1 Partitions generated by random sampling

Given random variables $X_1, \ldots, X_n$, let $F_n$ be the empirical distribution function of $X_1, \ldots, X_n$:

$$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} 1_{[X_i \leq x]}.$$

**Proposition 1** (de Finetti extension). Given an infinite exchangeable sequence $(X_n)$ With $\mathbb{E}\{X_1\} < \infty$, let $F$ be the random probability distribution from de Finetti’s theorem. Then with probability one,

$$\sup_x (F(x) - F_n(x)) \to 0 \text{ as } n \to \infty.$$

Remark: the rate of convergence implied by “uniformly in $x$” may depend on $F$.

**Lemma 1** (Glivenko-Cantelli theorem). Let $G$ be a distribution function. If $(X_n)_{n \in \mathbb{N}}$ are iid with distribution $G$ and $\mathbb{E}\{X_1\} < \infty$ then with probability one,

$$\lim_{n \to \infty} \sup_x (G(x) - G_n(x)) = 0.$$

**Proof of Proposition assuming lemma.** We have

$$\mathbb{P}\{ \lim_{n \to \infty} \sup_x (F(x) - F_n(x)) = 0 \} = \mathbb{E}\{ \mathbb{P}\{ \lim_{n \to \infty} \sup_x (F(x) - F_n(x)) = 0 \} | F \}.$$ 

But given $F$, the $(X_n)$ are iid with distribution $F$, so the Glivenko-Cantelli theorem implies that the right-hand side above is just $\mathbb{E}\{1 | F\} = 1$.

**Proof of Glivenko-Cantelli theorem.** By the strong law of large numbers, $F_n(x) \to F(x)$ with probability one. The idea is to apply this convergence at increasingly closely spaced, finite subsets of the range $[0, 1]$ of $F$.

Due to lack of continuity, we also need to consider $F_n(x^-) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{[X_i < x]}$. Of course, we can also apply the SLLN to see that $F_n(x^-) \to F(x^-)$ with probability one.

Now fix $k \in \mathbb{N}$, and for $j \in [k - 1]$ let $x_{j,k} = \inf\{x : F(x) > j/k\}$. (Note that some of the $x_{j,k}$ could be equal.) Also, let $x_{0,k} = -\infty$, $x_{k,k} = \infty$. Then apply the SLLN at each of $x_{j,k}$,
We obtain that there is some (random) time \( N_k \) such that for all \( n \geq N_k \), and for all \( j \in \{0, 1, \ldots, k\} \),

\[
|F_n(x_{j,k}) - F(x_{j,k})| < \frac{1}{k} \quad \text{and} \quad |F_n(x_{j,k}^{-}) - F(x_{j,k}^{-})| < \frac{1}{k}.
\]

This shows that the value of \( F_n \) is close to that of \( F \) at the points \( x_{j,k} \). For the remaining points, fix any \( x \in \mathbb{R} \) with \( x \in (x_{j-1,k}, x_{j,k}) \), for some \( j \in [k] \). If \( n \geq N_k \) we then have

\[
F_n(x) \leq F_n(x_{j,k}^{-}) \leq F(x_{j,k}^{-}) + \frac{1}{k} \quad \text{since } n \geq N_k.
\]

Since \( F \) is monotone nondecreasing, \( F(x_{j,k}^{-}) \leq F(x_{j-1,k}) + k^{-1} \), so

\[
F_n(x) \leq F(x_{j-1,k}) + \frac{2}{k} \leq F(x) + \frac{2}{k} \quad \text{since } x_{j-1,k} < x.
\]

Similarly,

\[
F_n(x) \geq F_n(x_{j,k-1}^{-}) \geq F(x_{j,k-1}^{-}) - \frac{1}{k} \quad \text{since } n \geq N_k
\]

\[
\geq F(x_{j,k}^{-}) - \frac{2}{k} \quad \text{by monotonicity}
\]

\[
\geq F(x) - \frac{2}{k} \quad \text{since } x_{j,k} > x.
\]

Thus, for all \( n \geq N_k \), \( \sup_x (F_n(x) - F(x)) \leq 2/k \). Since \( k \) was arbitrary, the conclusion follows.

Given a probability distribution \( F \) on \( \mathbb{R} \) and \( x \in \mathbb{R} \), \( c \in (0, 1) \), say that \( F \) has an atom of magnitude \( c \) at \( x \) if \( F(x^{-}) = F(x) - c \). We write \( (P_i^1, i \geq 1) = (P_i^1(F), i \geq 1) \) for the sequence of ranked atoms of \( F \), i.e. the atoms of \( F \) listed in decreasing order of magnitude.

