The Annals of Statistics
2002, Vol. 30, No. 3, 708–730

DEFICIENCY DISTANCE BETWEEN MULTINOMIAL AND MULTIVARIATE NORMAL EXPERIMENTS

BY ANDREW V. CARTER

University of California, Santa Barbara

The deficiency distance between a multinomial and a multivariate normal experiment is bounded under a condition that the parameters are bounded away from zero. This result can be used as a key step in establishing asymptotic normal approximations to nonparametric density estimation experiments. The bound relies on the recursive construction of explicit Markov kernels that can be used to reproduce one experiment from the other. The distance is then bounded using classic local-limit bounds between binomial and normal distributions. Some extensions to other appropriate normal experiments are also presented.

1. Introduction. In Blackwell–Le Cam decision theory, a statistical experiment is a set of probability distributions $\mathcal{P} = \{\mathbb{P}_\theta : \theta \in \Theta\}$ on a measurable space $(\mathcal{X}, \mathcal{A})$. Le Cam defined the deficiency, $\delta(\mathcal{P}, \mathcal{Q})$, to quantify the degree to which $\mathcal{P}$ can be approximated by a simpler experiment $\mathcal{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta\}$ on a different space $(\mathcal{Y}, \mathcal{B})$. If $\delta(\mathcal{P}, \mathcal{Q}) \le \varepsilon$ then every decision procedure in $\mathcal{Q}$ has a corresponding procedure in $\mathcal{P}$ that comes within ε of achieving the same risk for loss functions bounded by 1. The deficiency $\delta(\mathcal{Q}, \mathcal{P})$ is defined analogously, and the distance between the two experiments $\Delta(\mathcal{P}, \mathcal{Q})$ is the larger of the two deficiencies. If there are two sequences of experiments $\{\mathcal{P}_n\}$ and $\{\mathcal{Q}_n\}$ such that $\Delta(\mathcal{P}_n, \mathcal{Q}_n) \to 0$, then the sequences are asymptotically equivalent.

The most direct (and useful) way of bounding $\delta(\mathcal{P}, \mathcal{Q})$ is to propose a Markov kernel $K_x(B)$ from $\mathcal{X}$ to $\mathcal{B}$. A measure P on $(\mathcal{X}, \mathcal{A})$ and a kernel generate a measure on $(\mathcal{Y}, \mathcal{B})$ by

$$(KP)B = \int K_x(B) P(dx).$$

Le Cam (1964) showed that if there exists a Markov kernel K, independent of θ, such that $(K\mathbb{P}_\theta)$ is within ε of $\mathbb{Q}_\theta$ in total-variation distance for all θ, then $\delta(\mathcal{P}, \mathcal{Q}) \le \varepsilon$. Constructing a kernel that provides a connection between two experiments is also illuminating because a decision procedure $\sigma(y)$ in $\mathcal{Q}$ then implies a random decision procedure σK_x in $\mathcal{P}$.

Only recently have examples of deficiencies between nonparametric experiments been published. Brown and Low (1996) proposed a kernel which established a continuous Gaussian approximation to nonparametric regression models

Received December 2000; revised November 2001.

AMS 2000 subject classifications. Primary 62B15; secondary 62G20, 62G07.

Key words and phrases. Deficiency distance, nonparametric experiments, density estimation, multinomial distribution.

with normal errors. Grama and Nussbaum (1998) solved the same problem with nonnormal errors. Milstein and Nussbaum (1998) showed the equivalence of a continuous diffusion process and a nonparametric autoregression model, and Golubev and Nussbaum (1998) demonstrated that stationary Gaussian processes with unknown spectral densities are asymptotically equivalent to a discrete nonparametric regression experiment.

Nussbaum (1996) solved the harder problem of the asymptotic equivalence between a density estimation experiment and a Gaussian white noise experiment. Specifically, the density estimation experiment involved n i.i.d. observations from an unknown density f on [0, 1]. The asymptotically equivalent experiment contains the distributions of continuous Gaussian processes on [0, 1] with drifts that depend on f. Klemelä and Nussbaum (1998) proposed a specific kernel that would bound the distance between these experiments with a stronger restriction on the class of densities. I will propose a technically simpler approach to establishing normal approximations to density estimation experiments.

First, I will assume the class of densities is such that the problem can be reduced to bounding the distance between a multinomial experiment $\mathscr{P}$ and a multivariate normal experiment $\mathscr{Q}$. Conditions for this approximation can be found following a variation on the argument of Brown and Low (1996). In Section 8, a particular example of this sort of bound is given that is sufficient to imply equivalence of the nonparametric experiments from Theorem 1. Earlier arguments in Müller (1979) and Luckhaus and Sauermann (1989) give somewhat different conditions under which nonparametric experiments can be approximated by multinomials.

In experiment $\mathscr{P}$, the $\mathbb{P}_\theta$ distributions are multinomial $\mathscr{M}(n, [\theta_1, \theta_2, \dots, \theta_m])$ where the cell probabilities depend on the density f. Nussbaum (1996) requires the densities to be smooth and bounded away from 0 which translates to the assumptions that $\theta_i \approx \theta_{i+1}$ for neighboring cells and the ratio θ_i/θ_j is bounded. Theorem 1 only uses the boundedness property. [Carter (2001) extends this result by using the smoothness property.] Theorem 1 does not use the set of densities f as the parameter set; instead, every probability vector $(\theta_1, \theta_2, \dots, \theta_m)$ with bounded ratios is included. If the parameter set generated by the set of densities is smaller than Θ_R, the distance cannot be any bigger.

THEOREM 1. *Let $\mathscr{P} = \{\mathbb{P}_\theta : \theta \in \Theta_R\}$, where $\mathbb{P}_\theta = \mathscr{M}(n, \theta)$ and $\Theta_R \subset \mathbb{R}^m$ consists of all vectors of probabilities such that*

$$\frac{\max \theta_i}{\min \theta_i} \le R. \tag{1.1}$$

Let $\mathscr{Q} = \{\mathbb{Q}_\theta : \theta \in \Theta_R\}$ where $\mathbb{Q}_\theta$ is the multivariate normal distribution with the same means and covariances as $\mathbb{P}_\theta$. Then

$$\Delta(\mathscr{P}, \mathscr{Q}) \le C_R \frac{m \log m}{\sqrt{n}} \tag{1.2}$$

for a constant C_R that depends only on R.

This theorem can be used as the key step in reproducing Nussbaum's equivalence when his smoothness parameter α is greater than 1. The results in Carter (2001) can be used to extend the result to the case where $1/2 < \alpha < 1$. This theorem has the advantage that it relies only on the densities being bounded away from zero, and thus can be applied to a variety of density estimation experiments regardless of the underlying probability space. The drawback is that it puts a rather stringent condition on the rate the dimensions can increase, $m = o(\sqrt{n}\log n)$.

The $\mathbb{Q}_\theta$ distributions are not quite what is needed to directly approximate the white noise experiments, but $\mathcal{Q}$ is the most convenient normal experiment. Section 7 bounds the distance between $\mathcal{Q}$ and another multivariate normal experiment, from which a straightforward rescaling will reproduce the increments of a Gaussian process in Section 8.

In the following, generally only the bound on $\delta(\mathcal{P}, \mathcal{Q})$ is described. Section 5 will show how this bound on $\delta(\mathcal{P}, \mathcal{Q})$ can be used to bound $\delta(\mathcal{Q}, \mathcal{P})$ as well.

1.1. *Working with Markov kernels.* I will use a linear-functional notation for integration where

$$\mathbb{P}_\theta f(x) = \int f(x)\mathbb{P}_\theta(dx).$$

A Markov kernel will be written K_x^y where for each x it produces a measure on the $(\mathcal{Y}, \mathcal{B})$ space. The distribution generated by $\mathbb{P}_\theta$ and K_x^y will be written $\mathbb{P}_\theta K_x^y$ because the expectation of the set $B \in \mathcal{B}$ under this distribution is the $\mathbb{P}_\theta$ expectation of the $\mathcal{A}$-measurable function $K_x^y B$, that is,

$$\mathbb{P}_\theta K_x^y B = \mathbb{P}_\theta[K_x^y(B)].$$

Sometimes a superscript will be included on measures, $\mathbb{P}_\theta^x = \mathbb{P}_\theta$, akin to the notation used for kernels.

A Markov kernel can be considered a map between spaces of measures. Specifically, K_x^y is a map from measures on $(\mathcal{X}, \mathcal{A})$ to measures on $(\mathcal{Y}, \mathcal{B})$ where $K_x^y(\mu^x) = \mu^x K_x^y$. I will construct a kernel M_x^y that maps measures $\mathbb{P}_\theta^x$ to measures $\mathbb{P}_\theta^x M_x^y$ that approximate $\mathbb{Q}_\theta^y$.

