COUPLED SAMPLES IN SIMULATION

LUC DEVROYE

McGill University, Montreal, Canada

(Received July 1987; revisions received April, September 1988; accepted September 1989)

Assume that we wish to generate two samples of n independent identically distributed random variables, $(X_1, \ldots, X_n)$ and $(Y_1, \ldots, Y_n)$, where X_1 and Y_1 have densities f and g, respectively. If these samples are used in a simulation, and f is *close* to g, it is sometimes desirable to have close simulation results. This can be achieved by insisting that both samples agree in most of their components, that is, $X_i = Y_i$ for as many i as possible under the given distributional constraints. Samples with this property are said to be optimally coupled. In this paper, we propose and study various methods of coupling two samples, a sequence of samples and an infinite family of samples.

We call a sequence of n i.i.d. random variables $X_1, \ldots, X_n$ *a sample*. In this paper, we explore the possible uses and limitations of *coupled samples* in simulation, where the coupling between two samples $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ can be measured roughly by $N = \sum_{i=1}^{n} I_{[X_i \neq Y_i]}$. If the two samples have the same distribution, it is clearly possible to have $N = 0$, but for unequal distributions, there is a limitation as to how small N can be. N can be interpreted as the number of X_i's we have to replace by other values (Y_i's) to obtain a sample of Y_i's with the correct distribution. N is proportional to the amount of surgery we have to perform on the X_i sample in order to obtain a Y_i sample. Not surprisingly, the magnitude of N (or the amount of coupling) is related to the closeness of the distributions of X_1 and Y_1.

When we present experimental results for varying distributions, it is sometimes desirable to have close outcomes when distributions are close. This leads to smooth nonoscillatory experimental plots. This point will be picked up in more detail in Section 6. One way to strongly connect two experimental results is to use coupled samples or samples for which N is small. Schmeiser and Kachitvichyanukul (1986) present another method. Whitt (1976) has pointed out that maximally correlated samples can be obtained as follows: let F, G be the distribution functions for X_1 and X_2, let $U_1, \ldots, U_n$ be i.i.d. uniform [0, 1] random variables, and define $X_i = F^{-1}(U_i)$ and $Y_i = G^{-1}(U_i)$ for $1 \leq i \leq n$. When $F \neq G$ at every point, it is clear that this method yields $N = n$. However, when the samples are used to estimate a quantity such as the kth moment, both maximal correlation and coupling may give good results. Also note that our coupling is not restricted to univariate distributions, while multivariate maximal correlations are not obvious.

As a simple comparison of the maximal coupling and maximal correlation methods referred to above, suppose that we have a random variable X and we are interested in the probability that X belongs to a given set A. Repeated i.i.d. observations $X_1, \ldots, X_n$ are made and the unknown quantity $p = P(X \in A)$ is estimated by the proportion of X_i's falling in A, which we denote by p_n. Assume that we wish to estimate the probability of the set A when there is a slight change in the distribution of X. Let the new random variable be Y, and let the new probability and estimate be q and q_n. We want to keep $Ep_n = p$ and $Eq_n = q$, while correlating the results. For example, we could ask that $\mathrm{Var}(p_n - q_n)$ be small. If we were to draw a new sample for q_n, then $\mathrm{Var}(p_n - q_n) = \mathrm{Var}(p_n) + \mathrm{Var}(q_n) = p(1-p)/n + q(1-q)/n$. Importantly, this variance is unaffected by the closeness of the two distributions. As an extreme case, if X and Y are identically distributed, then $\mathrm{Var}(p_n - q_n) = 2p(1-p)/n$. We make the example more specific by letting X be uniformly distributed on [0, 1] and setting $Y = (1+\varepsilon)X$ for some small positive ε. Define the special set A as $(-\infty, a] \cup [1, \infty)$ where $a < 1/2$ is a constant. Let N_B be the number of U_i's falling in a set B. Then, the maximum correlation method just described uses the U_i's as the sample drawn from the distribution of X, and uses $(1+\varepsilon)U_1, \ldots, (1+\varepsilon)U_n$ as the Y-sample. Thus

$$p_n = \frac{N_{[0,a]}}{n}, \quad q_n = \frac{N_{[0,a/(1+\varepsilon)]} + N_{[1/(1+\varepsilon),1]}}{n}$$

Subject classifications: Probability: Kolmogorov entroy. Simulation, random variable generation: generating dependent samples. Statistics, correlation: coupled samples.

0030-364X/90/3801-0115 $01.25

so that $Ep_n = p = a$, $Eq_n = q = a + (1 - a)(\varepsilon/(1 + \varepsilon))$ and

$$p_n - q_n = \frac{N_{[a/(1+\varepsilon),1]} - N_{[1/(1+\varepsilon),1]}}{n}$$

$$\mathrm{Var}(p_n - q_n) = \frac{1}{n}\left((1+a)\frac{\varepsilon}{1+\varepsilon} - (1-a)^2\left(\frac{\varepsilon}{1+\varepsilon}\right)^2\right) \sim \frac{(1+a)\varepsilon}{n}$$

as $\varepsilon \downarrow 0$. In other words, we have achieved continuity in our experimental results; this is especially important when these results are graphically reported.

The method we are analyzing in this paper reduces in this simple example to the following technique: for each U_i, flip a coin which comes up *heads* with probability $\varepsilon/(1 + \varepsilon)$. The ith element of the Y sample is taken to be U_i (the ith element of the X-sample) when the ith coin comes up *tails*, and is defined as V_i, a newly generated uniform random variable on $[1, 1 + \varepsilon]$, otherwise. It is easy to see that p_n and q_n are individually distributed, as in the previous maximal correlation example. However, the samples agree in all but about $n\varepsilon/(1 + \varepsilon)$ of their elements, and this is beneficial when estimating probabilities of certain sets. Indeed, $p_n - q_n$ is equal to $-1/n$ times the number of *heads* that correspond to U_i's with $U_i > a$. First, it is guaranteed to be negative, which was not the case in the previous example. Furthermore, from properties of the binomial distribution

$$\mathrm{Var}(p_n - q_n) = \frac{(1 - a)\varepsilon}{n(1 + \varepsilon)}\left(1 - \frac{(1 - a)\varepsilon}{1 + \varepsilon}\right) \sim \frac{(1 - a)\varepsilon}{n} \quad \text{as } \varepsilon \downarrow 0.$$

For all $\varepsilon < 1$ and $a < ½$, the variance is smaller than for the maximal correlation case; as $\varepsilon \to 0$, the asymptotic ratio of the variances is $(1 - a)/(1 + a)$.

The main body of the paper is concerned with the generation of coupled samples. In Section 1, we present lower bounds for $E(N)$, the amount of coupling, in terms of the *distance* between the distributions. When the lower bound is achieved by a certain method, we say that the coupling is optimal (Section 2). Optimally coupled samples can be generated by the methods introduced in Section 3. In Section 4, we propose a method for the generation of coupled samples for all densities in infinite families of densities (such as the family of all Lipschitz densities on [0, 1], or the family of all gamma densities with a shape parameter between 1 and 1,000). Interestingly, even though there are an infinite number of members in these families, we can deduce all the samples from one data pool of the size $O(n)$. This device could prove useful in simulations in which distributions change slightly from simulation run to simulation run. In Section 5, we introduce the coupling coefficient, which tells us how far we are removed from optimal coupling with suboptimal methods. Another example of continuity in reporting experimental results is given in Section 6. Sections 7 through 9 establish the connection between the time and space complexities of the proposed coupling method for families of densities, and a measure of the *richness* of these families, the metric entropy.