Given \( \Pi_{\infty} \), an e.r.p. of \( \mathbb{N} \), by a diagonalization argument (special case of Skorohod’s theorem), we can assume that for all \( m \leq n \), we actually have \( \Pi_{m,n} \overset{a.s.}{=} \Pi_m \). Write \( (A_{n,i}^1) \) for the parts of \( \Pi_n \) in decreasing order of size (ties broken lexicographically), and write \( (N_{n,i}^1, i \geq 1) \) for the corresponding sizes (with \( N_{n,i}^1 = 0 \) if \( \Pi_n \) has less than \( i \) parts). This is all setup for the below theorem.
Theorem 1 (Kingman’s paintbox representation). Let $\Pi_{\infty}$ be an e.r.p. of $\mathbb{N}$. Then there is a random vector $P_i = (P_i, i \geq 1)$ such that for each $i \in \mathbb{N}^{>0}$,

$$\lim_{n \to \infty} \frac{N_{n,i}^{\downarrow}}{n} \overset{a.s.}{=} P_i^{\downarrow}.$$ 

Then given $P_i$, $\Pi_{\infty}$ is distributed as $\Pi_{\infty}((X_n)_{n \in \mathbb{N}})$, where the $X_n$ are iid with distribution function $F = F(P_i^{\downarrow})$.

Proof. Since we are working in a space where $\Pi_{m,n} \overset{a.s.}{=} \Pi_m$ for all $m \leq n$, with probability one $\Pi_{\infty}$ defines a partition of $\mathbb{N} = \{A_{\infty,1}, A_{\infty,2}, \ldots\}$, listed in order of appearance. (Put $n$ into whatever part it appears in in $\Pi_n$.)

Now let $(U_n)_{n \geq 1}$ be a sequence of independent Uniform$[0,1]$ random variables. Define a sequence $(X_m)_{m \geq 1}$ by setting $X_m = U_j$ if $m \in A_{j,\infty}$. Then $(X_m)_{m \geq 1}$ is exchangeable since for each $m$, $\Pi_m$ is exchangeable.

Also, for $n \geq 1$ and $i \geq 1$, write $\hat{U}_{n,i} = U_j$ if $A_{n,i} = A_j$. Then the number of $m \in [n]$ for which $X_m = \hat{U}_{n,i}$ is precisely $N_{n,i}^{\downarrow}$.

Write $F$ for the distribution function from Proposition 1, and $(P_i^{\downarrow})_{i \geq 1}$ for its ranked atoms. (We assume the atoms are all distinct for simplicity.) By Proposition 1, with probability one,

$$\sup_x (F(x) - F_n(x)) \to 0 \text{ as } n \to \infty.$$ (1)

Here $F$ has $F(0^-) = 0$, $F(1) = 1$, and for $0 \leq u \leq 1$,

$$F_n(u) = \frac{1}{n} \sum_{m=1}^{n} 1_{[X_m \leq u]} = \sum_{i=1}^{\infty} \frac{N_{n,i}^{\downarrow}}{n} 1_{[\hat{U}_{n,i} \leq u]}.$$ 

Given $F$, the $X_n$ are independent with distribution $F$. Since all the $X_j$ take values from among $(U_i)_{i \geq 1}$, there must be some permutation $\sigma : \mathbb{N} \to \mathbb{N}$ (which is a function of $F$) such that for all $i \geq 1$

$$\mathbb{P}\{X_1 = U_{\sigma(i)}\} = P_i^{\downarrow}.$$ 

Now fix $i \geq 1$ and let $\delta = \min(P_i^{\downarrow} - P_{i-1}^{\downarrow}, P_{i+1}^{\downarrow} - P_i^{\downarrow})$. By (1), with probability one there is $n_0$ such that for all $n \geq n_0$, we have $\sup_x (F(x) - F_n(x)) < \delta/4$. Then in particular, for all

3
\[ j < i, \]

\[ |\{ k \leq n : X_k = U_{\sigma(j)} \}| = F_n(U_{\sigma(j)}) - F_n(U_{\sigma(j)}^-) \]
\[ > F(U_{\sigma(j)}) - F(U_{\sigma(j)}^-) - \frac{\delta}{2} \]
\[ = P_j^1 - \frac{\delta}{2} \]
\[ \geq P_{i-1}^1 - \frac{\delta}{2} \]
\[ > P_i^1 + \frac{\delta}{2} \]
\[ = F(U_{\sigma(i)}) - F(U_{\sigma(i)}^-) + \frac{\delta}{2} \]
\[ > F_n(U_{\sigma(i)}) - F_n(U_{\sigma(i)}^-) \]
\[ = |\{ k \leq n : X_k = U_{\sigma(i)} \}|. \]

A similar argument shows that for all \( n \geq n_0 \) and \( j > i \),

\[ |\{ k \leq n : X_k = U_{\sigma(j)} \}| < |\{ k \leq n : X_k = U_{\sigma(i)} \}|. \]

It follows that for all \( n \geq n_0 \), we have \( |\{ k \leq n : X_k = U_{\sigma(i)} \}| = N_{n,i}^1 \), and so

\[ \lim_{n \to \infty} \frac{N_{n,i}^1}{n} \overset{a.s.}{=} \lim_{n \to \infty} \frac{|\{ k \leq n : X_k = U_{\sigma(i)} \}|}{n} \overset{a.s.}{=} P_i^1, \]

the last holding, conditional on \( F \), by the law of large numbers, and then unconditionally by taking the expectation over \( F \).