The symbol K_x^y will also be used to refer to the extensions of the kernel to measures on the product space $(\mathcal{X} \otimes \mathcal{Y}, \sigma(\mathcal{A} \otimes \mathcal{B}))$ with support on the set $\{X = x\}$ such that $K_x^y B = K_x^y\{y : (x, y) \in B\}$. This new kernel maps distributions on $(\mathcal{X}, \mathcal{A})$ to distributions on $(\mathcal{X} \otimes \mathcal{Y}, \sigma(\mathcal{A} \otimes \mathcal{B}))$ such that for $A \in \mathcal{A}$ and $B \in \mathcal{B}$,

$$\mathbb{P}_\theta^x K_x^y(AB) = \mathbb{P}_\theta^x(A[K_x^y B]).$$

2. The structure of the experiments. In bounding $\delta(\mathcal{P}, \mathcal{Q})$, the two main tasks are to construct a Markov kernel and then to bound the total-variation distance between the resulting distributions. It is much easier to bound the distance between experiments on a product space like $\mathbb{R}^m$ if the coordinates are independent. When there is independence, the kernel can be constructed from independent kernels on each coordinate, and then the total-variation distance can be bounded by the inequality $\| \prod P_i - \prod Q_i \| \le [\sum_i P_i \log(dP_i/dQ_i)]^{1/2}$ (see Section B.1 in the Appendix).

Unfortunately, the coordinates of a multinomial experiment are not independent because the coordinates are constrained to sum to n. However the multinomial can be rearranged to produce conditionally independent components.

Let $(X_1, X_2, \dots, X_m)$ be multinomial random variables with probabilities $(\theta_1, \dots, \theta_m)$ for m even. Let $T_j = X_{2j-1} + X_{2j}$. Conditional on this vector of sums $\mathbf{T}$, each pair (X_{2j-1}, X_{2j}) is independent of all the other pairs (X_{2k-1}, X_{2k}) for $j \neq k$.

$\mathbf{T}$ is also multinomially distributed except the dimension is now $m/2$, and the probabilities in this new multinomial are the sums of the original probabilities: $\mathbf{T} \sim \mathcal{M}(n, [\psi_1, \dots, \psi_{m/2}])$ where $\psi_j = \theta_{2j-1} + \theta_{2j}$.

Each pair (X_{2j-1}, X_{2j}) is binomially distributed conditional on $\mathbf{T}$,

$$X_{2j-1} \mid \mathbf{T} \sim \mathrm{Bin}(T_j, p_j) \quad \text{and} \quad X_{2j} = T_j - X_{2j-1},$$

where the conditional probability $p_j = \theta_{2j-1}/(\theta_{2j-1} + \theta_{2j})$.

There is an analogous structure in the normal experiments. Let $Y_1, Y_2, \dots, Y_m$ be multivariate normal random variables with the same means and covariances as the X_i's. Let $S_j = Y_{2j-1} + Y_{2j}$. Conditional on these sums, the pair (Y_{2j-1}, Y_{2j}) is independent of all the other pairs. The sums are multivariate normally distributed with dimension $m/2$,

$$S_j \sim \mathcal{N}(n\psi_j, n\psi_j[1 - \psi_j])$$

for $\psi_j = \theta_{2j-1} + \theta_{2j}$ as before. The covariance structure of the S_j's is the same as for the T_j's.

As before, the pairs (Y_{2j-1}, Y_{2j}) are conditionally independent given $\mathbf{S}$ where $Y_{2j-1} \mid \mathbf{S} \sim \mathcal{N}(S_j p_j, n\psi_j p_j q_j)$ and $Y_{2j} = S_j - Y_{2j-1}$. Notice that the conditional variance of $Y_{2j-1} \mid \mathbf{S}$ is not the same as for the multinomial, $\mathrm{Var}(X_{2j-1} \mid \mathbf{T}) = T_j p_j q_j$, but otherwise the two experiments have the same structure.

Let μ_θ^t and λ_θ^s be the distributions of the statistics $\mathbf{T}$ under $\mathbb{P}_\theta$ and $\mathbf{S}$ under $\mathbb{Q}_\theta$ respectively. The conditional distributions $\mathbb{P}_\theta\{\cdot|\mathbf{T}\}$ and $\mathbb{Q}_\theta\{\cdot|\mathbf{S}\}$ are regular and therefore have versions P_t^x and Q_s^y which are Markov kernels such that $\mathbb{P}_\theta = \mu_\theta^t P_t^x$ and $\mathbb{Q}_\theta = \lambda_\theta^s Q_s^y$.

If there are kernels K_t^s and $L_{s,t,x}^y$ such that $\mu_\theta^t K_t^s$ approximates λ_θ^s and $P_t^x L_{s,t,x}^y$ approximates Q_s^y for each pair (s, t), then a kernel M_x^y can be constructed that maps distributions $\mathbb{P}_\theta$ to distributions that approximate $\mathbb{Q}_\theta$.

LEMMA 1. *If the kernel K_t^s is such that*

$$\|\lambda_\theta^s - \mu_\theta^t K_t^s\| \le \varepsilon,$$

and the kernel $L^y_{s,t,x}$ is such that for each pair (s,t),

$$\|Q^y_s - P^x_t L^y_{s,t,x}\| \le \rho(s,t),$$

then there exists a kernel M^y_x such that

$$\|\mathbb{Q}_\theta - \mathbb{P}_\theta M^y_x\| \le \varepsilon + \mu^t_\theta K^s_t[\rho(s,t)].$$

The kernel M^y_x is the composition of the kernels K^s_t and $L^y_{s,t,x}$. The bound follows from some manipulation of the measures and applications of the triangle inequality for the total-variation distance. The proof is in Section A of the Appendix.

Thus, the task of bounding the distance between multinomials and multivariate normals is reduced to bounding the distance between independent binomials and normals ($\|Q^y_s - P^x_t L^y_{s,t,x}\|$, see Section 3) and the distance between a multinomial and multivariate normal with a smaller dimension ($\|\lambda^s_\theta - \mu^t_\theta K^s_t\|$). The distance between these new multinomials μ^t_θ and multivariate normals λ^s_θ can be bounded using the same strategy of conditioning on pairwise sums to create another set of independent binomials and normals and another, even smaller set of multinomials and multivariate normals. The condition (1.1) from Theorem 1 still applies to the smaller multinomial experiment,

$$\frac{\max \psi_j}{\min \psi_j} \le \frac{2\max\theta_i}{2\min\theta_i} \le R.$$

Thus, Section 4 uses induction on the dimension m to bound the distance between the experiments $\mathcal{P}$ and $\mathcal{Q}$.

3. The distance between independent binomials and normals.

The bound between the multinomial and normal experiments depends on bounding the distance between the conditional distributions P^x_t and Q^y_s. Because the P^x_t has support on the integers, the total-variation distance $\|P^x_t - Q^y_s\| = 1$. To counter this, a random perturbation is added to each of the coordinates of the P^x_t distribution. The kernel $L^y_{s,t,x}$, for a particular value of the vectors s, t, and x, is the distribution of $Y_{2j-1} = x_{2j-1} + U_j$ and $Y_{2j} = s_j - Y_{2j-1}$ for $j = 1, 2, \dots, m/2$ where the U_j are independent uniforms on $[-1/2, 1/2]$. Therefore the distribution generated by P^x_t and $L^y_{s,t,x}$ is absolutely continuous with respect to Q^y_s.

The distributions $P^x_t L^y_{s,t,x}$ and Q^y_s both have independent pairs of coordinates that sum to $\mathbf{S}$. Thus the total-variation between $P^x_t L^y_{s,t,x}$ and Q^y_s is less than the distance between the marginal distributions of the odd coordinates because the transformation $Y_{2j} = S_j - Y_{2j-1}$ reproduces the even coordinates, and a transformation applied to both distributions cannot increase the total-variation distance between those distributions. In general, for two measures μ_1 and μ_2 on $(\mathcal{X}, \mathcal{A})$ and a Markov kernel K^y_x from $\mathcal{X}$ to $(\mathcal{Y}, \mathcal{B})$,

$$\|\mu_1 K^y_x - \mu_2 K^y_x\| \le \|\mu_1 - \mu_2\|, \tag{3.1}$$

because the kernel K_x^y maps $\mathcal{B}$-measurable test functions into $\mathcal{A}$-measurable test functions.

As an intermediate step in bounding the distance between the odd coordinates of $P_t^x L_{s,t,x}^y$ and Q_s^y, let $\mathbb{Q}_{\mathbf{p}}$ be a multivariate normal distribution with $m/2$ independent $\mathcal{N}(t_j p_j, t_j p_j q_j)$ coordinates. The total-variation distance between $\mathbb{Q}_{\mathbf{p}}$ and the joint distribution of $m/2$ independent $\mathcal{N}(s_j p_j, n\psi_j p_j q_j)$ is less than

$$\left[\sum_{j=1}^{m/2} \frac{p_j}{q_j} \frac{(t_j - s_j)^2}{2n\psi_j} + \frac{3}{2}\left(1 - \frac{t_j}{n\psi_j}\right)^2\right]^{1/2} \tag{3.2}$$

using (C.1) in the Appendix.

