal{M}(n,\theta)$ has (i,j)th element $n\theta_i\{i=j\} - n\theta_i\theta_j$.

The central limit theorem ensures that $\mathcal{M}(n,\theta)$ is close to the $N(n\theta, nV_\theta)$, in the sense of weak convergence, for fixed m when n is large. In his doctoral dissertation, Andrew Carter (2000*a*) considered the deeper problem of bounding the Le Cam distance $\Delta(\mathcal{M},\mathcal{N})$ between models $\mathcal{M} := \{\mathcal{M}(n,\theta) : \theta \in \Theta\}$ and $\mathcal{N} := \{N(n\theta, nV_\theta) : \theta \in \Theta\}$, under mild regularity assumptions on Θ. For example, he proved that

<1> $$\Delta(\mathcal{M},\mathcal{N}) \le C'_\Theta \frac{m\log m}{\sqrt{n}} \qquad \text{provided } \sup_{\theta\in\Theta} \frac{\max_i \theta_i}{\min_i \theta_i} \le C_\Theta < \infty,$$

for a constant C'_Θ that depends only on C_Θ. From this inequality he was able to recover most of a result due to Nussbaum (1996), establishing an asymptotic

equivalence (in Le Cam's sense) between a density estimation model and a white noise model. By means of an extension (Carter & Pollard 2000) of Tusnady's lemma, Carter was also able to sharpen his bound under further "smoothness assumptions" on Θ, and thereby deduce the Nussbaum equivalence under the same conditions as Nussbaum (Carter 2001).

I feel that Carter's methods are a significant contribution to the study of the Le Cam distance. The following discussion is based on the December 2000 version of the Carter (2000*b*) preprint, but with many notational changes. For a more detailed account, the reader should consult Carter's preprints and dissertation.

The proof for <1> uses only the following basic results.

Facts

[1] Let $\mathfrak{U}$ denote the Uniform distribution on $(-1/2, 1/2)$. Then

$$H^2\left(\text{Bin}(n, p) \star \mathfrak{U}, N(np, npq)\right) \le \frac{C}{(1+n)pq},$$

where C is a universal constant.

[2] If X has a Bin(n, p) distribution then $(1+n)\mathbb{P}(1+X)^{-1} \le p^{-1}$.

[3] For all $\sigma_2^2 > 0$,

$$H^2\left(N(\mu_1, \sigma_1^2), N(\mu_2, \sigma_2^2)\right) \le \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2} + \frac{4|\sigma_1^2-\sigma_2^2|^2}{\sigma_2^4}.$$

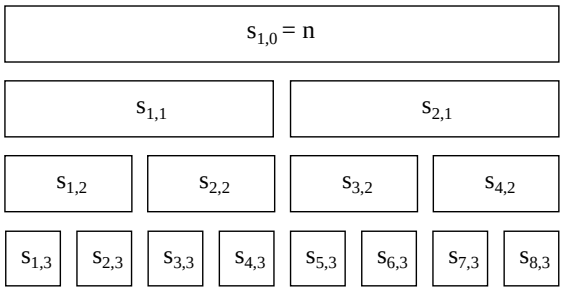
[4] For all probability measures $\{\alpha_i\}$ and $\{\beta_i\}$,

$$\| \otimes_i \alpha_i - \otimes_i \beta_i \|^2 \le 4H^2(\otimes_i \alpha_i, \otimes_i \beta_i) \le 4\sum\nolimits_i H^2(\alpha_i, \beta_i).$$

An outline of the proof for [1] appears in Section 5. See Problem [2] for [2], and Problem [1] for [3]. See UGMTP §3.3 & Problem 4.18 for [4].

For simplicity of exposition, I will assume the number of cells to be a power of 2, that is, $m = 2^M$ for some positive integer M. Write the multinomial counts as $s_{1,M}, \ldots, s_{m,M}$, and regard them as the coordinate maps on $\mathbb{Z}_+^m$, equipped with the probability measure $\mathbb{P}_\theta := \mathcal{M}(n, \theta)$.