In this paper, we hope to show the utility of coupled samples in simulation. The idea of coupling random variables, samples, or processes has been used in various other contexts as well. A partial listing of applications is given below.

Doeblin (1937) and later Pitman (1974, 1976), Griffeath (1975), Goldstein (1979) and Thorisson (1986) looked at coupled discrete time stochastic processes. The idea here is to consider two simultaneous realizations of these processes on the same probability space so that the processes eventually agree. This has been useful in proving ergodic theorems for these processes.

Consistency proofs in nonparametric density estimation are rendered very simple by the judicious use of coupled samples, by replacing a sample of an arbitrary density by a coupled sample of a close density with nice properties (e.g., a density with infinitely many bounded derivatives and compact support). See, for example, Devroye (1985).

Coupling inequalities such as Doeblin's have been used extensively by probabilists in the study of approximations of sums of independent discrete random variables by Poisson random variables, with key results reported in Lecam (1960), Serfling (1975) and Deheuvels and Pfeifer (1986).

In simulation, one is sometimes given a sample drawn from an unknown density f and asked to generate a new sample with this density. This, of course, is an impossible task. Yet, in the study of how well one can do, the concept of coupled samples pops up again. See, for example, Devroye (1986a).

1. LOWER BOUNDS FOR THE AMOUNT OF COUPLING

We begin with the following lower bound.

Theorem 1. *Let $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_n$ be samples with marginal probability measures μ and ν, and let*

$$N = \sum_{i=1}^{n} I_{[X_i \neq Y_i]}.$$

Then, if **P** *is the collection of all distributions of* $X_1, \ldots, X_n, Y_1, \ldots, Y_n$ *such that the* X_i*'s and* Y_i*'s are samples with the given marginals (i.e.,* **P** *covers all possible dependencies between the samples)*

$$\inf_{\mathbf{P}} E(N) = n \sup_A |\mu(A) - \nu(A)|$$

where A ranges over all Borel sets on the real line.

Proof. Note that

$$\begin{aligned} E(N) &= \sum_{i=1}^{n} P(X_i \neq Y_i) \\ &\geqslant \sum_{i=1}^{n} \sup_A P(X_i \in A, Y_i \notin A) \\ &\geqslant \sum_{i=1}^{n} \sup_A (P(X_i \in A) - P(Y_i \in A)) \\ &= \sum_{i=1}^{n} \sup_A (\mu(A) - \nu(A)) \\ &= n \sup_A (\mu(A) - \nu(A)) \end{aligned}$$

which shows one half of the inequality. For the other half, we need only construct a particular dependence structure for which equality can be obtained. Note that there exists a unique measure σ on the Borel sets of R such that σ and μ are mutually singular and

$$\nu(A) = \int_A f \, d\mu + \sigma(A)$$

for all Borel sets A (see e.g., Wheeden and Zygmund 1977, p. 181), where f is the Radon-Nikodym derivative of ν relative to μ on the set where $\mu \ll \nu$. In fact, by the Lebesgue decomposition theorem, $\sigma(A) = \nu(A \cap Z)$ where Z is a set for which $\mu(Z) = 0$ and $\nu(Z^c) = 0$ (Z^c is the complement of Z). We can construct three measures on the Borel sets of R, defined as

$$\mu_1(A) = \int_{A \cap Z^c} \min(1, f) \, d\mu$$

$$\mu_2(A) = \int_{A \cap Z} d\nu + \int_{A \cap Z^c} (f - 1)_+ \, d\mu$$

$$\mu_3(A) = \int_{A \cap Z^c} (1 - f)_+ \, d\mu.$$

If we divide $\mu_i(A)$ by $\mu_i(R)$ for each i, we obtain three probability measures. Note that $\mu_1 + \mu_2 \equiv \nu$ and $\mu_1 + \mu_3 \equiv \mu$, so that we must have $\mu_2(R) = \mu_3(R) = 1 - \mu_1(R) = \delta$ for some $\delta \in [0, 1]$.

The construction is based on n independent 4-tuples (U, V, W, Z) where V, W, Z are three independent random variables, one from each probability measure, and U is an independent Bernoulli random variable that takes the value 1 with probability δ and the value 0 probability $1 - \delta$. If $U = 1$, we set $(X, Y) = (W, Z)$, and if $U = 0$, we set $(X, Y) = (V, V)$. It is easy to see that if $U = 1$, we have $X \neq Y$ with probability one. Also, the marginal probability measures of the samples thus obtained are correct because X has probability measure $\mu_1 + \mu_2 = \nu$ and Y has probability measure $\mu_1 + \mu_3 = \mu$.

We observe that

$$\begin{aligned} &\sup_A |\mu(A) - \nu(A)| \\ &= \sup_A |\mu_2(A) - \mu_3(A)| \\ &= \sup_A \left| \int_{A \cap Z^c} (1 - f) \, d\mu - \nu(A \cap Z) \right| \\ &= \sup_A \left| \int_{A \cap Z^c} (1 - f)_+ \, d\mu - \nu(A \cap Z) \right. \\ &\qquad\qquad \left. - \int_{A \cap Z^c} (f - 1)_+ \, d\mu \right| \\ &\geqslant \int_{Z^c} (1 - f)_+ \, d\mu \quad (\text{take } A = \{f \leqslant 1\} \cup Z^c) \\ &= \mu_3(R) = \delta = P(U = 1) = P(X \neq Y) \\ &= EI_{[X_i \neq Y_i]} = E(N)/n \end{aligned}$$

the last because N is binomial (n, δ). Thus, this construction yields

$$E(N) \leqslant n \sup_A |\mu(A) - \nu(A)|$$

and, since the first part of the proof shows that the reverse inequality always holds, the construction gives equality as claimed.

Using Scheffe's theorem (Scheffe 1947), we have the following important corollary.

Corollary 1. *When two samples are required to have marginal densities f and g, respectively, then*

$$\inf_{\mathbf{P}} E(N) = \frac{n}{2} \int_{-\infty}^{\infty} |f(x) - g(x)| \, dx.$$

From now on, we will write $\int |f - g|$ and so forth when integration with respect to the Lebesgue measure is meant. The bounds of this section should guide us in deciding the optimality of a coupling scheme, where two samples are said to be *optimally coupled* when $E(N) = n \sup_A |\mu(A) - \nu(A)|$. Hidden in the proof of Theorem 1 is a construction that will allow us to optimally couple two samples.

In Theorem 1, we in fact reprove Doeblin's coupling inequality which states that for any two random variables X, Y, $P(X \neq Y) \geqslant \sup_A |\mu(A) - \nu(A)|$.

2. OPTIMAL COUPLING OF TWO SAMPLES

The construction given here is based upon the proof of Theorem 1. Instead of treating the general case of arbitrary probability measures μ and ν, we elected to deal with the more specific case when μ and ν have densities f and g, respectively. We can deal with the general case with equal ease.