Likewise, let $\mathbb{P}_{\mathbf{p}}$ be the distribution of $m/2$ independent binomial(t_j, p_j) coordinates and let $\mathbb{P}_{\mathbf{p}} \star \mathbf{U}$, be the result of convolving $\mathbb{P}_{\mathbf{p}}$ with independent uniform $[-1/2, 1/2]$ distributions. The total-variation distance between $\mathbb{Q}_{\mathbf{p}}$ and $\mathbb{P}_{\mathbf{p}} \star \mathbf{U}$ can be bounded using local-limit techniques of Prohorov (1961) or Feller (1968), pages 168–170.

LEMMA 2.

$$\|\mathbb{Q}_{\mathbf{p}} - \mathbb{P}_{\mathbf{p}} \star \mathbf{U}\| \le \left[\sum_{j=1}^{m/2} \frac{C}{t_j p_j q_j}\right]^{1/2}.$$

The proof appears in the Appendix, Section B.

Therefore, by (3.1) and the triangle inequality,

$$\begin{aligned} \|P_t^x L_{s,t,x}^y - Q_s^y\| &\le \left[\sum_{j=1}^{m/2} \frac{C}{t_j p_j q_j}\right]^{1/2} \\ &\quad + \left[\sum_{j=1}^{m/2} \frac{p_j}{q_j} \frac{(t_j - s_j)^2}{2n\psi_j} + \frac{3}{2}\left(1 - \frac{t_j}{n\psi_j}\right)^2\right]^{1/2} \end{aligned} \tag{3.3}$$

from Lemma 2 and (3.2).

4. Induction on the dimension of the observations.

The bound on $\delta(\mathcal{P}, \mathcal{Q})$ is derived using an inductive argument that proves the result for m-dimensional experiments from a bound on the distance between $m/2$-dimensional experiments. This assumes that m is even, but some minor alterations can take care of the extra observation. Besides, when the argument is being made for density-estimation experiments, we are free to choose the dimension of the experiments as $m = 2^k$ for some integer k.

LEMMA 3 (Inductive step). *For $m > 1$, if*

$$\sup_{\Theta_R} \|\mu_\theta^t K_t^s - \lambda_\theta^s\| \le C_R \frac{(m/2)\log(m/2)}{\sqrt{n}} \tag{4.1}$$

and the kernel K_t^s is such that $K_t^s\{|S_j - T_j| \le (\log_2 m)/2\} = 1$ for all j and t, then there exists a kernel M_x^y such that

$$\sup_{\Theta_R} \|\mathbb{P}_\theta M_x^y - \mathbb{Q}_\theta\| \le C_R \frac{m \log m}{\sqrt{n}} \tag{4.2}$$

and the kernel M_x^y is such that $M_x^y\{|X_i - Y_i| \le (\log_2 2m)/2\} = 1$ for all i and x.

PROOF. By Lemma 1, for each θ the distance is bounded by

$$\|\mathbb{P}_\theta M_x^y - \mathbb{Q}_\theta\| \le \|\mu_\theta^t K_t^s - \lambda_\theta^s\| + \mu_\theta^t K_t^s \|P_t^x L_{s,t,x}^y - Q_s^y\|, \tag{4.3}$$

where the first term is bounded by assumption (4.1).

The first step in bounding the expectation $\mu_\theta^t K_t^s \|P_t^x L_{s,t,x}^y - Q_s^y\|$ is to bound the expectation on the set $A^c = \bigcup_j \{T_j \le n\psi_j/2\}$. The total-variation distance is bounded by 1 and $\mu_\theta^t A^c \le m \exp[-Cn\psi_j]$, which is much smaller than the other terms when $\psi_j > R/(2m)$ and $m < \sqrt{n}$. Therefore, implicitly the rest of the calculations will assume that all the T_j's are larger than $n\psi_j/2$.

Section 3 bounded the second term by

$$\begin{aligned} \mu_\theta^t K_t^s \|P_t^x L_{s,t,x}^y - Q_s^y\| &\le \mu_\theta^t K_t^s \left[\sum_{j=1}^{m/2} \frac{C}{t_j p_j q_j}\right]^{1/2} \\ &\quad + \mu_\theta^t K_t^s \left[\sum_{j=1}^{m/2} \frac{3}{2}\left(1 - \frac{t_j}{n\psi_j}\right)^2 + \frac{p_j}{q_j}\frac{(s_j - t_j)^2}{2n\psi_j}\right]^{1/2}. \end{aligned}$$

By Jensen's inequality, the measure can be moved inside the square roots:

$$\begin{aligned} \mu_\theta^t K_t^s \|P_t^x L_{s,t,x}^y - Q_s^y\| &\le \left[\sum_{j=1}^{m/2} \mu_\theta^t \frac{C}{t_j p_j q_j}\right]^{1/2} \\ &\quad + \left[\sum_{j=1}^{m/2} \frac{3}{2}\mu_\theta^t\left(1 - \frac{t_j}{n\psi_j}\right)^2 + \mu_\theta^t K_s^t \frac{p_j}{q_j}\frac{(s_j - t_j)^2}{2n\psi_j}\right]^{1/2}. \end{aligned}$$

The terms that do not depend on the $\mathbf{S}$ are binomial(n, ψ_j) expectations,

$$\mu_\theta^t \left(1 - \frac{T_j}{n\psi_j}\right)^2 \left\{T_j > \frac{n\psi_j}{2}\right\} \le \frac{1 - \psi_j}{n\psi_j} \quad \text{and} \quad \mu_\theta^t T_j^{-1} \left\{T_j > \frac{n\psi_j}{2}\right\} \le \frac{2}{n\psi_j}$$

[Johnson and Kotz (1969), pages 73–75]. By assumption, the kernel K_t^s is such that $(S_j - T_j)^2 \leq (\log m)^2/4$, so

$$K_t^s \frac{p_j}{q_j} \frac{(S_j - T_j)^2}{n\psi_j} \leq \frac{p_j}{q_j} \frac{\log^2 m}{4n\psi_j}.$$

The restrictions on the parameter space Θ_R from Theorem 1 bound the most extreme cases in each sum:

$$\frac{p_j}{q_j} \leq R \quad \text{and} \quad \frac{1}{p_j q_j} = \frac{(\theta_{2j} + \theta_{2j-1})^2}{\theta_{2j}\theta_{2j-1}} \leq 4R^2 \quad \text{and} \quad \frac{1}{\psi_j} \leq \frac{R}{\max \psi_j} \leq \frac{Rm}{2}.$$

Thus, the sums are bounded by

$$\sum_j \frac{2C}{p_j q_j n\psi_j} \leq \frac{8R^2C}{n} \sum_j \frac{1}{\psi_j} \leq 8R^3C \frac{m^2}{4n},$$

$$\frac{3}{2} \sum_j \frac{1 - \psi_j}{n\psi_j} \leq \frac{1}{n} \sum_j \frac{1}{\psi_j} \leq \frac{3R}{2} \frac{m^2}{4n}$$

and

$$\sum_j \frac{p_j}{q_j} \frac{\log^2 m}{4n\psi_j} \leq \frac{R \log^2 m}{4n} \sum_j \frac{1}{\psi_j} \leq \frac{R^2}{4} \frac{m^2 \log^2 m}{4n}.$$

The total contribution from the conditional experiments is therefore bounded by

$$\mu_\theta^t K_t^s \| P_t^x L_x s, t, x^y - Q_s^y \| \leq \frac{m \log m}{2\sqrt{n}} \Big[\sqrt{8R^3C} + \sqrt{3R/2 + R^2/4} \Big].$$

Choose $C_R > \sqrt{8R^3C} + \sqrt{3R/2 + R^2/4}$.

Then plugging back into (4.3) produces the bound

$$\| \mathbb{P}_\theta M_x^y - \mathbb{Q}_\theta \| \leq \frac{C_R m \log(m/2)}{2\sqrt{n}} + \frac{C_R m \log m}{2\sqrt{n}} < \frac{C_R m \log m}{\sqrt{n}}$$

for all $\theta \in \Theta_R$ thus establishing (4.2).

To show that the kernel $M_x^y = K_s^t L_{s,t,x}^y$ fulfills the condition that $M_x^y\{|X_i - Y_i| \leq (\log_2 2m)/2\} = 1$ for all x_i, there are two cases to consider. If i is odd, then

$$|X_{2j-1} - Y_{2j-1}| \sim \text{uniform}[-1/2, 1/2]$$

and clearly $M_x^y\{|X_{2j-1} - Y_{2j-1}| \leq 1/2\} = 1$. If i is even, then

$$|X_{2j} - Y_{2j}| \leq |X_{2j-1} - Y_{2j-1}| + |T_j - S_j|$$

and, by assumption, $|T_j - S_j| \leq (\log_2 m)/2$ almost surely. Therefore,

$$|X_{2j} - Y_{2j}| \leq \frac{1}{2} + \frac{\log_2 m}{2} = \frac{\log_2 2m}{2}$$

almost surely for all i and every x. $\square$

5. A bound on $\delta(\mathcal{Q},\mathcal{P})$. The deficiency in the other direction $\delta(\mathcal{Q},\mathcal{P})$ can be bounded using the bound on the total-variation distance between $\mathbb{P}_\theta M_x^y$ and $\mathbb{Q}_\theta$. The key is to choose a kernel $\bar{M}_y^x$ that inverts the effects of the kernel M_x^y. If $\bar{M}_y^x$ is such that $\mathbb{P}_\theta M_x^y \bar{M}_y^x = \mathbb{P}_\theta$ for all θ then

$$\begin{aligned}\|\mathbb{P}_\theta - \mathbb{Q}_\theta \bar{M}_y^x\| &= \|\mathbb{P}_\theta M_x^y \bar{M}_y^x - \mathbb{Q}_\theta \bar{M}_y^x\| \\ &\le \|\mathbb{P}_\theta M_x^y - \mathbb{Q}_\theta\|\end{aligned}$$

by inequality (3.1). The kernel $\bar{M}_y^x$ will be constructed recursively just like the construction of M_x^y except the adding of the uniform perturbations will be inverted by rounding off to the nearest integer.

