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.

[§intro] ## 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

carter1 <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 $\sigma_2^2 > 0$,

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

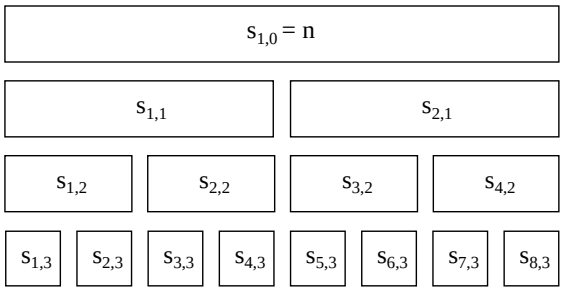
[4] For all probability measures $\{\alpha_i\}$ and $\{\beta_i\}$,

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

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.

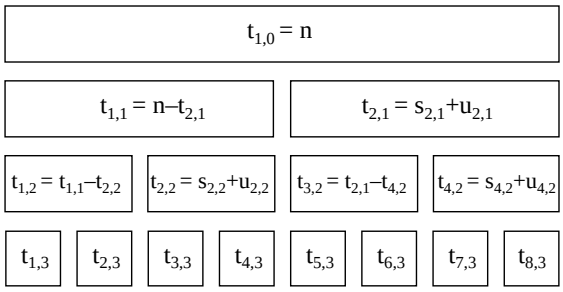
The simplicity of the method is largely due to the recursive structure. It replaces calculations involving the awkward dependences of the multinomial by simpler calculations based on conditional independence: given the counts in the $(M-1)$st row, the counts in the even-numbered boxes of the Mth row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

t-s <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.

[§product] ## 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,

Ps <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,

Qt <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.

product <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) $\|K\mu - \lambda\| \le \epsilon$

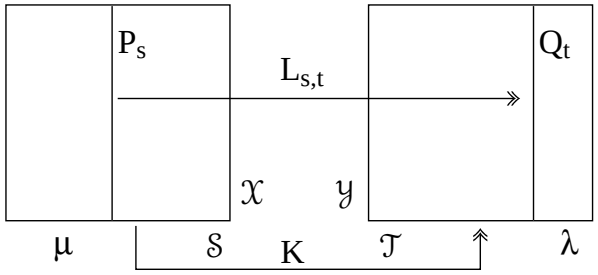
(ii) $\|L_{s,t}P_s - Q_t\| \le \rho(s, t)$ *for all* $s \in \mathcal{S}$ *and* $t \in \mathcal{T}$.

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.

[§mn.to.normal]

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

psi.ratio <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 $p_i := \theta_{2i}/\psi_i$,

p.bounds <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

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

[§normal.to.mn]

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.

[§hellinger]

5. Hellinger bound for smoothed Binomial

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

Hellinger.smoothed <8>

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

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 $\mathbb{P}\{\operatorname{Bin}(n, p) = k\}$, 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

weak.prohorov <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

sum.int <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.

step.approx <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

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

□

[§problems.multinomial]

6. Problems

hell.normal [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

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

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

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

prohorov [3] Give argument to derive <9> via Stirling.

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.ucsb.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.ucsb.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.

[§indep] **7. Independent normals**

Inequality <1> bounds the distance between the multinomial and multivariate normal models $\mathcal{M}$ and $\mathcal{N}$, where $\mathcal{N} := \{\mathbb{Q}_\theta : \theta \in \Theta\}$ with $\mathbb{Q}_\theta := N(n\theta, nV_\theta)$. Under $\mathbb{Q}_\theta$, the coordinates are dependent variables, whereas the increments of Nussbaum's white noise model are independent, with variances that do not depend on θ.

Carter completed his rederivation of Nussbaum's result by a sequence of steps leading from $\mathcal{N}$ to the white noise model. (He also presented analogous arguments for randomizing the white noise model to approximate $\mathcal{N}$, which I will not discuss. The ideas are almost the same for both directions.) He compared $\mathcal{N}$ with the white noise via two intermediate models:

(i) $\mathcal{N}_{\text{indep}} := \{\mathbb{N}_\theta : \theta \in \Theta\}$, where $\mathbb{N}_\theta = \otimes_{i\le m} N(n\theta_i, n\theta_i)$.