Consider a 4-tuple (U, V, W, Z) that consists of four independent random variables: U is Bernoulli with $P(U = 1) = 1 - P(U = 0) = \delta$, V has density $\min(f, g)/(1 - \delta)$, W has density $(f - g)_+/\delta$ and Z has density $(g - f)_+/\delta$, and $\delta = ½ \int |f - g|$ (as in the proof of Theorem 1). We construct (X, Y) from (U, V, W, Z) as

$$(X, Y) = \begin{cases} (V, V) & \text{if } U = 0 \\ (W, Z) & \text{if } U = 1 \end{cases}.$$

It is easy to verify that X has density f and Y has density g; for example, X has the mixture density

$$(1 - \delta) \frac{\min(f, g)}{1 - \delta} + \delta \frac{(f - g)_+}{\delta} = f.$$

Furthermore, if we define X and Y samples by creating n independent (X, Y) pairs, we have

$$E(N) = nP(X \neq Y) = n\delta = \frac{n}{2} \int |f - g|.$$

In other words, the samples thus obtained are optimally coupled. A similar construction exists for discrete distributions; for general probability measures, the construction is only slightly more complicated, but requires knowledge of the Radon-Nikodym derivative $d\mu/d\nu$ on the set on which $\mu \ll \nu$ (μ is absolutely continuous with respect to ν).

The construction given above is wasteful because we use only part of each 4-tuple, depending upon the value of U. Although it is rather trivial to give parsimonious constructions, we will do so, nevertheless, for the sake of future generalization.

Bernoulli-Based Construction

```
FOR $i$ := 1 TO $n$ DO
  Generate a uniform [0, 1] random variable $U$.
  IF $U < \delta$
    THEN
      Generate $(W, Z)$, where $W$ has density
      $(f - g)_+/\delta$ and $Z$ has density $(g - f)_+/\delta$
      RETURN $(X_i, Y_i) \leftarrow (W, Z)$
    ELSE
      Generate $V$ with density $\min(f, g)/(1 - \delta)$
      RETURN $(X_i, Y_i) \leftarrow (V, V)$.
```

The expected total number of generated variables of the type V, W or Z is $n(1 + ½ \int |f - g|)$. For two independent samples, we need $2n$ such random variables, n for each sample. The algorithm requires n random variables of the type U. Using the waiting time method (see p. 522 of Devroye 1986b), we can reduce this somewhat; the expected number of random variables needed when generating all the times until the next occurrence of a (W, Z) pair is $1 + \delta n$.

3. GENERATING A SEQUENCE OF COUPLED SAMPLES

Assume next that we wish to generate a sequence of coupled samples of size n with marginal densities f_1, f_2, ... where the f_i's are given densities. In many experiments, researchers wish to see how the variation of one or more parameters in a given model influences certain key quantities. In those situations, consecutive f_i's are close to each other, reflecting a gradual change in one or a few descriptors of the density. The first question here is that of the construction of a sequence of pairwise optimally coupled samples, that is, a sequence of samples such that for any $i \neq j$, the ith and jth samples are optimally coupled. The Bernoulli-based construction of the previous section is simply not applicable here. It is, however, possible to construct a sequence of samples such that two consecutive samples in the sequence are optimally coupled, that is, the ith and $i + 1$st samples are optimally coupled for all i. Another construction, in which all (infinite) pairs of samples are *almost* optimally coupled is presented in another section.

The idea is borrowed from the rejection method in random variate generation (see Bratley, Fox and Schrage 1987 or Devroye 1986b for general discussions): we associate with each X a random variate ξ where ξ is uniformly distributed on $[0, f(X)]$ and f is the density of X. The point (X, ξ) is uniformly distributed under the curve of f, i.e., it has the uniform distribution on the set $\{(x, y): x \in R, 0 \leq y \leq f(x)\}$ (see Figure 1). The sample for f_1 is obtained as follows.

Original Sample

```
FOR $i := 1$ TO $n$ DO
    Generate independent random variables $X$ and $U$
    where $X$ has density $f_1$ and $U$ is uniformly distributed on $[0, 1]$.
    RETURN $(X_i, \xi_i) \leftarrow (X, Uf_1(X))$.
```

To construct the $i + 1$st sample from the ith sample, we proceed as follows.

Construction of $i + 1$st Sample from ith Sample

```
[We are given the $i$th sample $(X_1, \xi_1), \ldots, (X_n, \xi_n)$.]
FOR $j := 1$ TO $n$ DO
    IF $\xi_j > f_{i+1}(X_j)$
        THEN
            Generate $X_j$ with density $(f_{i+1} - f_i)_+/\delta_i$ where
            $\delta_i \triangleq 1/2 \int |f_i - f_{i+1}|$.
            Generate $\xi_j$ uniformly on $[f_i(X_j), f_{i+1}(X_j)]$.
    RETURN $(X_j, \xi_j)$.
```

Since only a binomial $(n, 1/2 \int |f_i - f_{i+1}|)$ number of couples (X_j, ξ_j) is changed between the ith and $i + 1$st samples, we see that these samples are optimally coupled. In the construction of the $i + 1$st sample with $i > 0$, we need on the average $n \int |f_i - f_{i+1}|$ *new* random variables, two per affected couple. The expected total number of random variables generated in the course of the generation of random samples with densities $f_1, \ldots, f_k$ (with possibly $k = \infty$) is therefore

$$2n\left(1 + \sum_{i=1}^{k-1} \frac{1}{2} \int |f_i - f_{i+1}|\right).$$

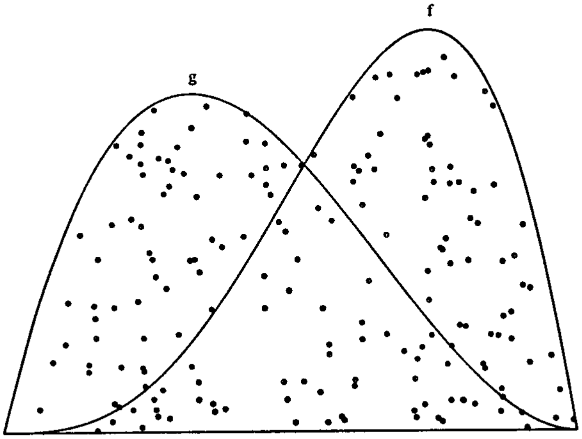

Figure 1. Two densities f and g are shown where the sample size is $n = 120$. (Sample from f consists of all x-coordinates of points under curve of f; sample from g consists of all x-coordinates of points under curve of g.)

In fact, for the expected number of replacements between the ith and kth samples (with $i < k$ arbitrary) we have

$$E(N) = \sum_{j=i}^{k-1} n \frac{1}{2} \int |f_j - f_{j+1}| \geq \frac{n}{2} \int |f_i - f_k|.$$

Hence, for a sequence of densities with no repetitions, optimal coupling is only guaranteed to occur between consecutive samples.

It is helpful to compare the expected number of generated random variables for $f_1, \ldots, f_k$ obtained by the given coupling method with that for independent samples (kn). The comparison is nearly always in favor of coupling because consecutive densities are usually close to each other. To illustrate this point, consider a gamma family with f_i equal to the gamma density with parameter i. By the local central limit theorem (see e.g., Petrov 1975, p. 213), it is known that $\int |f_i - \phi_i| \sim c/\sqrt{i}$ as $i \to \infty$ for some constant c, where ϕ_i is the normal density with the same mean and variance as f_i. Hence, we have $\int |f_i - f_{i+1}| \leq c/\sqrt{i}$ for some other constant c, and all $i \geq 1$. Thus, a conservative upper bound for the expected number of generated random variables is

$$2n \sum_{i=1}^{k} \frac{c}{\sqrt{i}} \leq 2n \int_0^k \frac{c}{\sqrt{u}}\, du = 4nc\sqrt{k}.$$

This is about $\sqrt{k}$ times less than for independent sampling. Nevertheless, the effort still tends to ∞ as $k \to \infty$. For some nice classes of densities, there is a way of constructing nearly optimal coupled samples based upon one original sample of size cn where c is some universal constant, despite the fact that the given classes may have an uncountably infinite number of members. This is dealt with in the next section.