To start, consider the case where the dimension $m = 2$. The original kernel produced $Y_1 = X_1 + U$ where $U \sim \text{uniform}[-1/2, 1/2]$. The kernel $\bar{M}_y^x$ rounds Y_1 off to the nearest integer to produce X_1 and X_2 is still $n - X_1$.

To construct $\bar{M}_y^x$ for an m-dimensional distribution, first assume there exists a kernel $\bar{K}_s^t$ such that for all $m/2$-dimensional multinomials μ_θ^t,

$$\mu_\theta^t K_t^s \bar{K}_s^t = \mu_\theta^t.$$

Then, the kernel $\bar{L}_{s,t,y}^x$ rounds off the Y_{2j-1} to produce X_{2j-1} and $X_{2j} = T_j - X_{2j-1}$. As before, the whole kernel is $\bar{M}_y^x = \bar{K}_s^t \bar{L}_{s,t,y}^x$.

To see that $\bar{M}_y^x$ inverts M_x^y, let T_j^* be a random variable distributed $\mu_\theta^t K_t^s \bar{K}_s^t$ and let X_i^* be distributed $\mathbb{P}_\theta M_x^y \bar{M}_y^x$. The odd coordinates are $X_{2j-1}^* = \text{round}[Y_{2j-1}]$ where $Y_{2j-1} = X_{2j-1} + U_j$ for $U_j \sim \text{uniform}[-1/2, 1/2]$, so that $X_{2j-1}^* = \text{round}[X_{2j-1} + U_j] = X_{2j-1}$. Furthermore, $X_{2j}^* = T_j^* - X_{2j-1}^*$ which has the same distribution as $X_{2j} = T_j - X_{2j-1}$ because T_j^* has the same distribution as T_j and $X_{2j-1}^* = X_{2j-1}$. Therefore, by induction,

$$\mathbb{P}_\theta M_x^y \bar{M}_y^x = \mathbb{P}_\theta, \tag{5.1}$$

and $\|\mathbb{P}_\theta - \mathbb{Q}_\theta \bar{M}_y^x\| \le C_R n^{-1/2} m \log m$.

6. Proof of Theorem 1. The proof is by induction on the dimension of the experiments. The kernel K_t^s is just the kernel M_x^y of the experiment for dimension $m/2$. All that is necessary beyond Lemma 3 is to establish a starting point for the induction, $m = 1$.

When $m = 1$, all the distributions in both experiments put probability 1 on the point n. These are noninformative experiments because the distributions do not depend on θ, and the distance between them is 0 [cf. Torgersen (1991), pages 225 and 226, Example 6.2.1]. The kernel K_s^t is the identity, thus $|S_1 - T_1| = 0$ almost surely K_s^t. Therefore, if $m = 2$ then the assumptions in Lemma 3 hold, and Theorem 1 is established by induction. □

7. Some other normal experiments. The normal experiment $\mathcal{Q}$ is convenient because it has the same moments as the multinomial and thus approximates $\mathcal{P}$ well. However, if the multivariate normal distributions are to represent the increments of a continuous Gaussian process then the normal coordinates must be independent and have a variance that does not depend on θ. To this end, the experiment $\mathcal{Q}$ is approximated by the experiment $\widetilde{\mathcal{Q}}$ with distributions $\widetilde{\mathbb{Q}}_\theta$ which are products of m independent $\mathcal{N}(n\theta_i, n\theta_i)$ distributions, and, furthermore, $\widetilde{\mathcal{Q}}$ is approximated by $\mathcal{Q}^*$ with $\mathbb{Q}_\theta^*$ distributions having independent $\mathcal{N}(\sqrt{n\theta_i}, 1/4)$ coordinates. I will show that $\Delta(\mathcal{Q}, \widetilde{\mathcal{Q}}) \le C_R n^{-1/2} m \log m$ and $\Delta(\widetilde{\mathcal{Q}}, \mathcal{Q}^*) \le C_R n^{-1/2} m \log m$ which implies, along with Theorem 1, that $\Delta(\mathcal{P}, \mathcal{Q}^*) \le 3C_R n^{-1/2} m \log m$.

7.1. *Approximation by an experiment with independent coordinates.* The technique for approximating the $\mathbb{Q}_\theta$ by distributions with independent coordinates is similar to the Poissonization technique that is used in density estimation, in that it randomizes the total to produce independence among the subtotals. It is easier to establish this approximation between normal experiments, as opposed to multinomials and Poissons, because working with real numbers rather than integers allows more flexibility in the choice of transformations.

The distributions $\widetilde{\mathbb{Q}}_\theta$ with independent $\mathcal{N}(n\theta_i, n\theta_i)$ distributions are related to the $\mathbb{Q}_\theta$ distributions in that $\mathbb{Q}_\theta = \widetilde{\mathbb{Q}}_\theta\{\cdot | \sum Y_i = n\}$. Thus both $\mathbb{Q}_\theta$ and $\widetilde{\mathbb{Q}}_\theta$ have the same conditional distributions given the vector of pairwise sums $\mathbf{S}$, even though under the $\widetilde{\mathbb{Q}}_\theta$ distributions the sums S_j are distributed independently $S_j \sim \mathcal{N}(n\psi_j, n\psi_j)$.

Given these similarities between the conditional structures of $\mathcal{Q}$ and $\widetilde{\mathcal{Q}}$, an induction argument similar to the proof of Theorem 1 is used to construct the kernel and to bound $\delta(\mathcal{Q}, \widetilde{\mathcal{Q}})$. Let $\tilde{\mathbf{S}}$ be the sums of the pairs in the $\widetilde{\mathbb{Q}}_\theta$ distributions. Let $\tilde{\lambda}_\theta^{\tilde{s}}$ be the distribution of $\tilde{\mathbf{S}}$ and let the Markov kernel $\widetilde{Q}_{\tilde{s}}^{\tilde{y}}$ be a version of the conditional distribution of $\widetilde{\mathbb{Q}}_\theta$ given $\tilde{\mathbf{S}}$.

First, consider the case where $m = 1$. These experiments are trivial but they are the starting point for the induction argument. The parameter is restricted to $\theta_1 = 1$, so all the distributions in $\mathcal{Q}$ put probability 1 on n and all the distributions in $\widetilde{\mathcal{Q}}$ are $\mathcal{N}(n, n)$. These experiments are noninformative and the distance between them is 0 [Torgersen (1991), pages 225 and 226]. The Markov kernel $\tilde{K}_n^{\tilde{n}}$ that maps from $\mathbb{Q}_\theta$ to $\widetilde{\mathbb{Q}}_\theta$ for $m = 1$ produces an independent $\tilde{n} \sim \mathcal{N}(n, n)$.

The induction step starts with the assumption that there is a kernel $\tilde{K}_s^{\tilde{s}}$ such that

$$\sup_{\Theta_R} \| \tilde{\lambda}_\theta^{\tilde{s}} - \lambda_\theta^s \tilde{K}_s^{\tilde{s}} \| \le C_R \sqrt{\frac{m/2}{n}}$$

and that

$$\tilde{K}_s^{\tilde{s}} \left\{ \frac{\tilde{s}_j}{s_j} = \frac{\tilde{n}}{n} \right\} = 1 \qquad \text{for all } s \text{ and } j = 1, \dots, m/2,$$

where $\tilde{n}$ is the sum of all the $\tilde{s}_j$. The kernel $\tilde{K}_n^{\tilde{n}}$ meets these conditions for $m = 1$.