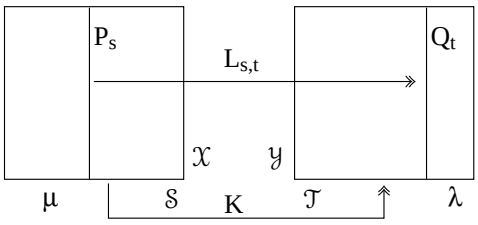
(ii) $\mathcal{N}_{\text{stabil}} := \{\widetilde{\mathbb{N}}_\theta : \theta \in \Theta\}$, where $\widetilde{\mathbb{N}}_\theta = \otimes_{i\le m} N(\sqrt{n\theta_i}, 1/4)$.

Finally, he used the random vectors with distribution $\widetilde{\mathbb{N}}_\theta$ to obtain (by interpolation) processes with continuous paths, which are close to the white noise processes.

I will discuss the interpolation in Section 9, the comparison of $\mathcal{N}_{\text{indep}}$ and $\mathcal{N}_{\text{stabil}}$ in Section 8, and the comparison of $\mathcal{N}$ and $\mathcal{N}_{\text{indep}}$ in the current Section.

Think of each $\mathbb{Q}_\theta$ as a probability measure on $\mathcal{X} := \mathbb{R}^m$ and each $\mathbb{N}_\theta$ as a probability measure on $\mathcal{Y} := \mathbb{R}^m$. Under $\mathbb{Q}_\theta$, the sum $s := \sum_{i\le m} x_i$ has a distribution μ that is degenerate at n. Under $\mathbb{N}_\theta$, the sum $t := \sum_{i\le m} y_i$ has distribution $\lambda := N(n, n)$. Moreover, under $\mathbb{N}_\theta$, the conditional distribution of y given t is $\mathbb{Q}_{\theta,t} := N(t\theta, nV_\theta)$. Note that $\mathbb{Q}_{\theta,n} = \mathbb{Q}_\theta$. The models $\mathcal{N}_{\text{indep}}$ from $\mathcal{N}$ have the same form of conditional distribution given the totals; they differ only in the distributions of the sums.

We are again in the situation covered by Lemma <5>, with $P_s := \mathbb{Q}_{\theta,s}$ and $Q_t := \mathbb{Q}_{\theta,t}$. Under μ, only the value $s = n$ is relevant, but under λ we need to consider all values of t.



We can take K_x as the $N(n, n)$ distribution for all x. (Of course, only the value $x = n$ is really needed.) We must choose the randomization $L_{s,t}$ to control

$$\|L_{n,t}N(n\theta, nV_\theta) - N(t\theta, nV_\theta)\|.$$

REMARK. It might seem that we could take $L_{n,t}$ as the identity map, since the t should be close to n with high probability under λ. However, closeness of means would not suffice, because V_θ is singular: if $t \ne n$ then $\|\mathbb{Q}_{\theta,t} - \mathbb{Q}_{\theta,n}\| = 2$. We need to match the means.

Choose $L_{n,t}$ as the map that corresponds to multiplication of a random vector by the constant t/n. That is, take $L_{n,t,x}$ as the point mass at tx/n. Then take

$$\rho(n,t) := \|N(t\theta, (t^2/n)V_\theta) - N(t\theta, nV_\theta)\| = \|N(0,(t/n)^2 nV_\theta) - N(0, nV_\theta)\|.$$

In general, for any $m \times m$ covariance matrix W and any constant c, both the $N(0,W)$ and the $N(0,c^2W)$ are image measures of products of independent normals under the linear map $W^{1/2}$. Total variation distance can only be decreased by such a map. Thus

$$\begin{aligned}\|N(0,\bar t^2 nV_\theta) - N(0,nV_\theta)\|^2 &\le \|N(0,\bar t^2 I_m - N(0,I_m)\|^2 \qquad \text{where } \bar t := t/n\\ &\le 4mH^2\left(N(0,1), N(0,\bar t^2)\right)\\ &\le 8m\left(\bar t^2 - 1\right)^2 \qquad \text{by Problem [1].}\end{aligned}$$

Under K_n, the random variable $\bar t$ has a $N(1,1/n)$ distribution. Thus

$$(\mu \otimes K\rho(s,t))^2 \le 8m\mathbb{P}\left((1+N(0,1/n))^2 - 1\right)^2 \le Cm/n,$$

for some constant C.