The main innovation in Carter's method is a recursive argument based on a decomposition of the multinomial into a collection of (conditional) Binomials. Inequality <1> will be derived by reducing the problem for a multinomial on m cells to an analogous problem for $m/2$ cells, then $m/4$ cells, and so on. Eventually we reach the trivial case with one cell, where the multinomial and multivariate normal models coincide. The argument is easiest to describe with the help of the following picture (for $m = 2^M = 8$), whose bottom row contains the multinomial counts $\{s_{j,M} : 1 \le j \le 2^M\}$ for all m cells.



The counts $\{s_{j,M-1} : 1 \le j \le 2^{M-1}\}$ in the $m/2$ boxes of the next row are obtained by adding together pairs of counts from the bottom row: $s_{j,M-1} := s_{2j-1,M} + s_{2j,M}$. And so on, until all n observations are collected in the single box in the top row. The picture could also be drawn as a binary tree, with the count at each parent node equal to the sum of the counts at the two children.

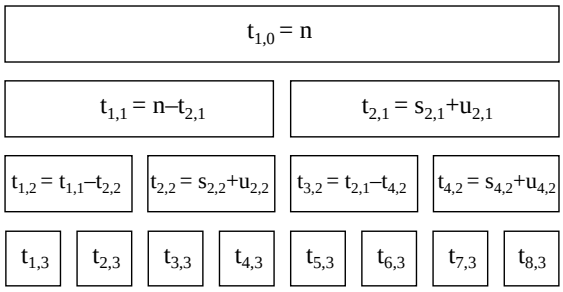
The simplicity of the method is largely due to the recursive structure. It replaces calculations involving the awkward dependences of the multinomial by simpler calculations based on conditional independence: given the counts in the $(M-1)$st row, the counts in the even-numbered boxes of the Mth row are independent Binomials.

> REMARK. Even though the picture seems to suggest some linear ordering of the cells of the multinomial, there is no such assumption behind <1>. The pairings could be made in an arbitrary fashion. There is no implied neighborhood structure on the cells. However, for the more precise results of Carter (2001), an ordering is needed to make sense of the idea of smoothness—similarity of probabilities for neighboring cells.

There is a corresponding recursive decomposition of the multivariate normal into a collection of normals, with the even-numbered variables in each row being conditionally independent given the values in the previous row.

The randomization to make the bottom row of the multinomial picture close, in a total variation sense, to the bottom row of the multivariate normal picture uses convolution smoothing, which is easiest to explain by means of a collection of independent observations $\{u_{j,k}\}$ drawn from $\mathfrak{U}$.

The smoothing works downwards from the top of the picture. We first define $t_{2,1} := s_{2,1} + u_{2,1}$ and then $t_{1,1} := n - t_{2,1}$ in order that the counts still sum to n. For the next row we define $t_{2j,2} := s_{2j,2} + u_{2j,2}$ and $t_{2j-1,2} := t_{j,1} - t_{2j,2}$, for $j = 1, 2$. That is, we directly smooth the counts in the even-numbered boxes, then adjust the counts in the odd-numbered boxes to ensure that the variable in each box is still equal to the sum of the variables in the two boxes beneath it. And so on.



These operations serve to define a Markov kernel $\mathbb{K} := \{\mathbb{K}_s : s \in \mathbb{Z}_+^m\}$ for which the joint distribution of the variables in the Mth row equals $\mathbb{K}\mathbb{P}_\theta$. The kernel corresponds to a convolution, $t_{j,M} = s_{j,M} + W_{j,M}$, where each $W_{j,m}$ is a sum of at most M of the independent $u_{i,j}$ variables. In consequence,

<2> $$\mathbb{K}_s^t \left(t_{j,M} - s_{j,M}\right)^2 \le M/12 \qquad \text{for all } j.$$

REMARK. Note that t_j is not normally distributed under $\mathcal{M}$. We should be careful when referring to the random variables t_j to specify the underlying probability measure.