The induction proceeds by the use of Lemma 1,

$$\|\widetilde{\mathbb{Q}}_\theta - \mathbb{Q}_\theta \widetilde{M}_y^{\tilde{y}}\| \leq \|\tilde{\lambda}_\theta^{\tilde{s}} - \lambda_\theta^s \tilde{K}_s^{\tilde{s}}\| + \lambda_\theta^s \tilde{K}_s^{\tilde{s}} \|\widetilde{Q}_{\tilde{s}}^{\tilde{y}} - Q_s^y \tilde{L}_{\tilde{s},s,y}^{\tilde{y}}\|.$$

The $\widetilde{Q}_{\tilde{s}}^{\tilde{y}}$ distribution has independent pairs of coordinates $(\tilde{Y}_{2j-1}, \tilde{Y}_{2j})$ where $\tilde{Y}_{2j-1} \sim \mathcal{N}(\tilde{s}_j p_j, n\psi_j p_j q_j)$. Thus the kernels $\widetilde{Q}_{\tilde{s}}^{\tilde{y}}$ and Q_s^y are the same, but the bound depends on the distance between the distributions for $\tilde{s} \neq s$. The kernel $\tilde{L}_{\tilde{s},s,y}^{\tilde{y}}$ is a nonrandom kernel that corresponds to rescaling each coordinate by the factor $\tilde{s}/s$, that is, $\tilde{L}_{\tilde{s},s,y}^{\tilde{y}}\{\tilde{Y}_i = Y_i(\tilde{S}_{\lceil i/2 \rceil}/S_{\lceil i/2 \rceil})\} = 1$. The kernel preserves the property that $\tilde{Y}_{2j} + \tilde{Y}_{2j-1} = \tilde{S}_j$ so by the same argument as in Section 3 it is only necessary to bound the total variation distance between the odd coordinates. Q_s^y and $\tilde{L}_{\tilde{s},s,y}^{\tilde{y}}$ generate distributions such that the odd coordinates are distributed $\mathcal{N}(\tilde{s}_j p_j, (\tilde{s}_j/s_j)^2 n\psi_j p_j q_j)$. Therefore, the distance between the conditional distributions is bounded using the inequalities in the Appendix, Section C, by

$$\|\widetilde{Q}_{\tilde{s}}^{\tilde{y}} - Q_s^y \tilde{L}_{\tilde{s},s,y}^{\tilde{y}}\| \leq \left[\sum_{j=1}^{m/2} \left(1 - \frac{\tilde{s}_j^2}{s_j^2}\right)^2\right]^{1/2}.$$

By the assumption on the kernel $\tilde{K}_s^{\tilde{s}}$ and Jensen's inequality,

$$\lambda_\theta^s \tilde{K}_s^{\tilde{s}} \left[\sum_{j=1}^{m/2} \left(1 - \frac{\tilde{s}_j^2}{s_j^2}\right)^2\right]^{1/2} \leq \left[\sum_{j=1}^{m/2} \tilde{K}_n^{\tilde{n}} \left(1 - \frac{\tilde{n}^2}{n^2}\right)^2\right]^{1/2} \leq \sqrt{\frac{Cm}{n}}$$

for a constant $C > 4$. Combined with the induction assumption on the size of $\|\tilde{\lambda}_\theta^{\tilde{s}} - \lambda_\theta^s K_s^{\tilde{s}}\|$, this establishes the bound

$$\|\widetilde{\mathbb{Q}}_\theta - \mathbb{Q}_\theta \widetilde{M}_y^{\tilde{y}}\| \leq C_R \sqrt{\frac{m/2}{n}} + \sqrt{\frac{Cm}{n}} \leq C_R \sqrt{\frac{m}{n}},$$

as long as $C_R > C(1 - 1/\sqrt{2})^{-2}$. Furthermore, the kernel $\widetilde{M}_y^{\tilde{y}}$ is such that

$$\frac{\tilde{Y}_i}{Y_i} = \frac{\tilde{S}_{\lceil i/2 \rceil}}{S_{\lceil i/2 \rceil}} = \frac{\tilde{n}}{n} \qquad \text{almost surely.}$$

Thus the kernel $\widetilde{M}_y^{\tilde{y}}$ can be used as the kernel $K_s^{\tilde{s}}$ in establishing the result for the larger, $2m$-dimensional experiments.

7.2. *A variance stabilizing transformation.* The $\widetilde{\mathcal{Q}}$ experiment with m independent observations $\mathcal{N}(n\theta_j, n\theta_j)$ is a problem because the variance depends on the parameter. The solution is to apply a variance stabilizing transformation as the kernel to produce a multivariate normal location experiment $\mathcal{Q}^*$. The coordinates of both $\widetilde{\mathcal{Q}}$ and $\mathcal{Q}^*$ are independent so the kernel can be constructed from independent transformations of each coordinate and an inductive argument is not necessary.

The transformed normals are no longer normally distributed, but for large n they admit normal approximations. Klemelä and Nussbaum [(1998), pages 18–20] describe a similar method of getting a bound on the deficiency between the transformed Y and its normal approximation.

Each coordinate $\tilde{Y}_i$ is transformed by taking the square root of the positive part of $\tilde{Y}_i$. Negative values of $\tilde{Y}_i$ are unlikely so they do not contribute much to the distance. The density of $\sqrt{(\tilde{Y}_i)_+}$ is

$$f(y) = \frac{2}{\sqrt{2\pi}} \exp\left[-\frac{(y^2 - n\theta_i)^2}{2n\theta_i}\right] \frac{y}{\sqrt{n\theta_i}} \qquad \text{for } y > 0,$$

and there is a point mass at $\{\tilde{Y}_i = 0\}$ with small probability. The marginal density of $\mathbb{Q}_\theta^* \sim \mathcal{N}(\sqrt{n\theta_i}, 1/4)$ is

$$g(y) = \frac{2}{\sqrt{2\pi}} \exp[-2(y - \sqrt{n\theta_i})^2].$$

Let the set A be a symmetric set around $\sqrt{n\theta_i}$ with high probability under $\mathbb{Q}_\theta^*$,

$$A = \{y : |y - \sqrt{n\theta_i}| \le \sqrt{n\theta_i}/2\}.$$

To use the inequality (B.1), it is necessary to bound

$$\mathbb{Q}_\theta^* A \log \frac{d\mathbb{Q}_\theta^*}{d\widetilde{\mathbb{Q}}_\theta} = -\mathbb{Q}_\theta^* A\left[\log(y/\sqrt{n\theta_i}) + 2(y - \sqrt{n\theta_i})^2 - \frac{(y^2 - n\theta_i)^2}{2n\theta_i}\right].$$

Let $\xi = 2(y - \sqrt{n\theta_i})$. The first term can be bounded using the Taylor expansion around $\xi = 0$,

$$\log\left(\frac{y}{\sqrt{n\theta_i}}\right) = \log\left(1 + \frac{\xi}{2\sqrt{n\theta_i}}\right) \ge \frac{\xi}{2\sqrt{n\theta_i}} - \frac{\xi^2}{8n\theta_i} + \frac{\xi^3}{24(n\theta_i)^{3/2}} - \frac{\xi^4}{32(n\theta_i)^2}$$

for $|\xi| \le \sqrt{n\theta_i}$. Thus

$$-\mathbb{Q}_\theta^* A \log(y/\sqrt{n\theta_i}) \le \frac{1}{8n\theta_i} + \frac{3}{32(n\theta_i)^2}.$$

Substituting ξ into the other terms yields

$$\frac{(y^2 - n\theta_i)^2}{2n\theta_i} - 2(y - \sqrt{n\theta_i})^2 = \frac{\xi^3}{4\sqrt{n\theta_i}} + \frac{\xi^4}{32n\theta_i},$$

which has $\mathbb{Q}_\theta^*$ expectation

$$\mathbb{Q}_\theta^* A\left(\frac{\xi^3}{4\sqrt{n\theta_i}} + \frac{\xi^4}{32n\theta_i}\right) \le \frac{3}{32n\theta_i}.$$

So the total-variation distance between $\mathbb{Q}_\theta^*$ and the image of the square root under $\widetilde{\mathbb{Q}}_\theta$ is bounded using (B.1) by

$$\left[2\mathbb{Q}_\theta^* A^c + \sum_i \frac{7}{32n\theta_i} + \frac{3}{32(n\theta_i)^2}\right]^{1/2}.$$

Using standard tail bounds $\mathbb{Q}_\theta^* A^c \le \exp[-n/m]$, and, as in the proof of Lemma 3, $\sum(n\theta_i)^{-1} \le Rm^2/n$. Therefore, the total-variation distance is less than $C_R mn^{-1/2}$.

8. Applying Theorem 1 to nonparametric experiments. The intent of bounding the distance between multinomials and multivariate normals is to make assertions about density estimation experiments. The multinomial experiment can be seen as the result of grouping independent observations from a continuous density into subsets. Likewise, the normals are approximately the increments of a continuous Gaussian process.

The bound between $\mathcal{P}$ and $\mathcal{Q}^*$ does not depend on the specific sample space of the original density estimation experiment. The properties of the sample space only enter into the problem when approximating the density estimation experiment by the multinomial experiment $\mathcal{P}$ and the related problem of approximating the continuous Gaussian experiment by its increments.

Typically, a smoothness condition on the densities is sufficient to show asymptotic equivalence, as long as m grows sufficiently fast with n. Brown and Zhang (1998) showed that a smoothness condition on the densities is necessary for the asymptotic equivalence of the density estimation experiment and the normal experiment.