It follows that $\delta(\mathcal{N}, \mathcal{N}_{\text{indep}})$ is bounded by a constant multiple of $\sqrt{m/n}$, which is smaller than the bound in <1>. A similar argument gives a similar bound for $\delta(\mathcal{N}_{\text{indep}}, \mathcal{N})$. When combined with <1>, these bounds give, for some new constant C,

carter2 <12> $$\Delta(\mathcal{M}, \mathcal{N}_{\text{indep}}) \le C\frac{m\log m}{\sqrt{n}} \qquad \text{provided } \sup_{\theta\in\Theta}\frac{\max_i \theta_i}{\min_i \theta_i} \le C_\Theta < \infty.$$

[§sqrt] 8. Variance stabilizing transformations

Each $\mathbb{N}_\theta$ is a product measure, $\otimes_i N(n\theta_i, n\theta_i)$, but the variances still depend on θ. To remove this dependence, Carter (following the lead of Nussbaum) applied the classical method for variance stabilization.

Suppose X has a $N(\mu,\mu)$ distribution for a large positive μ. Let g be a smooth, increasing function. Then the random variable $Y = g(X)$ is approximated by $g(\mu) + g'(\mu)(X-\mu)$, which has a normal distribution with mean $g(\mu)$ and variance $\mu g'(\mu)^2$. If we choose g so that $g'(\mu)$ is proportional to $1/\sqrt{\mu}$, then the variance of the approximation will not depend on μ.

Clearly we should choose g so that $g(x) = \sqrt{x}$, at least for large positive x. For example, we could define $g(x) = \sqrt{x}\{x \ge 1\} + h(x)\{x < 1\}$, for an arbitrary smooth h that makes g behave smoothly at $x = 1$. Of course, the behavior for small x should be irrelevant—because the important contributions come from a region near μ—but it is inconvenient to have to worry about the possibility of taking the square root of a negative value for X.

If we apply the transformation g to each coordinate, we should obtain a new distribution close (under $\mathbb{N}_\theta$) to $\widetilde{\mathbb{N}}_\theta := \otimes_i N(\sqrt{n\theta_i}, 1/4)$. More formally, we have a map $G : \mathbb{R}^m \to \mathbb{R}^m$ for which we hope $\|G\mathbb{N}_\theta - \widetilde{\mathbb{N}}_\theta\|$ is small for all θ in Θ.

Using Taylor expansions, after separating out contributions from the lower tails, Carter was able to derive a bound on $D\left(g(N(\mu,\mu))\|N(\sqrt{\mu},1/4)\right)$. This bound gives the desired control over total variation distance between the product measures. And so on. In short, Carter showed that $\Delta(\mathcal{N}_{\text{indep}},\mathcal{N}_{\text{stabil}})$ is also small enough to be absorbed into the bound from <12>, leading to

carter3 <13>
$$\Delta(\mathcal{M},\mathcal{N}_{\text{stabil}}) \le C\frac{m\log m}{\sqrt{n}} \qquad \text{provided } \sup_{\theta\in\Theta}\frac{\max_i \theta_i}{\min_i \theta_i} \le C_\Theta < \infty.$$

for some new constant C.

Carter's method is satisfactory, but I feel it would be more elegant if we could appeal to some general result about variance stabilization rather than having to derive a special case. I outline below what I would like to include in the final version of these Paris notes. I invite elegant solutions from the audience before the end of May.

The general problem

Let X be a random variable whose distribution P has density f with respect to Lebesgue measure on $\mathbb{R}$. Let g be a smooth, increasing function on $\mathbb{R}$. Suppose f concentrates most of its mass near a value μ. The classical delta-method then asserts that the distribution of $g(X)$ should be close to the distribution Q of $Z := g(\mu) + g'(\mu)(X-\mu)$, which has density

$$q(z) = \frac{1}{g'(\mu)} f\left(\mu + \frac{z-g(\mu)}{g'(\mu)}\right)$$

with respect to Lebesgue measure.