The smoothing and the conditional independence will allow us to invoke Fact [1] repeatedly to bound the total variation distance between the conditional distribution of the smoothed multinomials and the conditional distributions for the normals. We then need to piece together the resulting bounds, using the method presented in the next Section.

2. Conditioning

Write r for $m/2$. To simplify notation, omit both the subscript indicating the choice of θ and the subscript indicating the row number, writing $s := (s_1, s_2, \ldots, s_r)$ for the counts in the $(M-1)$st row, and

$$\gamma(s, x) := (s_1 - x_1, x_1, \ldots, s_j - x_j, x_j, \ldots, s_r - x_r, x_r)$$

for the counts in the Mth row. Under the $\mathcal{M}$ model, the distribution for s is $\mu := \mathcal{M}(n, \psi)$, where $\psi_j := \theta_{2j-1} + \theta_{2j}$ for $j = 1, 2, \ldots, r$. Think of μ as a probability measure on $\mathcal{S} := \mathbb{Z}_+^r$. The conditional distribution for the Mth row, given s, is clearly determined by the conditional distribution of $x := (x_1, \ldots, x_r)$ given s,

<3> $$P_s := \otimes_{j=1}^r \text{Bin}(s_j, p_j) \qquad \text{where } p_j := \theta_{2j}/\psi_j.$$

The family $\mathcal{P} := \{P_s : s \in \mathcal{S}\}$ is a Markov kernel from $\mathcal{S}$ to $\mathcal{X} := \mathbb{Z}_+^r$. The joint distribution for s and x is the probability measure $\widetilde{\mathbb{P}} := \mu \otimes \mathcal{P}$, defined formally by

$$\widetilde{\mathbb{P}}^{s,x} f(s, x) := \mu^s P_s^x f(s, x) \qquad \text{for } f \in \mathcal{M}^+(\mathcal{S} \times \mathcal{X}).$$

The joint distribution, $\mathbb{P}$, for all the counts in the Mth row is the image of $\widetilde{\mathbb{P}}$ under the (one-to-one) linear map γ.

There is a similar decomposition for the normal model. Under $\mathcal{N}$, the $(M-1)$st row $t := (t_1, \ldots, t_r)$ has distribution $\lambda := N(n\psi, nV_\psi)$, a probability measure on $\mathcal{T} := \mathbb{R}^r$. The conditional distribution for the Mth row,

$$\gamma(t, y) := (t_1 - y_1, y_1, \ldots, t_j - y_j, y_j, \ldots, t_r - y_r, y_r),$$

given t is determined by the conditional distribution of $y := (y_1, \ldots, y_r)$ given t,

<4> $$Q_t := \otimes_{j=1}^r N(t_j p_j, n\psi_j p_j q_j) \qquad \text{where } q_j := 1 - p_j.$$

Notice that the jth factor in the product that defines Q_t has a slightly different form from $N(s_j p_j, s_j p_j q_j)$, the natural normal approximation for the jth factor in the product that defines P_s. (Of course the conditional variance for the normal model could not be a multiple of t_j, because there is a nonzero probability that $t_j < 0$ under $\mathcal{N}$.) The difference will cause only minor problems because $s_j \approx t_j \approx n\psi_j$ under $\mathbb{P}$.

The family $\mathcal{Q} := \{Q_t : t \in \mathcal{T}\}$ is a Markov kernel from $\mathcal{T}$ to $\mathcal{Y} := \mathbb{R}^r$. The joint distribution for t and y is the probability measure $\widetilde{\mathbb{Q}} := \lambda \otimes \mathcal{Q}$. The joint distribution, $\mathbb{Q}$, for all the variables in the Mth row is the image of $\widetilde{\mathbb{Q}}$ under the (one-to-one) linear map γ.