A class of smooth, differentiable densities $f \in \mathcal{F}(\gamma, \varepsilon, M)$ on the interval [0, 1] such that $\varepsilon < f < M$ and

$$|f'(x) - f'(y)| \le M|x-y|^\gamma \qquad \text{for all } (x, y)$$

provides an example to which Theorem 1 can be shown to apply for $\gamma, \varepsilon > 0$. This is a subset of a Hölder ball with exponent $\alpha > 1$ as in Klemelä and Nussbaum (1998). These densities generate probabilities

$$\theta_i = \int_{(i-1)/m}^{i/m} f,$$

which are between ε/m and M/m so that

$$\frac{\max \theta_i}{\min \theta_i} \le \frac{M}{\varepsilon} = R.$$

Thus the probabilities generated by $\mathcal{F}(\gamma, \varepsilon, M)$ are a subset of Θ_R.

8.1. *Density estimation.* Let $\bar{\mathcal{P}}$ be the density estimation experiment which observes n independent observations from the distribution P_f with density f. This $\bar{\mathcal{P}}$ can be approximated by the multinomial $\mathcal{P}_m$. The multinomial observations are just the counts within the intervals $[(i-1)/m, i/m]$ so that $\delta(\bar{\mathcal{P}}, \mathcal{P}_m) = 0$. The difficulty is in generating the original n observations from the m counts on each subinterval.

Let X^* be a single observation from the discrete distribution that puts probability θ_i at the midpoint x_i^* of the subintervals, $x_i^* = (2i-1)/(2m)$. Add to X^* an independent V^* with density

$$\frac{dP_{V^*}}{d\lambda}(x) = m - m^2|x| \qquad \text{for } -\frac{1}{m} \le x \le \frac{1}{m}.$$

Then the density of the random variable $X^* + V^*$ is the convolution of the discrete distribution with these triangles or simply a linear interpolation between the values at the midpoints, $f^*(i) = m\theta_i$,

$$\hat{f}(x) = f^*(i)[x - x_{i+1}^*] - f^*(i+1)[x - x_i^*] \qquad \text{for } x_i^* \le x \le x_{i+1}^*,\ 1 \le i \le m-1.$$

To avoid putting any probability outside the interval [0, 1], if $X^* = \frac{1}{2m}$ or $X^* = m - \frac{1}{2m}$ and $X^* + V^* \notin [0, 1]$ then reflect the value back into the interval: $|X^* + V^*|$ for $i = 1$, or $1 - |(X^* + V^*) - 1|$ for $i = m$. Thus, near the edges, $\hat{f}(x)$ is a constant equal to $f^*(1)$ for $x < 1/m$ or $f^*(m)$ for $x > (m-1)/m$.

The multinomial experiment $\mathcal{P}_m$ is a sufficient statistic for n copies of the discrete distribution of X^*. Thus adding n independent V^*'s as above describes a randomization which produces n independent copies of $\hat{f}$. This approximates $\bar{\mathcal{P}}$ with an error less than

$$H(P_f^n, P_{\hat{f}}^n) \le \sqrt{n} H(f, \hat{f}) \le \frac{\sqrt{n}}{\sqrt{\varepsilon}} \|f - \hat{f}\|_2. \tag{8.1}$$

A function approximation bound is necessary.

LEMMA 4. *For* $f \in \mathcal{F}(\gamma, M, \varepsilon)$, *and* $\hat{f}$ *defined above, the squared* L_2 *distance is bounded by*

$$\sup_{f \in \mathcal{F}} \|f - \hat{f}\|_2^2 \le M m^{-2\gamma-2}.$$

for $0 < \gamma \le 1/2$.

Therefore, the distance is bounded by

$$\Delta(\bar{\mathcal{P}}, \mathcal{P}_m) \le C m^{-\gamma-1} \sqrt{n}, \tag{8.2}$$

where C is a constant that depends on ε and M.

PROOF. The smoothness condition on $\mathcal{F}$ can be used to bound the error in a Taylor expansion,

$$\begin{aligned} |f(t+\delta) - f(t) - \delta f'(t)| &= |f(t) + \delta f'(t^*) - f(t) - \delta f'(t)| \\ &\le M\delta^{\gamma+1} \end{aligned} \tag{8.3}$$

because there exists a t^* such that $|t^* - t| \le \delta$ by the mean value theorem.

This implies a bound on $|f(x_i^*) - \hat{f}(x_i^*)|$,

$$\begin{aligned} |f(x_i^*) - \hat{f}(x_i^*)| &= |f(x_i^*) - f^*(i)| \\ &= \left| f(x_i^*) - m\int_{(i-1)/m}^{i/m} [f(x_i^*) + (x - x_i^*)f'(x_i^*) + \xi(x)]\,dx \right| \\ &= \left| m \int_{(i-1)/m}^{i/m} \xi(x)\,dx \right| \\ &\le Mm^{-\gamma-1}. \end{aligned} \tag{8.4}$$

The function $\xi(x)$ is the error in the Taylor expansion bounded in (8.3).

Between successive x_i^*'s the density $\hat{f}$ is linear. The original density is within $2Mm^{-\gamma-1}$ of a straight line between $f(x_i^*)$ and $f(x_{i+1}^*)$,

$$\begin{aligned} f(x) &= f(x_i^*) + (x - x_i^*)f'(x_i^*) + \xi \\ &= f(x_i^*) + (x - x_i^*)f'(t) + (x - x_i^*)(f'(x_i^*) - f'(t)) + \xi, \end{aligned}$$

where t is the point between x_i^* and x such that $f'(t)$ is the slope of the line between $f(x_i^*)$ and $f(x_{i+1}^*)$. By (8.3), $|\xi|$ and $|(x - x_i^*)(f'(x_i^*) - f'(t))|$ are both less than $Mm^{-\gamma-1}$.

The total error between the densities at any point x is thus

$$|f(x) - \hat{f}(x)| \le 3Mm^{-\gamma-1} \qquad \text{for } \frac{1}{m} \le x \le \frac{m-1}{m}.$$

There is a bit of a complication at the edges of the intervals. The density $\hat{f}$ is defined to be a constant $f^*(1)$ or $f^*(m)$ at either edge. A rougher bound on $|f - \hat{f}|$ of M/m applies at the edges because the derivative is bounded by M. Therefore, the total error is

$$\begin{aligned} \int (f - \hat{f})^2 &= \int_0^{1/m} + \int_{1/m}^{(m-1)/m} + \int_{(m-1)/m)}^1 (f - \hat{f})^2 \\ &\le Mm^{-3} + Mm^{-3} + 4Mm^{-2\gamma-2} \le Mm^{-2\gamma-2} \end{aligned}$$

as long as $\gamma \le 1/2$. □

8.2. *The normal experiment.* The multivariate normal can be approximated by a continuous Gaussian process,

$$Y(t) = \int_0^t f^{1/2}\,dt + \frac{1}{2\sqrt{n}} W(t),$$

where $W(t)$ is the standard Brownian motion process. The distributions of these continuous processes form an experiment $\bar{\mathcal{Q}}$.

Note. The functions $g = f^{1/2}$ are members of the smoothness class $\mathcal{F}(\gamma, \sqrt{\varepsilon}, \frac{M}{\sqrt{\varepsilon}})$.

The increments of the $Y(t)$ process over the intervals are

$$\hat{Y}_i \equiv Y(i/m) - Y([i-1]/m) \sim \mathcal{N}\biggl(\int_{(i-1)/m}^{i/m} f^{1/2}, \frac{1}{4nm}\biggr)$$

and the distributions of these increments form an experiment $\hat{\mathcal{Q}}$ such that $\delta(\bar{\mathcal{Q}}, \hat{\mathcal{Q}}) = 0$.

Rescaling these increments, $\sqrt{mn}\hat{Y}_i$, generates approximately the same distributions as

$$Y_i \sim \mathcal{N}\bigl(\sqrt{n\theta_i}, \tfrac{1}{4}\bigr).$$

The difference between the means is

$$\begin{aligned}\frac{n^{1/2}}{m^{1/2}}\biggl|\sqrt{f^*(i)} - m\int_{(i-1)/m}^{i/m} f^{1/2}\biggr| &\le \frac{n^{1/2}}{m^{1/2}}\Bigl|\sqrt{f(x_i^*)} - \sqrt{f^*(i)}\Bigr| \\ &\quad + \frac{n^{1/2}}{m^{1/2}}\biggl|m\int_{(i-1)/m}^{i/m}\Bigl(\sqrt{f^*(i)} - \sqrt{f(x)}\Bigr)\,dx\biggr|.\end{aligned}$$

Each of these two terms is less than $n^{1/2}m^{-1/2}(M\varepsilon^{1/2}m^{-\gamma-1})$ by a simple variation of the reasoning in (8.4).