4. LARGE FAMILIES OF DENSITIES

In this section, we deal with the problem of the construction of nearly optimal coupled samples for given families of densities F. One of the conditions that will be imposed is that the class F be L_1 *totally bounded*, that is, for every $\varepsilon > 0$, we can find a finite number of densities $f_1, \ldots, f_m$ such that

$$\sup_{f \in \mathbf{F}} \inf_{1 \leq i \leq m} \int |f - f_i| \leq \varepsilon.$$

To put it differently, **F** can be covered with a finite number of ε-balls, regardless of how small $\varepsilon > 0$ is. We will discuss several totally bounded classes in Section 9. For example, it suffices to consider all the Lipschitz densities with support on [0, 1], where a Lipschitz density f satisfies $\sup_{x,y}|f(x) - f(y)| \leq C$ for some finite constant C. However, the class of all densities on [0, 1] that are bounded by 2 is not totally bounded. Also, the class of all normal densities with zero mean is not totally bounded. The Frechet-Kolmogorov theorem (see p. 275 of Yosida 1980) provides us with easy-to-verify necessary and sufficient conditions for total boundedness:

$$1)\quad \lim_{t\to 0} \sup_{f\in \mathbf{F}} \int |f(x+t) - f(x)|\, dx = 0;$$

$$2)\quad \lim_{t\to\infty} \sup_{f\in \mathbf{F}} \int_{|x|>t} f(x)\, dx = 0.$$

The first of these conditions is satisfied, for example, if every f is unimodal, and there is a global finite bound for all f.

In addition, we assume that there exists a dominating curve g such that

$$\sup_{f\in\mathbf{F}} f(x) \leq g(x) \quad \text{for all } x$$

$$\int g(x)\, dx = G < \infty;$$

The latter condition is necessary to be able to employ some rejection-like method. Note that the existence of such a g automatically implies that condition 2 in the Frechet-Kolmogorov theorem is satisfied.

As in the previous section, we store pairs (X_i, ξ_i), but the difference is that we now store one large array of pairs in a preprocessing step, and that all the X_i's needed for any $f \in \mathbf{F}$ at some point in the future will be drawn from this large pool.

Generation of Data Pool

NOTE: the size of the pool is m (which is slightly larger than Gn, see below).

```
FOR $i := 1$ TO $m$ DO
    Generate $X_i$ with density $g/G$.
    Generate a uniform [0, 1] random variate $U$.
    Set $\xi_i \leftarrow Ug(X_i)$.
```

We note that the (X_i, ξ_i) pairs are uniformly distributed under the curve of g.

Retrieving a Sample for Density f

```
$i, j \leftarrow 0$ ($i, j$ are counters)
WHILE $i < n$ DO
    $j \leftarrow j + 1$
    IF $j > m$ THEN STOP (the algorithm fails to return a sample)
    IF $\xi_j < f(X_j)$ THEN $i \leftarrow i + 1$, RETURN $X_j$.
```

It is easy to see that the X_j's thus returned form an i.i.d. sequence drawn from f (see Figure 2). Furthermore, when we require a sample from a density f^* that is close to f in the L_1 sense, it is very likely that most of the X_j's in one sample also will be present in the other sample. We will see that the samples are nearly optimally coupled.

The more serious problem we have to address is that of the choice of m. It is possible to choose m such that

$$P\,(m \text{ does not suffice for some } f \in \mathbf{F}) \leq e^{-cn}$$

for some constant c depending upon the richness of the class **F** only. In Theorem 6, an explicit bound is derived in which c depends upon G and another measure of the complexity of **F**, namely the metric entropy of **F**, which will be introduced in Section 7. In any case, if during the construction of a sample, we exhaust the data pool, it is always possible to generate

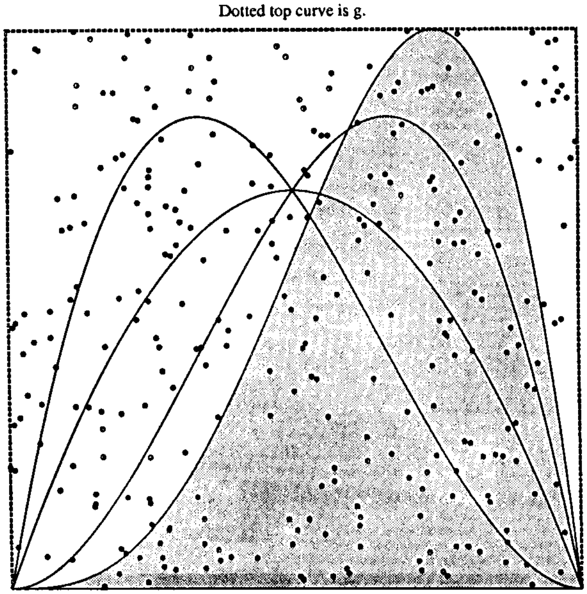

Figure 2. Four densities f and a data pool of 300 points are shown where first n points falling under f define a sample for f.

additional (X_i, ξ_i) pairs, which should be appended to the data pool. Preferably, the data pool should be implemented as a linked list. We will call M the (random) minimal number of pairs needed so that for every $f \in \mathbf{F}$, the number of points under the curve of f is at least n.

There are a number of ways to deal with the rare event if the algorithm fails. It is possible, for example, to reject the entire collection of samples, and restart the algorithm, and to repeat this until the algorithm successfully halts. If necessary, m can be increased after each rejection, so as to increase the probability of success.

5. AMOUNT OF COUPLING

We will measure the goodness of a certain method of obtaining two coupled samples with probability measures μ and ν, respectively, by the ratio

$$D(\mu, \nu) \triangleq \frac{E(N)}{n \sup_A |\mu(A) - \nu(A)|}$$

which we shall call the *coupling coefficient*. For any method, the coupling coefficient is a number at least equal to one. We are interested in the coupling coefficient with the data pool method described in the previous section. We can assume without loss of generality that $m = \infty$ if the suggestion to use a linked list is followed.

Theorem 2. *For the data pool method with $m = \infty$*

$$1 \leq \sup_{f, f^* \in \mathbf{F}, f \neq f^*} D(f, f^*) \leq 2.$$

Proof. The infinite data pool model allows us to ignore all (X_i, ξ_i)'s that do not fall under the curve of $\max(f, f^*)$. Thus, we assume without loss of generality that the index i refers to the ith pair that falls under $\max(f, f^*)$. Let us introduce the cardinalities L_k, M_k and N_k, where L_k is the number of (X_i, ξ_i) pairs with $1 \leq i \leq k$ which do not fall under the curve of f^* but under the curve of f. Similarly, M_k refers to those pairs that fall under f^* but not under f, and finally, N_k counts the number of pairs that lie under $\min(f, f^*)$. For fixed k, it is easy to see that (L_k, M_k, N_k) is multinomially distributed with parameters k, $\delta/(1 + \delta)$, $\delta/(1 + \delta)$, and $(1 - \delta)/(1 + \delta)$ where $\delta = \frac{1}{2} \int |f - f^*|$.