Section 1 defined a randomization $\mathbb{K}$ for which we hope $\|\mathbb{K}\mathbb{P} - \mathbb{Q}\|$ is small. Because the map γ is one-to-one, it makes no difference whether we work with $\widetilde{\mathbb{P}}$ and $\widetilde{\mathbb{Q}}$, rather than with $\mathbb{P}$ and $\mathbb{Q}$, when we calculate total variation distances. More precisely,

$$\|\mathbb{K}\mathbb{P} - \mathbb{Q}\| = \|\widetilde{\mathbb{K}}\widetilde{\mathbb{P}} - \widetilde{\mathbb{Q}}\| \qquad \text{where } \widetilde{\mathbb{K}} = \gamma^{-1}\mathbb{K}\gamma.$$

The construction from Section 1 can also be interpreted recursively. If we stop at the $(M-1)$st row, we have a randomization K for which we hope $\|K\mu - \lambda\|$ is small. The full randomization $\mathbb{K}$ is then obtained by further convolution smoothing to generate the Mth row of smoothed multinomials. In effect, we build $\mathbb{K}$ from K and a new randomization, which we can interpret as an attempt to match the conditional distribution P_s with the conditional distribution Q_t.

This interpretation fits with a general method for building randomizations a layer at a time. The following Lemma is written using notation borrowed from the preceding paragraphs, but it applies beyond the multinomial/multivariate normal problem. For the special case at hand, the randomization L in the Lemma consists of independent convolution smoothing of the x_j distributions. In general, L could also be allowed to depend on s and t.

<5> **Lemma.** *Let μ be a probability measure on $\mathcal{S}$, and $\mathcal{P} := \{P_t : t \in \mathcal{T}\}$ be a probability kernel from $\mathcal{T}$ to $\mathcal{X}$; and let λ be a probability measure on $\mathcal{T}$, and $\mathcal{Q} := \{Q_t : t \in \mathcal{T}\}$ be a probability kernel from $\mathcal{T}$ to $\mathcal{Y}$. Define $\widetilde{\mathbb{P}} := \mu \otimes \mathcal{P}$ and $\widetilde{\mathbb{Q}} = \lambda \otimes \mathcal{Q}$. Suppose there exist Markov kernels K, from $\mathcal{S}$ to $\mathcal{T}$, and, $L_{s,t}$, from $\mathcal{X}$ to $\mathcal{Y}$ for each (s, t), such that*

(i) $\|K\mu - \lambda\| \le \epsilon$

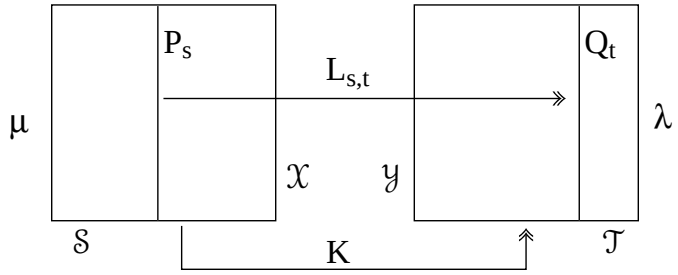
(ii) $\|L_{s,t}P_s - Q_t\| \le \rho(s, t)$ *for all* $s \in \mathcal{S}$ *and* $t \in \mathcal{T}$.

Let $\widetilde{\mathbb{K}}$ be the Markov kernel, from $\mathcal{S} \times \mathcal{X}$ to $\mathcal{T} \times \mathcal{Y}$, defined by

$$\widetilde{\mathbb{K}}^{t,y}_{s,x} f(t,y) := K^t_s L^y_{s,t,x} f(t,y) \qquad \text{for all } f \in \mathcal{M}^+(\mathcal{T} \times \mathcal{Y}).$$

Then

$$\|\widetilde{\mathbb{K}}\widetilde{\mathbb{P}} - \widetilde{\mathbb{Q}}\| \le \epsilon + (\mu \otimes K)^{s,t}\, \rho(s,t).$$