The Hellinger distance between the multivariate normals is less than the sum of the squared distances between the means

$$\begin{aligned}H^2\bigl(\{Y_i\}_{i=1}^n, \{\bar{Y}_i\}_{i=1}^n\bigr) &\le 2\sum_{i=1}^m \frac{n}{m}\biggl(\sqrt{f^*(i)} - m\int_{(i-1)/m}^{i/m} f^{1/2}\biggr)^2 \\ &\le n\frac{M^2}{\varepsilon}m^{-2\gamma-2}.\end{aligned}$$

Therefore,

$$\Delta(\hat{\mathcal{Q}}_m, \mathcal{Q}_m^*) \le Cn^{1/2}m^{-\gamma-1}. \tag{8.5}$$

To bound $\delta(\hat{\mathcal{Q}}_m, \bar{\mathcal{Q}})$, the transformation is a bit more involved. The $\hat{Y}_i$ provide approximations at the midpoints of the intervals. Then, in analogy to the density estimation situation let $V_i(x)$ be the function

$$V_i(x) = m - m^2|x - x_i^*|$$

for $(i-3)/(2m) \leq x \leq (i+1)/(2m)$ and 0 elsewhere. Define $V_0(t)$ and $V_n(t)$ as $1/m$ on the half intervals at either edge. Then let

$$Y^*(t) = \int_0^t \left(\sum_{i=1}^m \hat{Y}_i V_i(x) \right) dx + \frac{1}{2\sqrt{n}} \sum_{i=1}^m \frac{1}{\sqrt{m}} B_i(t),$$

where the B_i are independent zero mean Gaussian processes with variances

$$\operatorname{Var}(B_i) = \int_0^t V_i - \left[\int_0^t V_i \right]^2.$$

These processes can be constructed from a standard Brownian bridge $B(t)$ via

$$B_i(t) = B\left(\int_0^t V_i \right).$$

The drift of the Y^* process is

$$\sum_{i=1}^m V_i(x) \mathbb{E}(\hat{Y}_i) = \sum_{i=1}^m V_i(x) \int_{(i-1)/m}^{i/m} f^{1/2} \equiv \hat{g}(x),$$

where $\hat{g}$ is a linear interpolation between the midpoints like $\hat{f}$ except for the function $g = f^{1/2}$.

The variance in the process comes from two sources, the variance of the observed $\hat{Y}_i$,

$$\operatorname{Var}\left(\int_0^t \left(\sum_{i=1}^m \hat{Y}_i V_i(x) \right) dx \right) = \sum_{i=1}^m \frac{1}{4mn} \left[\int_0^t V_i(t) \right]^2$$

and the contribution from the sum of the B_i,

$$\operatorname{Var}\left(\frac{1}{2\sqrt{n}} \sum_{i=1}^m \frac{1}{\sqrt{m}} B_i(t) \right) = \frac{1}{4mn} \sum_{i=1}^m \left[\int_0^t V_i(t) - \left(\int_0^t V_i(t) \right)^2 \right].$$

The result is that

$$\operatorname{Var}(Y^*(t)) = \frac{1}{4n} \int_0^t \sum_{i=1}^m \frac{1}{m} V_i(t) = \frac{t}{4n}.$$

Therefore,

$$Y^*(t) = \int_0^t \hat{g}(x)\, dx + \frac{1}{2\sqrt{n}} W(t). \tag{8.6}$$

The total variation distance between $Y^*(t)$ constructed this way and the Gaussian process $Y(t)$ is on the order of

$$2\sqrt{n} \|\hat{g} - g\|_2$$

and thus, applying Lemma 4 for $g \in \mathcal{F}(\gamma, \sqrt{\varepsilon}, M/\sqrt{\varepsilon})$,

$$\Delta(\bar{\mathcal{Q}}, \hat{\mathcal{Q}}) \leq C\sqrt{n}m^{-\gamma-1}, \tag{8.7}$$

where C depends on M and ε.

Therefore, combining (8.5) and (8.7), the distance

$$\Delta(\bar{\mathcal{Q}}, \mathcal{Q}_m^*) \leq 2Cn^{1/2}m^{-\gamma-1}. \tag{8.8}$$

8.3. *Choosing m*. Combining the results in (8.2) and (8.8) along with Theorem 1, the deficiency distance is

$$\Delta(\bar{\mathcal{P}}, \bar{\mathcal{Q}}) \leq Cn^{1/2}m^{-\gamma-1} + C_R\frac{m \log m}{n^{1/2}}.$$

This bound goes to zero when m is chosen to be $n^{1/2-\zeta}$ for $\zeta < \gamma/2$. Furthermore,

$$\Delta(\bar{\mathcal{P}}, \bar{\mathcal{Q}}) \leq Cn^{-\gamma/(\gamma+2)} \log n$$

when the dimension is chosen to be $m = n^{1/(2+\gamma)}$.

APPENDIX

A. Proof of Lemma 1. The two kernels, K_s^t and $L_{s,t,x}^y$ can be combined to form a single kernel M_x^y by

$$M_x^y B = K_x^{s,t,x}[L_{s,t,x}^y B],$$

where the kernel $K_x^{s,t,x}$ is defined for $A \in \sigma(S) \times \sigma(T) \times \mathcal{A}$ by

$$K_x^{s,t,x} A = K_{T(x)}^s \{s : (s, T(x), x) \in A\}.$$

Thus the kernel K_t^s is extended to measures on the product space $\mathcal{S} \otimes \mathcal{T} \otimes \mathcal{X}$ that have support on the set $\{T = t,\ X \in T^{-1}(t)\}$.

The first step toward bounding $\|\mathbb{P}_\theta M_x^y - \mathbb{Q}_\theta\|$ is to express the $\mathbb{P}_\theta M$ distribution as

$$\mathbb{P}_\theta M_x^y = \mu_\theta^t P_t^x K_t^s L_{s,t,x}^y = \mu_\theta^t K_t^s P_t^x L_{s,t,x}^y,$$

where the change of order is justified because for any particular value of t the joint distribution $P_t^x K_t^s$ makes S and X independent.

Therefore, for any $\mathcal{B}$-measurable function $g(y)$ such that $|g(y)| \leq 1$,

$$\begin{aligned}|\mathbb{Q}_\theta g(y) - \mathbb{P}_\theta M g(y)| &= \left|\lambda_\theta^s Q_s^y g(y) - \mu_\theta^t K_t^s P_t^x L_{s,t,x}^y g(y)\right| \\ &\leq \left|\lambda_\theta^s Q_s^y g(y) - \mu_\theta^t K_t^s Q_s^y g(y)\right| \\ &\quad + \left|\mu_\theta^t K_t^s Q_s^y g(y) - \mu_\theta^t K_t^s P_t^x L_{s,t,x}^y g(y)\right|\end{aligned}$$

by the triangle inequality.

In the first term, $Q_s^y g(y)$ is a $\sigma(S)$-measurable function such that $|Q_s^y g(y)| \le 1$. Thus

$$\big|\lambda_\theta^s[Q_s^y g(y)] - \mu_\theta^t K_t^s[Q_s^y g(y)]\big| \le \|\lambda_\theta^s - \mu_\theta^t K_t^s\| \le \varepsilon.$$

In the second term,

$$\begin{aligned}\big|\mu_\theta^t K_t^s[Q_s^y g(y)] - \mu_\theta^t K_t^s[P_t^x L_{s,t,x}^y g(y)]\big| &\le \mu_\theta^t K_t^s\big|Q_s^y g(y) - P_t^x L_{s,t,x}^y g(y)\big| \\ &\le \mu_\theta^t K_t^s \|Q_s^y - P_t^x L_{s,t,x}^y\| \\ &\le \mu_\theta^t K_t^s \rho(s,t).\end{aligned}$$

Thus, $|\mathbb{Q}_\theta g(y) - \mathbb{P}_\theta M g(y)| \le \varepsilon + \mu_\theta^t K_t^s \rho(s,t)$ for any $|g| \le 1$, and therefore the total-variation distance is also bounded

$$\|\mathbb{Q}_\theta - \mathbb{P}_\theta M_x^y\| \le \varepsilon + \mu_\theta^t K_s^t \rho(s,t). \qquad \square$$

B. The distance between a binomial and normal.

B.1. *A bound on the total-variation distance between product distributions.* The total-variation distance between distributions P and Q that are dominated by μ is defined to be

$$\|P - Q\| = \frac{1}{2}\mu\left|\frac{dP}{d\mu} - \frac{dQ}{d\mu}\right|.$$

This distance is bounded by the Hellinger distance,

$$\|P - Q\| \le H(P,Q) = \left[2 - 2P\sqrt{\frac{dQ}{dP}}\right]^{1/2} \le \sqrt{2}.$$

If there is a set A where the distributions are close on A and A^c is small then a useful bound on the Hellinger distance is

$$H^2(P,Q) \le 2 - 2PA\sqrt{\frac{dQ}{dP}}.$$

For the likelihood ratios it will be convenient to use the Kullback–Leibler divergence [Kullback (1967)] $P \log dP/dQ$. The divergence bounds the Hellinger distance,

$$\begin{aligned}H^2(P,Q) &\le 2 - 2PA\sqrt{\frac{dQ}{dP}} \\ &= 2 - 2PA + 2PA\left(1 - \sqrt{\frac{dQ}{dP}}\right) \\ &\le 2PA^c + PA\log\frac{dP}{dQ}.\end{aligned}$$

This bound is especially useful if P and Q are product measures because then

$$PA\log\frac{dP}{dQ} = \sum_i P_i A \log\frac{dP_i}{dQ_i}.$$

Therefore,

$$\|P - Q\| \le \left[2PA^c + \sum_i P_i A \log\frac{dP_i}{dQ_i}\right]^{1/2}. \tag{B.1}$$

B.2. *The local-limit theory bound between binomial and normal densities.* For Lemma 2, the total-variation distance between a product of normals and a product of smoothed binomials is needed. To apply the inequality (B.1), let

$$A = \bigcap_{i=1}^{m} \{x_i : |x_i - n_i p_i| \le (n_i p_i q_i)^{2/3}\}.$$

Standard tail bounds on the binomial show that $\mathbb{P}_{\mathbf{p}}A^c \le m\exp[-C(n/m)^{1/3}]$ which is smaller than the rest of the terms in the bound.