Let τ be the stopping time defined by the first occurrence of $\max(L_k, M_k) + N_k = n$. Then it is easy to see that N_τ is equal to the number of common pairs in both samples when we stop, and that

$$D(f, f^*) = \frac{E(N)}{n\delta} = \frac{E(\max(L_\tau, M_\tau))}{n\delta}$$
$$\leq \frac{E(L_\tau + M_\tau)}{n\delta} = \frac{2E(L_\tau)}{n\delta} \leq \frac{2E(L_{\tau'})}{n\delta}$$

where τ' is defined by the first occurrence of $L_k + N_k = n$. Clearly, $L_{\tau'}$ is binomial (n, δ), so that the upper bound of Theorem 2 is obtained without further ado.

The situation is much better than predicted by Theorem 2, in view of the following.

Theorem 3. *For any f, f^* with $\delta = \frac{1}{2} \int |f - f^*| > 0$, we have*

$$D(f, f^*) \leq 1 + \min\left(1, \sqrt{\frac{2}{n\delta}}\right).$$

Proof. We inherit the notation of the proof of Theorem 2. We begin with

$$D(f, f^*) = \frac{E(N)}{n\delta} = \frac{E(\max(L_\tau, M_\tau))}{n\delta}$$
$$\leq \frac{E(\max(L_{\tau'}, M_{\tau'}))}{n\delta}$$
$$\leq \frac{E(L_{\tau'})}{n\delta} + \frac{E(\max(0, M_{\tau'} - L_{\tau'}))}{n\delta}$$
$$= 1 + \frac{E(\max(0, M_{\tau'} - L_{\tau'}))}{n\delta}$$
$$\leq 1 + \frac{\sqrt{E((M_{\tau'} - L_{\tau'})^2)}}{n\delta}$$

(by the Cauchy-Schwarz inequality).

Using the independence of $M_{\tau'}$ and $L_{\tau'}$, the fact that $L_{\tau'}$ is binomial (n, δ), and that $M_{\tau'}$ is negative binomial $(n, 1/(1 + \delta))$ minus n, we have as upper bound

$$1 + \frac{\sqrt{(EM_{\tau'} - EL_{\tau'})^2}}{n\delta} + \frac{\sqrt{\mathrm{Var}(M_{\tau'}) + \mathrm{Var}(L_{\tau'})}}{n\delta}$$
$$= 1 + \frac{\sqrt{n\delta(1 + \delta) + n\delta(1 - \delta)}}{n\delta}$$
$$= 1 + \frac{\sqrt{2n\delta}}{n\delta}.$$

Here we use well known formulas for the moments of the negative binomial distribution (see e.g., Johnson and Kotz 1969, pp. 124–126).

6. CONTINUITY IN REPORTING EXPERIMENTAL RESULTS

When we report experiments in the form of a finite number of data points plotted in an xy-diagram, where the x-axis represents a change in distribution (such as a different parameter setting), it is to our advantage to use coupled samples rather than independent samples (one per reported data point). One limited example was reported in the Introduction. To make this case in all generality is rather difficult, so let us consider a particular situation in which we associate with each sample an average

$$\bar{h}_f = \frac{1}{n}\sum_{i=1}^{n} h(X_i)$$

where the X_i's are i.i.d. random variables with a certain density $f \in \mathbf{F}$ and h is a fixed bounded function. It is more aesthetically pleasing, and the experimental curves come across as more believable, when close densities yield close data points. Let us measure the closeness of data points for densities f and f^* by

$$Z(f, f^*) \triangleq \operatorname{Var}(\bar{h}_f - \bar{h}_{f^*})$$

where the X_i's form a sample with density f, and $\bar{h}_{f^*}$ is defined as $\bar{h}_f$, based upon an i.i.d. sample $(Y_1, \ldots, Y_n)$ with marginal density f^*. Notice that in the definition of $Z(f, f^*)$ we have automatically compensated for the fact that $h(X_1)$ and $h(Y_1)$ may have different means. The experimental results reported are *continuous* with respect to the L_1 norm if for fixed f and variable f^*, $\int |f - f^*| \to 0$ implies $Z(f, f^*) \to 0$ for all bounded functions h. We have the following theorem.

Theorem 4. *When the X_i and Y_i samples are independent, then*

$$Z(f,f^*) = \frac{1}{n}(\operatorname{Var}(h(X_1)) + \operatorname{Var}(h(Y_1))).$$

When the samples have coupling coefficient $D(f, f^*)$, we have

$$Z(f,f^*) \leq (4\|h\|_\infty)^2\left(\left(\frac{1}{2}\int |f-f^*|\, D(f,f^*)\right)^2 + \frac{1}{n}\left(\frac{1}{2}\int |f-f^*|\, D(f,f^*)\right)\right).$$

If the coupling coefficient is uniformly bounded over all f, f^*, the experimental results are continuous with respect to the L_1 norm.

Proof. The first statement is obvious by the independence of the samples. For the second part of the theorem, we have

$$E\left(\frac{1}{n}\sum_{i=1}^{n}(h(X_i) - Eh(X_i)) - \frac{1}{n}\sum_{i=1}^{n}(h(Y_i) - Eh(Y_i))\right)^2$$

$$= E\left(\frac{1}{n}\sum_{i=1}^{n}((h(X_i) - Eh(X_i)) - (h(Y_i) - Eh(Y_i)))I_{[X_i \neq Y_i]}\right)^2$$

$$\leq (4\|h\|_\infty)^2 E\left(\frac{1}{n}\sum_{i=1}^{n} I_{[X_i \neq Y_i]}\right)^2$$

$$= (4\|h\|_\infty)^2\left(E^2\left(\frac{1}{n}\sum_{i=1}^{n} I_{[X_i \neq Y_i]}\right) + \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} I_{[X_i \neq Y_i]}\right)\right)$$

$$\leq (4\|h\|_\infty)^2\left(\left(\frac{1}{2}\int |f-f^*|\, D(f,f^*)\right)^2 + \frac{1}{n}\left(\frac{1}{2}\int |f-f^*|\, D(f,f^*)\right)\right).$$

Note that for the maximum correlation method, no simple upper bound for $Z(f, f^*)$ in terms of $\int |f - f^*|$ can be given unless, at the very least, some assumptions about the smoothness of h are made. No smoothness conditions are present in Theorem 4.

7. SIZE OF THE DATA POOL AND METRIC ENTROPY

It is convenient if we choose m, the size of the data pool, such that the probability that some f has less than n points under its curve is small. This probability cannot be evaluated by Bonferroni's inequality because $\mathbf{F}$ can have an infinite number of member densities. If A_f is the set of all (x, y) with $0 \leq y \leq f(x)$, $x \in R$, and μ_m is the empirical measure defined by the data pool (X_i, Y_i), $1 \leq i \leq m$ (the empirical measure puts mass $1/m$ at each data point), then the *probability* *of eventual failure* is

$$P\left(\inf_{f\in \mathbf{F}} m\mu_m(A_f) < n\right)$$

$$= P\left(\inf_{f\in \mathbf{F}} (\mu_m(A_f) - \mu(A_f)) < \frac{n}{m} - \frac{1}{G}\right)$$

where μ is the uniform measure under the curve of g (i.e., the set $\{(x, y)\colon 0 \leq y \leq g(x), x \in R\}$), and where we use the obvious fact that $\mu(A_f) = \int f / \int g = 1/G$. Thus, we need to bound a uniform deviation of an empirical measure from its mean in such a way that the given probability tends to zero as $m \to \infty$, regardless of how large n is. Unfortunately, this is not always possible.