Proof. Remember that (i) and (ii) mean that

$$\sup_{|h|\le 1} |\mu^s K^t_s h(t) - \lambda^t h(t)| \le \epsilon$$
$$\sup_{|g|\le 1} |P^x_s L^y_{s,t,x} g(y) - Q^y_t g(y)| \le \rho(s,t) \qquad \text{for all } s,t.$$

For each function f on $\mathcal{T} \times \mathcal{Y}$ with $|f| \le 1$, define $h(t) := Q^y_t f(t,y)$ and $g_t(y) := f(t,y)$. Note that $|h| \le 1$ and $|g_t| \le 1$ for every t. Thus

$$\begin{aligned}
&|\widetilde{\mathbb{P}}^{x,s}\widetilde{\mathbb{K}}^{t,y}_{x,s} f(t,y) - \widetilde{\mathbb{Q}}^{t,y} f(t,y)| \\
&\quad = |\mu^s P^x_s K^t_s L^y_{s,t,x} f(t,y) - \mu^s K^t_s h(t) + \mu^s K^t_s h(t) - \lambda^t Q^y_t f(t,y)| \\
&\quad \le \mu^s K^t_s |P^x_s L^y_{s,t,x} g_t(y) - Q^y_t g_t(y)| + |\mu^s K^t_s h(t) - \lambda^t h(t)| \\
&\quad \le \mu^s K^t_s \rho(s,t) + \epsilon.
\end{aligned}$$

□ Take a supremum over f, with $|f| \le 1$, to complete the proof.

REMARK. Notice that the bound involves an integral with respect to $\mu \otimes K$, a distribution on $\mathcal{S} \times \mathcal{T}$. For the setting described in Section 1, this probability measure gives t has a multinomial distribution and s a "smoothed multinomial" distribution. That is, it refers to the joint distribution between the variables corresponding to the $(M-1)$st rows of the two pictures in that Section. As such, it combines the recursive effects of the smoothings that define the randomizations between all of the preceding rows.

3. From multinomial to multivariate normal

For the purposes of describing a recursive bound, add a subscript to the models, writing $\mathcal{M}_k$ for the multinomial with 2^k cells and $\mathcal{N}_k$ for the corresponding multivariate normal model. That is, $\mathcal{M}_M$ is the multinomial model with $m := 2^M$ cells in the Mth row, subject to the constraint $\max_i \theta_i / \min_i \theta_i \le C_\Theta$. The requirement $\sum_i \theta_i = 1$ gives

$$\frac{1}{C_\Theta m} \le \theta_i \le \frac{C_\Theta}{m} \qquad \text{for all } i.$$

The operation that combines counts from pairs of cells creates a new multinomial model $\mathcal{M}_{M-1}$ with 2^{M-1} cells and cell probabilities $\psi_j := \theta_{2j-1} + \theta_{2j}$ for $j = 1, 2, \ldots, 2^{M-1}$. Notice that $\max_i \psi_i \le 2 \max_i \theta_i$, and $\min_i \psi_i \ge 2 \min_i \theta_i$, from which it follows that

<6> $$\max_j \psi_j / \min_j \psi_j \le C_\Theta.$$

That is, all that changes in going from the Mth to the $(M-1)$st model is a halving of the number of cells. Similar considerations give a bound for the conditional probabilities $p_i := \theta_{2i}/\psi_i$,

<7> $$\frac{1}{1+C_\Theta} \le p_j \le \frac{C_\Theta}{1+C_\Theta} \qquad \text{for all } j.$$

Lemma <5> gives a bound for the Le Cam distance $\delta(\mathcal{M}_M, \mathcal{N}_M)$ if we add back the θ subscripts. The ϵ in the Lemma corresponds to $\delta(\mathcal{M}_{M-1}, \mathcal{N}_{M-1})$. The function $\rho(s,t)$ bounds the distance between products of smoothed multinomials and products of normals. By means of the Facts from Section 1 we can find a simple expression for $\rho(s,t)^2$:

$$\begin{aligned}
&\| \otimes_{j=1}^{m/2} \left(\text{Bin}(s_j, p_j) \star \mathfrak{U}\right) - \otimes_{j=1}^{m/2} N(t_j p_j, n\psi_j p_j q_j)\|^2 \\
&\quad \le 2\| \otimes_{j=1}^{m/2} \left(\text{Bin}(s_j, p_j) \star \mathfrak{U}\right) - \otimes_{j=1}^{m/2} N(s_j p_j, s_j p_j q_j)\|^2 \\
&\qquad + 2\| \otimes_{j=1}^{m/2} N(s_j p_j, s_j p_j q_j) - \otimes_{j=1}^{m/2} N(t_j p_j, n\psi_j p_j q_j)\|^2 \\
&\quad \le 8 \sum\nolimits_{j=1}^{m/2} H^2\left(\text{Bin}(s_j, p_j) \star \mathfrak{U}, N(s_j p_j, s_j p_j q_j)\right) \\
&\qquad + 8 \sum\nolimits_{j=1}^{m/2} H^2\left(N(s_j p_j, s_j p_j q_j), N(t_j p_j, n\psi_j p_j q_j)\right) \qquad \text{by Fact [4]} \\
&\quad \le 8 \sum\nolimits_{j=1}^{m/2} \frac{C}{(1+s_j) p_j q_j} + \frac{(s_j p_j - t_j p_j)^2}{2n\psi_j p_j q_j} + \frac{4|s_j p_j q_j - n\psi_j p_j q_j|^2}{(n\psi_j p_j q_j)^2}.
\end{aligned}$$

The last inequality comes from Facts [1] and [3].

Use <7> to tidy up the constants, suggesting the choice

$$\rho(s,t)^2 := C' \sum\nolimits_{j=1}^{m/2} \left(\frac{1}{1+s_j} + \frac{(s_j - t_j)^2}{n\psi_j} + \frac{(s_j - n\psi_j)^2}{(n\psi_j)^2} \right),$$

where C' is a constant depending only on C and C_Θ. Under the distribution $\Gamma := \mu \otimes K$, the random variable s_j has a $\text{Bin}(n, \psi_j)$ distribution and, by the analog of <2> for the $(M-1)$st row, $\Gamma(s_j - t_j)^2 \le (M-1)/2$. Invoking Fact [2] and the bound $\psi_j \ge 2/(mC_\Theta)$ we then get

$$\begin{aligned}
(\Gamma\rho(s,t))^2 \le \Gamma\rho(s,t)^2 &:= C' \sum\nolimits_{j=1}^{m/2} \left(\frac{1}{(1+n)p_j} + \frac{M-1}{2n\psi_j} + \frac{n\psi_j(1-\psi_j)}{(n\psi_j)^2} \right) \\
&\le C''(m/2)\,(1/n + Mm/n + m/n)\,.
\end{aligned}$$

Substitution into the result from Lemma <5> then leaves us with a recursive inequality,

$$\delta(\mathcal{M}_M, \mathcal{N}_M) \le \delta(\mathcal{M}_{M-1}, \mathcal{N}_{M-1}) + C'''m\sqrt{M/n},$$

with C''' yet another constant depending only on C_Θ.

A similar relationship holds between each pair of successive rows, with the same constants in each case because the cell probabilities always satisfy the analog of <6>. Substituting repeatedly, noting that $\delta(\mathcal{M}_0, \mathcal{N}_0) = 0$ because both models are degenerate, we then get

$$\delta(\mathcal{M}_M, \mathcal{N}_M) \le \delta(\mathcal{M}_0, \mathcal{N}_0) + C''' \sum_{k=1}^{M} 2^k \sqrt{k/n} \le C'_\Theta m \sqrt{\frac{\log m}{n}}.$$