For our purposes we need the distributions close in the total variation sense. More precisely, it will be useful to have a good bound for

$$H^2(g(P),Q) = H^2(P,g^{-1}(Q)).$$

The measure $g^{-1}(Q)$, which is the distribution of $g^{-1}(Z)$, has a density

$$f_0(x) := q\left(g(x)\right)g'(x) = \frac{g'(x)}{g'(\mu)} f\left(\kappa(x)\right) \qquad \text{where } \kappa(x) := \mu + \frac{g(x)-g(\mu)}{g'(\mu)}.$$

Notice that κ is an increasing function for which $\kappa(x) = x + o(|x-\mu|)$ near μ.

Write $\xi(x)$ for $\sqrt{f(x)}$ and $\gamma(x)$ for $\sqrt{g'(x)/g'(\mu)}$. Then

hell.delta <14>
$$H^2(P,g^{-1}(Q)) = \int \left(\xi(x) - \gamma(x)\xi\left(\kappa(x)\right)\right)^2 dx$$

REMARK. I have the feeling that there should be a neat general bound for the right-hand side of <14>, perhaps something involving $\int \dot{\xi}(x)^2\,dx$. One should be able to bound crudely the contributions from outside some neighborhood U of μ. For the delta-method to work, the derivative $g'(x)$ must stay close to $g'(\mu)$ on the neighborhood. Perhaps a tractable expression involving $(x-\mu)g''(x)$ could be found. That suggests we could hope for a final bound involving something like $\sup_{x\in U}|g''(x)|$ and $\int (x-\mu)^2 f(x)\,dx$. I would start by splitting <14> into a sum of two terms, obtained by adding and subtracting $\xi(x)\gamma(x)$ inside the square.

[§interpolate]

9. Intepolation of increments

The problem solved by Nussbaum (1996) involves the asymptotic equivalence between $\{P_f^n : f \in \mathcal{F}\}$ and the the model $\mathcal{W}_n := \{\mathbb{W}_{n,f} : f \in \mathcal{F}\}$, where $\mathcal{F}$ is a family of smooth density functions on $[0,1]$. That is, the first experiment corresponds to samples of size n from the distribution P_f with density f with respect to Lebesgue measure $\mathfrak{m}$ on $[0,1]$, and $\mathbb{W}_{n,f}$ denotes the probability measure on $C[0,1]$ defined by the white noise process $2\sqrt{n}F_f(t) + W_t$ for $0 \le t \le 1$, where the drift function is defined as

drift.f <15>
$$F_f(t) = \int_0^t \sqrt{f(x)}\,dx \qquad 0 \le t \le 1.$$

The process $\{W_t : 0 \le t \le 1\}$ is a Brownian motion with continuous sample paths. started from $W_0 \equiv 0$.

With Carter's approach, we discretize the observations from P_f by grouping them into m disjoint cells, intervals of length $1/m$, thereby defining a vector of counts with a multinomial distribution, $\mathcal{M}(n,\theta)$, where the vector $\theta := (\theta_1, \ldots, \theta_m)$ actually depend on the underlying density. That is,

theta.f <16>
$$\theta_i := \int \{x \in J_i\} f(x)\,dx \qquad \text{for } i = 1, \ldots, m \qquad \text{where } J_i := \left(\frac{i-1}{n}, \frac{i}{n}\right].$$

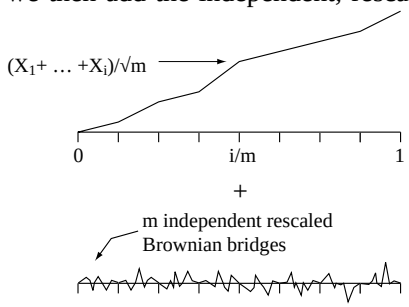
REMARK. Perhaps I should write θ_f to indicate the dependence on f when discussing the application of Carter's general inequality <1> to the Nussbaum problem.

From <16> we have

theta.approx <17>
$$n\theta_i \approx (n/m)\, f(i/m) \qquad \text{for } i = 1, \ldots, m.$$

The measure $\widetilde{\mathbb{N}}_\theta$ corresponds to independent observations $N(\sqrt{n\theta_i}, 1/4)$. To simplify the notation, I will multiply the observations by 2, making $\mathbb{N}_\theta$ correspond to independent random variables $X_i \sim N\left(2\sqrt{n\theta_i}, 1\right)$.