Consider, for example, the class of all densities on $[0, 1]$ that are bounded by $1 + \varepsilon$ where $\varepsilon > 0$ is arbitrary. For every m, n, it is possible to find a density in this class (for example, a density of height $1 + \varepsilon$ with support on a set of Lebesgue measure $1/(1 + \varepsilon)$) that has no (X_i, Y_i) under its curve. Hence, the probability of eventual failure is 1 for all m, n.

The previous example illustrates the need to limit the *size* of the class $\mathbf{F}$. It seems that one appropriate measure of the size of $\mathbf{F}$ is the *metric entropy* (with bracketing) defined as follows. We fix $\varepsilon > 0$ and find the minimum number of pairs of functions g_i, h_i with $1 \leq i \leq N_\varepsilon$ such that for every $f \in \mathbf{F}$

$$g_i \leq f \leq h_i \quad \text{for some } i$$

and

$$\int |h_i - g_i| \leq \varepsilon.$$

In other words, each density is close to a sandwiching pair of functions in the L_1 sense. The logarithm of N_ε, the minimal number of such pairs for a fixed ε, is the metric entropy. It is important to note that the metric entropy is a function of ε and the class $\mathbf{F}$. The N_ε balls with center h_i and radius ε cover $\mathbf{F}$ in the L_1 sense. Hence, $\mathbf{F}$ is L_1 totally bounded if $N_\varepsilon < \infty$. We will only consider classes that have $N_\varepsilon < \infty$ for all $\varepsilon > 0$. Note that this excludes some simple classes such as those defined by mere translations or rescalings of a fixed density.

Also, if we work on $[0, 1]$, and the centers of the ε-balls that cover a totally bounded set are $f_1, \ldots, f_k$, then the pairs $(f_1 - \varepsilon, f_1 + \varepsilon), \ldots, (f_k - \varepsilon, f_k + \varepsilon)$ can be used in the definition of the metric entropy with bracketing with 2ε instead of ε. In other words, $N_{2\varepsilon} \leq D_\varepsilon$ where D_ε is the minimum number of ε-balls needed to cover $\mathbf{F}$ (either in L_1 or L_∞). This tool is useful in obtaining upper bounds for $N_{2\varepsilon}$.

The idea of using the metric entropy in the study of the uniform deviation of empirical measures goes back to Blum (1955) and Dehardt (1971) in the Blum-Dehardt law of large numbers, which generalizes the Glivenko-Cantelli lemma. Applied to the situation at hand, it implies that if the metric entropy of $\mathbf{F}$ is finite for all $\varepsilon > 0$, then

$$\sup_{f\in\mathbf{F}} |\mu_m(A_f) - \mu(A_f)| \to 0 \quad \text{almost surely as } n \to \infty.$$

We want a bit more information in the form of an explicit inequality. The inequality of Theorem 5 is obtained by standard techniques found for example in Dudley (1984), Gine and Zinn (1984) and Yukich (1985).

Theorem 5. *Let* $\mathbf{F}$ *be a family with finite metric entropy, and assume that every* $f \in \mathbf{F}$ *is bounded by* g *where* $\int g = G < \infty$. *Let* μ *be the uniform probability measure under the curve of* g. *For every* $\varepsilon > 0$

$$P\left(\sup_{f\in\mathbf{F}} |\mu_m(A_f) - \mu(A_f)| \geq \frac{\varepsilon}{G}\right)$$

$$\leq 4N_{\varepsilon/6}\exp\left(-\frac{2}{9}\, m\left(\frac{\varepsilon}{G}\right)^2\right).$$

Proof. Let us introduce the sets B_f and C_f which are defined as follows. For fixed $\varepsilon > 0$, consider a collection of $N_{\varepsilon/6}$ pairs of functions (g_i, h_i) with the properties laid out in the definition of the metric entropy for $\mathbf{F}$. For $f \in \mathbf{F}$, we find the first pair for which $g_i \leq f \leq h_i$, and call the sets of points that fall under the curves of h_i and g_i, B_f and C_f, respectively. Thus, $C_f \subseteq A_f \subseteq B_f$ for all f. In particular $\mu(C_f - B_f) = \int (h_i - g_i)/G \leq \varepsilon/(6G)$. Clearly

$$|\mu_m(A_f) - \mu(A_f)|$$

$$\leq |\mu_m(A_f) - \mu_m(B_f)|$$

$$+ |\mu_m(B_f) - \mu(B_f)| + |\mu(B_f) - \mu(A_f)|$$

$$\leq \mu_m(B_f - C_f) + \mu(B_f - C_f) + |\mu_m(B_f) - \mu(B_f)|$$

$$\leq (\mu_m(B_f - C_f) - \mu(B_f - C_f))$$

$$+ 2\mu(B_f - C_f) + |\mu_m(B_f) - \mu(B_f)|.$$

Thus, for all $\varepsilon > 0$

$$\begin{aligned} &P\left(\sup_{f\in \mathbf{F}} |\mu_m(A_f) - \mu(A_f)| \geq \frac{\varepsilon}{G}\right) \\ &\leq P\left(\sup_{f\in \mathbf{F}} |\mu_m(B_f - C_f) - \mu(B_f - C_f)| \geq \frac{\varepsilon}{3G}\right) \\ &\quad + P\left(\sup_{f\in \mathbf{F}} |\mu_m(B_f) - \mu(B_f)| \geq \frac{\varepsilon}{3G}\right) \\ &\leq 4N_{\varepsilon/6}\exp(-2m(\varepsilon/(3G))^2) \end{aligned}$$

by Hoeffding's inequality (Hoeffding 1963) and the fact that there are at most $N_{\varepsilon/6}$ different pairs of sets (C_f, B_f).

For some classes **F**, it is advantageous to use Vapnik-Chervonenkis type inequalities (see e.g., Vapnik and Chervonenkis 1971 a, b, 1981, Devroye 1982 or Alexander 1984), but doing so will add little to the ideas developed here.

We are ready to apply Theorem 5 to the problem at hand, the choice of m. In particular, we have the following.

Theorem 6. *Let* **F** *be a class of densities with finite metric entropy, and let* $\sup_{f\in\mathbf{F}} f \leq g$ *with* $\int g = G < \infty$. *Let* $m > n$ *and* $\varepsilon \in (0, 1)$ *be related via the equality*

$$m = \frac{nG}{1 - \varepsilon}.$$

Then

$$P\left(\inf_{f\in \mathbf{F}} m\mu_m(A_f) < n\right) \leq 4N_{\varepsilon/6}\exp\left(-\frac{2}{9}m\left(\frac{\varepsilon}{G}\right)^2\right).$$

Proof. Theorem 6 follows without work from Theorem 5 and some inequalities obtained at the top of this section.