Let $b(k)$ be the binomial density, $\binom{n}{k}p^n q^{n-k}$ for $k = 0, \dots, n$, and let $b(x)$ be the density that is equal to $b(k)$ for $|x - k| \le 1/2$. If the U_j are independent uniform$[-1/2, 1/2]$ distributions, then $b(x_j)$ is the density of a single coordinate of $\mathbb{P}_p \star \mathbf{U}$.

Let

$$\varphi(x) = \frac{1}{\sqrt{2\pi npq}}\exp\left[-\frac{(x-np)^2}{2npq}\right],$$

the density of a $\mathcal{N}(np, npq)$.

By (B.1), it is enough to bound

$$(\mathbb{P}_p \star \mathbf{U})A\log\left[\frac{b(x_j)}{\phi(x_j)}\right]$$

for each coordinate.

Prohorov (1961) approximates the log of the likelihood ratio at each integer by

$$\log\left[\frac{b(k)}{\varphi(k)}\right] = C\frac{p-q}{6\sigma}[z^3 - 3z] + O\big([z^4 + z^2 + 1]\sigma^{-2}\big)$$

for k an integer in A and

$$\sigma = \sqrt{npq} \quad \text{and} \quad z = \frac{k - np}{\sigma}.$$

For noninteger values of x, the density $b(x)$ is equal to $b(k)$ where k is the closest integer to x. For the normal density,

$$\begin{aligned}\log\frac{\varphi(k)}{\varphi(x)} &= \frac{1}{2\sigma^2}\big[(x-np)^2-(k-np)^2\big]\\ &= \frac{1}{2\sigma^2}\big[(x-k)^2+2(x-k)(k-np)\big].\end{aligned}$$

So the integral is

$$(\mathbb{P}_{\mathbf{p}}\star\mathbf{U})A\log\frac{b(x_j)}{\varphi(x_j)} = (\mathbb{P}_{\mathbf{p}}\star\mathbf{U})A\bigg[\log\frac{b(K)}{\varphi(K)}+\log\frac{\varphi(K)}{\varphi(x_j)}\bigg],$$

where K is the integer closest to x_j,

$$\begin{aligned}&(\mathbb{P}_{\mathbf{p}}\star\mathbf{U})A\bigg[\log\frac{b(K)}{\varphi(K)}+\log\frac{\varphi(K)}{\varphi(x)}\bigg]\\ &\qquad\le \mathbb{P}_{\mathbf{p}}A\bigg[\frac{p-q}{6\sigma}(z^3-3z)+O\bigg(\frac{z^4+z^2+1}{\sigma^2}\bigg)\bigg]\\ &\qquad\quad+\frac{1}{2\sigma^2}(\mathbb{P}_{\mathbf{p}}\star\mathbf{U})\big[(x-K)^2+2(x-K)(K-np)\big].\end{aligned}$$

The moments of the binomial are [Johnson and Kotz (1969), pages 50–82]

$$\begin{aligned}&\mathbb{P}_p zA\le\frac{np}{\sigma}\mathbb{P}A^c, &&\mathbb{P}_p z^2A\le 1,\\ &\mathbb{P}_p z^3A\le\frac{p-q}{\sigma}+n^3\mathbb{P}A^c, &&\mathbb{P}_p z^4A\le 3+\frac{1-6pq}{\sigma^2},\end{aligned}$$

where $\mathbb{P}_pA^c$ is small. Thus,

$$\mathbb{P}_pA\bigg[\frac{p-q}{6\sigma}(z^3-3z)+O\bigg(\frac{z^4+z^2+1}{\sigma^2}\bigg)\bigg]=\frac{(p-q)^2}{6\sigma^2}+O(\sigma^{-2}).$$

The other expectation is computed using the fact that $x-K$ is uniformly distributed over $[-1/2,1/2]$ and is independent of x. Thus,

$$\begin{aligned}&(\mathbb{P}_p\star\mathbf{U})\big[(x-K)^2+2(x-K)(K-np)\big]\\ &\qquad=(\mathbb{P}_p\star\mathbf{U})\big[(x-K)^2-2(x-K)^2+2(x-K)(x-np)\big]=-\tfrac{1}{12}\end{aligned}$$

and this expectation can be ignored.

All the contributions to the distance are less than $C\sigma^{-2}$. Therefore, the bound between the product experiments is

$$\|\mathbb{P}_{\mathbf{p}}\star\mathbf{U}-\mathbb{Q}_{\mathbf{p}}\|\le\bigg[\sum_j\frac{C}{t_jp_jq_j}\bigg]^{1/2}.$$

C. Total-variation distance between normals. It is convenient to deal with normal experiments because the distance between them is bounded rather easily. Let $Q_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $Q_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$. Then the Hellinger affinity is

$$Q_1\sqrt{\frac{dQ_2}{dQ_1}} = \sqrt{\frac{2\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}}\exp\biggl[-\frac{(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)}\biggr],$$

so that the Hellinger distance between normals is bounded by

$$H^2(Q_1, Q_2) \le 2\biggl(1-\frac{\sigma_1^2}{\sigma_2^2}\biggr)^2 + \frac{(\mu_1-\mu_2)^2}{2\sigma_2^2}.$$

This inequality bounds the total-variation distance between $m/2$ independent $\mathcal{N}(s_j p_j, n\psi_j p_j q_j)$ and $\mathcal{N}(t_j p_j, t_j p_j q_j)$ distributions by

$$\text{(C.1)} \qquad \biggl[\sum_j \frac{p_j}{q_j}\frac{(t_j-s_j)^2}{2n\psi_j} + 2\biggl(1-\frac{t_j}{n\psi_j}\biggr)^2\biggr]^{1/2}$$

because the Hellinger distance is greater than total-variation distance, and the squared Hellinger distance between product measures is less than the sum of the squared Hellinger distances.

Acknowledgments. I want to thank my thesis advisor David Pollard for all his help in every stage of writing this paper. I would also like to thank an Associate Editor for his helpful suggestions.

REFERENCES

BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

BROWN, L. D. and ZHANG, C.-H. (1998). Asymptotic nonequivalence of nonparametric experiments when the smoothness index is 1/2. *Ann. Statist.* **26** 279–287.

CARTER, A. V. (2001). Deficiency distance between multinomial and multivariate normal experiments under smoothness constraints on the parameter set. Technical Report, UCSB. Available at www.pstat.ucsb.edu/faculty/carter/research.html.

FELLER, W. (1968). *An Introduction to Probability Theory and Its Applications* **1**, 3rd ed. Wiley, New York.

GOLUBEV, G. K. and NUSSBAUM, M. (1998). Asymptotic equivalence of spectral density and regression estimation. Technical Report 420, Weierstrass Institute, Berlin.

GRAMA, I. and NUSSBAUM, M. (1998). Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Related Fields* **111** 167–214.

JOHNSON, N. L. and KOTZ, S. (1969). *Distributions in Statistics*: *Discrete Distributions*. Houghton Mifflin, Boston.

KLEMELÄ, J. and NUSSBAUM, M. (1998). Constructive asymptotic equivalence of density estimation and Gaussian white noise. Discussion paper 53, Sonderforschungsbereich 373, Humboldt Univ., Berlin.

KULLBACK, S. (1967). A lower bound for discrimination in terms of variation. *IEEE Trans. Inform. Theory* **13** 126–127.

LE CAM, L. (1964). Sufficiency and approximate sufficiency. *Ann. Math. Statist.* **35** 1419–1455.

LUCKHAUS, S. and SAUERMANN, W. (1989). Multinomial approximations for nonparametric experiments which minimize the maximal loss of Fisher information. *Probab. Theory Related Fields* **81** 159–184.

MILSTEIN, G. and NUSSBAUM, M. (1998). Diffusion approximation for nonparametric autoregression. *Probab. Theory Related Fields* **112** 535–543.

MÜLLER, D. W. (1979). Asymptotically multinomial experiments and the extension of a theorem of Wald. *Z. Wahrsch. Verw. Gebiete* **50** 179–204.

NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.

PROHOROV, YU. V. (1961). Asymptotic behavior of the binomial distribution. *Select. Transl. Math. Statist. Probab.* **1** 87–96. Amer. Math. Soc., Providence, RI [(1953). *Uspekhi. Mat. Nauk.* **8** 135–142 (in Russian)].

TORGERSEN, E. (1991). *Comparison of Statistical Experiments*. Cambridge Univ. Press.

DEPARTMENT OF STATISTICS
AND APPLIED PROBABILITY
UNIVERSITY OF CALIFORNIA
SANTA BARBARA, CALIFORNIA 93106
E-MAIL: carter@pstat.ucsb.edu