[§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 $\mathbb{N}_\theta = \otimes_{i\le m} N(n\theta_i, n\theta_i)$.

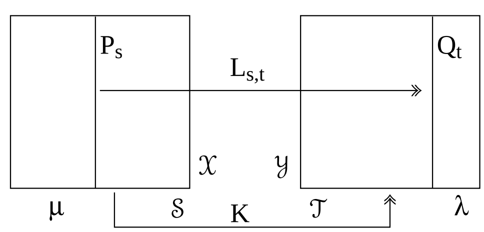
(ii) $\mathcal{N}_{\text{stabil}} := \{\widetilde{\mathbb{N}}_\theta : \theta \in \Theta\}$, where $\widetilde{\mathbb{N}}_\theta = \otimes_{i\le m} N(\sqrt{n\theta_i}, 1/4)$.

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 $N(0, c^2 W)$ 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

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

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

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 $\theta := (\theta_1, \dots, \theta_m)$ actually depend on the underlying density. That is,

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

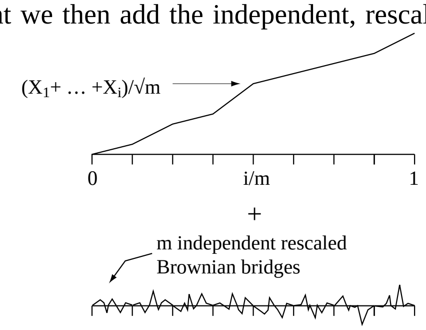
REMARK. Perhaps I should write θ_f to indicate the dependence on f when discussing the application of Carter's general inequality <1> to the Nussbaum problem.

From <16> we have

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

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 $\{X_i : i = 1, \dots, m\}$. 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 $i = 0, 1, \dots, m$. 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 $\xi_1, \dots, \xi_m$ are independent $N(0, 1)$

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

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

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

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

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.