4. From multivariate normal to multinomial (sketch)

The argument for approximating the multinomial by a randomization of the multivariate normal is very similar to the argument used in Section 3. We start from the decomposition $\{s_{j,k} : 1 \le j \le 2^k;\ k = 0, \ldots, M\}$ corresponding to the multivariate normal under the model $\mathcal{N}$. Instead of smoothing, we discretize by means of the function $[\cdot]$ that rounds a real number to its nearest integer. That is, for each row we define each $t_{2j,k} := [s_{2j,k}]$, then adjust the adjacent $s_{2j-1,k}$ to keep the sum $t_{2j-1,k} + t_{2j,k}$ equal to $t_{j,k}$. For the last row we have $t_{j,M} = s_{j,M} + V_{j,M}$ with each $V_{j,M}$ a sum of at most M terms of the form $\pm\big(s_{j,k} - [s_{j,k}]\big)$. The summands defining $V_{j,M}$ are dependent. Carter used the conservative bound

$$|t_{j,M} - s_{j,M}| \le M \qquad \text{for all } j,$$

which led to a factor of $\log m$ in <1>, rather than the factor $(\log m)^{1/2}$ suggested by the calculations in Section 3.

> REMARK. I cannot imagine that there is enough dependence between the fractional parts of the $s_{j,k}$ to make $\operatorname{var}\big(V_{j,M}\big)$ grow significantly faster than M. However, it does not seem worthwhile devoting great effort to improve the $\log m$ to a $(\log m)^{1/2}$, when it is not clear whether the $m/\sqrt{n}$ might not be improved by a more subtle randomization.

The argument via Lemma <5> proceeds much as before, except that now we need to bound the distance between the Binomial and the rounded normal. Actually, the bounds developed in Section 3 still apply, because the rounding operation can only decrease Hellinger distances.

5. Hellinger bound for smoothed Binomial

Fact [1] gives a bound for the Hellinger distance between a smooth Binomial and its approximating normal,

<8> $$H^2\big(\text{Bin}(n, p) \star \mathfrak{U},\ N(np, npq)\big) \le \frac{C}{(1+n)pq},$$

where C is a universal constant.

By choosing C large enough, we ensure that the asserted inequality is trivially true when npq is smaller than any fixed σ_0^2. Thus we need only consider the case where $\sigma^2 := npq \ge \sigma_0^2$. Write b_k for $\mathbb{P}\{\text{Bin}(n, p) = k\}$, and $g(x)$ for the

corresponding $N(np, npq)$ density. Using elementary Calculus and the Stirling formula, Prohorov (1961) proved that

$$\log \frac{b_k}{g(k)} = \frac{(q-p)}{6\sigma}(x^3 - x) + O\left(\frac{1+x^4}{\sigma^2}\right) \qquad \text{where } x := \frac{k-np}{\sigma}.$$

Actually, for our purposes it would suffice to have

<9> $$\log \frac{b_k}{g(k)} = R(x) \qquad \text{with } |R(x)| \le \frac{\pi(|x|)}{\sigma} \text{ for } |x| \le \sqrt{3\log\sigma},$$

for some polynomial $\pi(\cdot)$.

For each function f on the real line, write $\widetilde{f}$ for the associated step function,

$$\widetilde{f}(x) := \sum\nolimits_{k\in\mathbb{Z}} \{k - \tfrac{1}{2} < x \le k + \tfrac{1}{2}\} f(k).$$

The left-hand side of <8> is bounded above by

<10> $$2\sum_{k\in\mathbb{Z}} \left(b_k^{1/2} - g(k)^{1/2}\right)^2 + 2\int_{-\infty}^{\infty} \left(g(x)^{1/2} - \widetilde{g}(x)^{1/2}\right)^2 dx.$$

For the first term in <10>, split the sum according to whether k is in the range $A := \{k : |k - np| \le 3\sigma\sqrt{\log\sigma}\}$ or not. The contribution from $k \notin A$ is less than $\sum_{k\notin A}(b_k + g(k))$, which is easily bounded by a multiple of σ^{-2} using standard tail bounds for the Binomial and normal. The contribution from $k \in A$ equals

$$\begin{aligned}\sum\nolimits_{k\in A} g(k)\left(\frac{b_k^{1/2}}{g(k)^{1/2}} - 1\right)^2 &= \sum\nolimits_{k\in A} g(k)|e^{R(x)/2} - 1|^2 \qquad \text{by <9>} \\ &\le C_0\sigma^{-2}\sum\nolimits_{k\in A} g(k)\pi(x)^2,\end{aligned}$$

for a universal constant C_0. The sum is bounded by a multiple of $\mathbb{P}\pi(|Z|)^2$, with Z distributed $N(0,1)$.