Theorem 6 shows the importance of taking G as small as possible. Ideally, we put $g = \sup_{f\in\mathbf{F}} f$, so that G is minimal, but such a function g is often not easy to derive in practice.

For a good choice of ε in $m = nG/(1 - \varepsilon)$, we need some good estimate or bound for $N_{\varepsilon/6}$. One can play it conservatively and take ε constant, unrelated to m. For example, with $\varepsilon = 0.5$ ($m = 2nG$), the probability of eventual failure of the method does not exceed

$$4N_{1/12}\exp\left(-\frac{m}{18G^2}\right).$$

This tends to zero exponentially fast with n, with the coefficients in the rate of decrease directly affected by the *rejection constant* G and the *metric entropy* $\log(N_{1/12})$. It is also possible to let ε vary with m in such a way that $\varepsilon \to 0$, so that asymptotically $m \sim nG$. This amounts to moderate savings spacewise.

Finally, with the maximum correlation method discussed in the Introduction, a data pool of size $O(n)$ is needed. However, k samples require kn (often expensive) inversions of a distribution function. The method described by us requires a data pool of size $O(n)$ (with a larger constant hidden in the big oh), but k samples are obtained at a cost of kn evaluations of a density.

8. EXPECTED VALUE OF THE SIZE OF THE DATA POOL

Assume that we run the algorithm starting with $m = 0$, and that we enlarge the data pool as required. If we continue to do this ad infinitum, then the number of data pairs in the pool eventually reaches a random (but finite) limit M with probability one. If we knew M beforehand (that is, if we could peek at the future), then we would not be stuck with the problem of choosing m. Nevertheless, M is vitally important for storage purposes. Its size is again related to n, G and the metric entropy of **F**. We have the following upper bound for $E(M)$.

Theorem 7. *Assume that* **F** *has finite metric entropy for all* $\varepsilon > 0$, *and that* $G < \infty$. *Then* $E(M) \sim nG$ *as* $n \to \infty$. *Furthermore*

$$\begin{aligned} &E(M) - nG \\ &\quad \leq 72G^2N_{1/12}\exp\left(-\frac{n}{9G}\right) + 2 \\ &\qquad + \inf_{\varepsilon\in(0,1)} \left\{\frac{nG\varepsilon}{1 - \varepsilon} + 36G^2N_{\varepsilon/6}\exp\left(-\frac{4n\varepsilon^2}{9G}\right)\right\}. \end{aligned}$$

If $N_u \leq \exp(Cu^{-b})$ *for* $u > 0$ *for some finite* $C, b > 0$, *then the right-hand side is*

$$O(n^{(b+1)/(b+2)}) \quad \textit{as } n \to \infty.$$

Proof. Clearly, for any constant K

$$E(M) = \sum_{k=0}^{\infty} P(M > k) \leq K + \sum_{k=K}^{\infty} P(M > k).$$

For fixed very small $\varepsilon > 0$, the first term can be made arbitrarily close to nG in the relative sense, and the second term tends to zero since $K \to \infty$. This concludes the first part of the proof of Theorem 7.

Taking $K = \lceil nG/(1 - \varepsilon)\rceil$ for $\varepsilon \in (0, 1)$, we can apply the estimate of Theorem 6, and obtain

$$E(M) \leqslant \left\lceil \frac{nG}{1-\varepsilon} \right\rceil$$

$$+ \sum_{k=K}^{\infty} 4N_{(1-nG/k)/6} \exp\left(-\frac{2}{9} k \left(\frac{1-nG/k}{G}\right)^2\right)$$

$$\leqslant \frac{nG}{1-\varepsilon} + 2$$

$$+ \int_K^{\infty} 4N_{(1-nG/u)/6} \exp\left(-\frac{2}{9} u \left(\frac{1-nG/u}{G}\right)^2\right) du$$

(because N_ε decreases as $\varepsilon\uparrow$)

$$= \frac{nG}{1-\varepsilon} + 2$$

$$+ \int_{\varepsilon^*}^{1} \frac{4nG}{(1-\nu)^2} N_{\nu/6} \exp\left(-\frac{2}{9} \frac{nG}{1-\nu} \left(\frac{\nu}{G}\right)^2\right) d\nu$$

(where ε^* satisfies $K = nG/(1-\varepsilon^*)$)

$$\leqslant \frac{nG}{1-\varepsilon} + 2$$

$$+ \int_{\varepsilon}^{1} \frac{4nG}{(1-\nu)^2} N_{\nu/6} \exp\left(-\frac{2}{9} \frac{nG}{1-\nu} \left(\frac{\nu}{G}\right)^2\right) d\nu$$

$$\leqslant \frac{nG}{1-\varepsilon} + 2$$

$$+ \int_{\varepsilon}^{1/2} \frac{4nG}{(1-\nu)^2} N_{\nu/6} \exp\left(-\frac{2}{9} \frac{nG}{1-\nu} \left(\frac{\nu}{G}\right)^2\right) d\nu$$

$$+ \int_{1/2}^{1} \frac{4nG}{(1-\nu)^2} N_{\nu/6} \exp\left(-\frac{2}{9} \frac{nG}{1-\nu} \left(\frac{\nu}{G}\right)^2\right) d\nu$$

$$\leqslant \frac{nG}{1-\varepsilon} + 2 + \int_{\varepsilon}^{1/2} 16nGN_{\nu/6} \exp\left(-\frac{4n\nu^2}{9G}\right) d\nu$$

$$+ \int_{1/2}^{1} \frac{4nG}{(1-\nu)^2} N_{1/12} \exp\left(\frac{n}{18G(1-\nu)}\right) d\nu$$

$$\leqslant \frac{nG}{1-\varepsilon} + 2 + 16nGN_{\varepsilon/6} \int_{\varepsilon}^{\infty} \exp\left(\frac{4n\nu^2}{9G}\right) d\nu$$

$$+ 4nGN_{1/12} \int_{2}^{\infty} \exp\left(-\frac{n\nu}{18G}\right) d\nu$$

$$\leqslant \frac{nG}{1-\varepsilon} + 2 + 36G^2 N_{\varepsilon/6} \exp\left(-\frac{4n\varepsilon^2}{9G}\right)$$

$$+ 72G^2 N_{1/12} \exp\left(-\frac{n}{9G}\right).$$

Finally, the minimization of the right-hand side with respect to ε is a simple exercise in analysis.

The expected size of the final data pool is roughly of the order of nG whenever the class has finite G and finite metric entropy. In fact, for most classes of densities, we have $N_u \leqslant \exp(C/u^b)$ for some constants C, $b > 0$ (see the next section), so that $E(M) - nG = O(n^{(b+1)/(b+2)})$. The richness of the class has a direct bearing upon the required storage.

9. EXAMPLES OF FAMILIES OF DENSITIES

Explicit estimates of the metric entropy have been obtained by Kolmogorov and Tikhomirov (1961), Lorentz (1966), Dudley (1974, 1984) and others. Families with finite metric entropy include those shown in Table I.