We need a randomization (not depending on f) that will build a process with a distribution close to $\mathbb{W}_{n,f}$, starting from $\{X_i : i = 1, \ldots, m\}$. The obvious method is to interpolate between the partial sums of the X_i's, to build a piecewise linear continuous function with value $(X_1 + \ldots + X_m)/\sqrt{m}$ at i/m, for $i = 0, 1, \ldots, m$. On each linear segment we then add the independent, rescaled Brownian bridges.



In fact, this procedure gives us a white noise with drift. To understand why, it helps to write $X_i = 2\sqrt{n\theta_i} + \xi_i$, where $\xi_1, \ldots, \xi_m$ are independent $N(0,1)$

variables, and express the Brownian bridges as Gaussian process whose covariances are determined by the measures

nui <18> $$\nu_i = \text{uniform distribution on } J_i, \text{ for } i = 1, \ldots, m.$$

Notice that $\mathfrak{m} = \sum_{i \le m} \nu_i/m$. Take B_i as the centered Gaussian process with continuous paths and covariances

BB.cov <19> $$\operatorname{cov}(B_i(s), B_i(t)) = \nu_i[0, s \wedge t] - \nu_i[0, s]\nu_i[0, t] \qquad \text{for } 0 \le s, t \le 1.$$

The interpolated process is then $X(t) := \sum_{i \le m} X_i \nu_i[0, t]/\sqrt{m}$ and the randomization is given by the Gaussian process $\sum_{i \le n} B_i(t)/\sqrt{m}$. That is, we hope to approximate $\mathbb{W}_{n,f}$ by the distribution of the process

$$Z_n(t) := m^{-1/2} \sum\nolimits_{i \le m} \left(\left(2\sqrt{n\theta_i} + \xi_i \right) \nu_i[0, t] + B_i(t) \right).$$

The gaussian process $B_i(t) + \xi_i \nu_i[0, t]$ has covariance $\nu_i[0, s \wedge t]$. The standardized sum of such processes has covariance

$$m^{-1} \sum\nolimits_{i \le m} \nu_i[0, s \wedge t] = \mathfrak{m}[0, s \wedge t] = s \wedge t.$$

That is, the standardized sum is a Brownian motion. The process Z_n has the same distribution as

$$W_t + m^{-1/2} \sum\nolimits_{i \le m} 2\sqrt{n\theta_i}\, \nu_i[0, t] \qquad \text{for } 0 \le t \le 1.$$

The slope of the drift is a step function, taking values

$$m^{-1/2} 2\sqrt{n\theta_i}\, m \approx 2\sqrt{n f(i/m)} \qquad \text{for } x \in J_i$$

The approximation comes from <17>.

Thus we are left with the task of bounding the total variation distance between $\mathbb{W}_{n,\widetilde{f}}$ and $\mathbb{W}_{n,f}$, where $\widetilde{f}$ is a step function that approximates f. In fact, it is not hard to find an explicit expression for $\|\mathbb{W}_{n,\widetilde{f}} - \mathbb{W}_{n,f}\|$ involving the $\mathcal{L}^2(\mathfrak{m})$ distance between the square roots of f and $\widetilde{f}$. By such means, we could calculate a bound on $\delta(\mathcal{N}_{\text{stabil}}, \mathcal{W}_n)$. However, there is a problem.

The method outlined in the preceding Sections is intended to reproduce Nussbaum's result only for the case where f has a bounded derivative that satisfies a Lipschitz condition of order $\alpha - 1$, for some $1 < \alpha \le 2$. (See the next Chapter for refinements to cover $1/2 < \alpha \le 1$.) For such f, the step function approximation is too crude to establish the desired bound for $\delta(\mathcal{N}_{\text{stabil}}, \mathcal{W}_n)$. Instead, we must use an interpolation that corresponds to a smoother approximating $\widetilde{f}$.

As Carter showed, such an improvement is easily achieved. He replaced the uniform distributions by a family of probability distributions with continuous densities with respect to $\mathfrak{m}$, for which it is still true that $\mathfrak{m} = \sum_{i \le m} \nu_i/m$. The new interpolating functions $\nu_i[0, t]$ lead a better approximation for f, by taking advantage of its assumed smoothness. Very elegant.

REMARK. Undoubtedly the improved method corresponds to some simple wavelet fact. I would be pleased to have the connection explained to me.