The second term in <10> is disposed of by the following elementary calculation.

<11> **Lemma.** *Let f be the $N(\mu, \sigma^2)$ density. Then*

$$\int \left(f(x)^{1/2} - \widetilde{f}(x)^{1/2}\right)^2 dx \le \int \frac{f'(t)^2}{f(t)}\, dt = 1/(4\sigma^2).$$

Proof. Write h for $f^{1/2}$. Note that $h'(t) = f'(t)/2h(t)$. For $k \le x \le k + 1/2$ we have

$$|h(x) - h(k)| = |\int_k^x h'(t)\, dt| \le \int_k^{k+1/2} |h'(t)|\, dt.$$

There is a similar bound when $k \ge x \ge k - 1/2$. Thus

$$\begin{aligned}\int |h(x) - \bar{h}(x)|^2\, dx &\le \sum\nolimits_{k\in\mathbb{Z}} \int \{k - \tfrac{1}{2} < x \le k + \tfrac{1}{2}\}|h(x) - h(k)|^2\, dx \\ &\le \sum\nolimits_{k\in\mathbb{Z}} \left(\int \{|t-k| \le \tfrac{1}{2}\}|h'(t)|\, dt\right)^2 \le \int h'(t)^2\, dt.\end{aligned}$$

□

6. Problems

[1] Hellinger distance between normals.

(i) Show that

$$H^2\left(N(\mu_1,\sigma^2),N(\mu_1,\sigma^2)\right)=2-2\exp\left(-(\mu_1-\mu_2)^2/8\sigma^2\right).$$

(ii) Show that

$$H^2\left(N(\mu,\sigma_1^2),N(\mu,\sigma_2^2)\right)=2-2\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}.$$

(iii) Deduce that

$$\begin{aligned}H^2&\left(N(\mu_1,\sigma_1^2),N(\mu_2,\sigma_2^2)\right)\\&\le 2\left(2-2\exp\left(-(\mu_1-\mu_2)^2/8\sigma_2^2\right)\right)+2\left(2-2\sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2+\sigma_2^2}}\right)\\&\le\frac{(\mu_1-\mu_2)^2}{2\sigma_2^2}+\frac{4|\sigma_1^2-\sigma_2^2|^2}{\sigma_2^4}\end{aligned}$$

[2] Suppose X has a Bin(n,p) distribution. Show that

$$\mathbb{P}(1+X)^{-1}\le\int_0^1\mathbb{P}s^X\,ds\le\frac{1}{(n+1)p}.$$

References

Carter, A. (2000*a*), Asymptotic equivalence of nonparametric experiments, PhD thesis, Yale.

Carter, A. (2000*b*), Deficiency distance between multinomial and multivariate normal experiments, Technical report, University of California, Santa Barbara. [http://www.pstat.ucb.edu/~carter].

Carter, A. (2001), Le Cam distance between multinomial and multivariate normal experiments under smoothness constraints on the parameter set, Technical report, University of California, Santa Barbara. [http://www.pstat.ucb.edu/~carter].

Carter, A. & Pollard, D. (2000), Tusnády's inequality revisited, Technical report, Yale University. [http://www.stat.yale.edu/~pollard].

Nussbaum, M. (1996), 'Asymptotic equivalence of density estimation and gaussian white noise', *Annals of Statistics* **24**, 2399–2430.

Prohorov, Y. V. (1961), 'Asymptotic behavior of the binomial distribution', *Selected Translations in Mathematical Statistics and Probability* **1**, 87–95.