Table I shows that manageable classes of densities can be obtained by introducing smoothness conditions (such as in the first example), by drawing densities from a restricted parametric class (such as in the second example), or by imposing monotonicity conditions (as in the third and fourth examples). The various functions $c(\cdot)$ shown in the bounds can be obtained explicitly by analyzing the proofs of the original papers. For the Lipschitz class, see Kolmogorov and Tikhomirov, Devroye (1987) or Dudley (1984). For the monotone class, we refer to Dudley (1984), who reports computations done by Birgé. Finally, for the bounded variation class, and general smoothness classes, the reader can consult Clements (1963). We note that for the parametric class given above N_ε is polynomial in $1/\varepsilon$, whereas for the larger nonparametric classes, it is exponential in

Table I
Families With Finite Metric Entropy

F	Upper Bound for N_ε
All densities f on [0, 1] with $k-1$ absolutely continuous derivatives, and $f^{(k)}$ satisfying a Lipschitz condition with constants C, α: $\sup_{x,j\leqslant k} \lvert f^{(j)}(x)\rvert + \sup_{x\neq y} \frac{\lvert f^{(k)}(x) - f^{(k)}(y)\rvert}{\lvert x-y\rvert^{\alpha}} \leqslant C$	$\exp(c(\alpha + k, C)/\varepsilon^{1/(\alpha+k)})$
All gamma densities with shape parameter contained in a finite interval $[a, b]$ with $a > 0$	$c(a, b)/\varepsilon$
All nonincreasing nonnegative functions on [0, 1] bounded by M	$\exp(c(M)/\varepsilon)$
All densities on [0, 1] bounded by M and of total variation at most B, where B and M are finite constants	$\exp(c(M, B)/\varepsilon)$

$1/\varepsilon$. There are also classes that hover "in between" in richness, such as some classes of analytic functions, for which N_ε is exponential in $\log^2(1/\varepsilon)$ (see e.g., Lorentz or Kolmogorov and Tikhomirov).

For the monotone and bounded variation classes, we have $E(M) = nG + O(n^{2/3})$, while for the Lipschitz class, we obtain $E(M) = nG + O(n^{(1+\alpha+k)/(1+2(\alpha+k))})$.

ACKNOWLEDGMENT

I wish to thank the referees for their invaluable suggestions.

REFERENCES

ALEXANDER, K. S. 1984. Probability Inequalities for Empirical Processes and a Law of the Iterated Logarithm. *Ann. Prob.* **12,** 1041–1067.

BLUM, J. R. 1955. On the Convergence of Empiric Distribution Functions. *Ann. Math. Stat.* **26,** 527–529.

BRATLEY, P., B. L. FOX AND L. E. SCHRAGE. 1987. *A Guide to Simulation*, 2nd ed. Springer-Verlag, New York.

CLEMENTS, G. F. 1963. Entropies of Several Sets of Functions. *Pacific J. Math.* **13,** 1085–1097.

DEHARDT, J. 1971. Generalizations of the Glivenko-Cantelli Theorem. *Ann. Math. Stat.* **42,** 2050–2055.

DEHEUVELS, P., AND D. PFEIFER. 1986. A Semigroup Approach to Poisson Approximation. *Ann. Prob.* **14,** 663–676.

DEVROYE, L. 1982. Bounds for the Uniform Deviation of Empirical Measures. *J. Multivariate Anal.* **12,** 72–79.

DEVROYE, L. 1985. A Note on the L1 Consistency of Variable Kernel Estimates. *Ann. Stat.* **13,** 1041–1049.

DEVROYE, L. 1986a. Sample-Based Non-Uniform Random Variate Generation. In *Proceedings of the 1986 Winter Simulation Conference*, ed. J. Wilson, J. Henriksen, and S. Roberts, pp. 260–265, IEEE.

DEVROYE, L. 1986b. *Nonuniform Random Variate Generation*. Springer-Verlag, New York.

DEVROYE, L. 1987. *A Course in Density Estimation.* Birkhauser Verlag, Boston.

DOEBLIN, W. 1937. Exposé de la Théorie des Chaines Simples Constantes de Markov à un Nombre Fini d'États. *Revue Mathématique de l'Union Interbalkanique* **2,** 77–105.

DUDLEY, R. M. 1974. Metric Entropy of Some Classes of Sets With Differentiable Boundaries. *J. Approx. Theory* **10,** 227–236.

DUDLEY, R. M. 1984. Empirical Processes. Ecole de Probabilite de St.-Flour 1982, Lecture Notes in Mathematics 1097, Springer-Verlag, New York.

GINE, E., AND J. ZINN. 1984. Some Limit Theorems for Empirical Processes. *Ann. Prob.* **12,** 929–989.

GOLDSTEIN, S. 1979. Maximal Coupling. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **46,** 193–204.

GRIFFEATH, D. 1975. A Maximal Coupling for Markov Chains. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **31,** 95–106.

HOEFFDING, W. 1963. Probability Inequalities for Sums of Bounded Random Variables. *J. Am. Stat. Assoc.* **58,** 13–30.

JOHNSON, N. L., AND S. KOTZ. 1969. *Distributions in Statistics: Discrete Distributions.* John Wiley, New York.

KOLMOGOROV, A. N., AND V. M. TIKHOMIROV. 1961. ε-Entropy and ε-Capacity of Sets in Function Spaces. *Trans. Am. Math. Soc.* **17,** 277–364.

LECAM, L. 1960. An Approximation Theorem for the Poisson Binomial Distribution. *Pacific J. Math.* **10,** 1181–1197.

LORENTZ, G. G. 1966. *Approximation of Functions.* Holt, Rinehart & Winston, New York.

PETROV, V. V. 1975. *Sums of Independent Random Variables.* Springer-Verlag, Berlin.

PITMAN, J. W. 1974. Uniform Rates of Convergence for Markov Chain Transition Probabilities. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **29,** 193–227.

PITMAN, J. W. 1976. On Coupling of Markov Chains. *Zeitschrift fur Wahrscheinlichkeitstheorie und verwandte Gebiete* **35,** 315–322.

SCHEFFE, H. 1947. A Useful Convergence Theorem for Probability Distributions. *Ann. Math. Stat.* **18,** 434–458.

SCHMEISER, B. W., AND V. KACHITVICHYANUKUL. 1986. Correlation Induction Without the Inverse Transformation. *Proceedings of the 1986 Winter Simulation Conference*, Washington, D.C.

SERFLING, R. J. 1975. A General Poisson Approximation Theorem. *Ann. Prob.* **3,** 726–731.

THORISSON, H. 1986. On Maximal and Distributional Coupling. *Ann. Prob.* **14,** 873–876.

VAPNIK, V. N., AND A. YA. CHERVONENKIS. 1971a. On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities. *Theory Prob. Its Appl.* **16,** 264–280.

VAPNIK, V. N., AND A. YA. CHERVONENKIS. 1971b. Theory of Uniform Convergence of Frequencies of Events to Their Probabilities and Problems of Search for an Optimal Solution From Empirical Data. *Automation and Remote Control* **32,** 207–217.

VAPNIK, V. N., AND A. YA. CHERVONENKIS. 1981. Necessary and Sufficient Conditions for the Uniform Convergence of Means to Their Expectations. *Theory Prob. Its Appl.* **26,** 532–553.

WHEEDEN, R. L., AND A. ZYGMUND. 1977. *Measure and Integral.* Marcel Dekker, New York.

WHITT, W. 1976. Bivariate Distributions With Given Marginals. *Ann. Stat.* **4,** 1280–1289.

YOSIDA, K. 1980. *Functional Analysis.* Springer-Verlag, Berlin.

YUKICH, J. E. 1985. Laws of Large Numbers for Classes of Functions. *J. Multivariate Anal.* **17,** 245–260.