18.175 Theory of Probability
Fall 2008

Section 5

Bernstein Polynomials. Hausdorff and de Finetti theorems.

Let us look at some applications related to the law of large numbers. Consider an i.i.d. sequence of real valued r.v. (X_i) with distribution $\mathbb{P}_\theta$ from a family of distributions parametrized by $\theta \in \Theta \subseteq \mathbb{R}$ such that

$$\mathbb{E}_\theta X_i = \theta, \sigma^2(\theta) := \mathrm{Var}_\theta(X_i) \le K < +\infty.$$

Let $\bar{X}_n = \frac{1}{n}\sum_{i\le n} X_i$. The following holds.

Theorem 6 *If* $u : \mathbb{R} \to \mathbb{R}$ *is uniformly continuous and bounded then* $\mathbb{E}_\theta u(\bar{X}_n) \to u(\theta)$ *uniformly over* Θ.

Proof. For any $\varepsilon > 0$,

$$\begin{aligned} |\mathbb{E}_\theta u(\bar{X}_n) - u(\theta)| &\le \mathbb{E}_\theta |u(\bar{X}_n) - u(\theta)| \\ &= \mathbb{E}_\theta |u(\bar{X}_n) - u(\theta)| \Big(\mathrm{I}(|\bar{X}_n - \theta| \le \varepsilon) + \mathrm{I}(|\bar{X}_n - \theta| > \varepsilon) \Big) \\ &\le \max_{|x-\theta|\le\varepsilon} |u(x) - u(\theta)| + 2\max_x |u(x)| \mathbb{P}_\theta(|\bar{X}_n - \theta| > \varepsilon) \\ &\le \delta(\varepsilon) + 2\|u\|_\infty \frac{1}{\varepsilon^2} \mathbb{E}_\theta(\bar{X}_n - \theta)^2 \le \delta(\varepsilon) + \frac{2\|u\|_\infty K}{n\varepsilon^2}, \end{aligned}$$

where $\delta(\varepsilon)$ is the modulus of continuity of u. Letting $\varepsilon = \varepsilon_n \to 0$ so that $n\varepsilon_n^2 \to \infty$ finishes the proof. □

Example. Let (X_i) be i.i.d. with Bernoulli distribution $B(\theta)$ with probability of success $\theta \in [0, 1]$, i.e.

$$\mathbb{P}_\theta(X_i = 1) = \theta, \ \ \mathbb{P}_\theta(X_i = 0) = 1 - \theta,$$

and let $u : [0, 1] \to \mathbb{R}$ be continuous. Then, by the above Theorem, the following *Bernstein polynomials*

$$B_n(\theta) := \mathbb{E}_\theta u(\bar{X}_n) = \sum_{k=0}^n u\Big(\frac{k}{n}\Big) \mathbb{P}_\theta\Big(\sum_{i=1}^n X_i = k\Big) = \sum_{k=0}^n u\Big(\frac{k}{n}\Big) \binom{n}{k} \theta^k (1-\theta)^{n-k} \to u(\theta)$$

uniformly on $[0, 1]$.

Example. Let (X_i) have Poisson distribution $\Pi(\theta)$ with intensity parameter $\theta > 0$ defined by

$$\mathbb{P}_\theta(X_i = k) = \frac{\theta^k}{k!} e^{-\theta} \text{ for integer } k \ge 0.$$

Then it is well known (and easy to check) that $\mathbb{E}_\theta X_i = \theta, \sigma^2(\theta) = \theta$ and the sum $X_1 + \ldots + X_n$ has Poisson distribution $\Pi(n\theta)$. If u is bounded and continuous on $[0, +\infty)$ then

$$\mathbb{E}_\theta u(\bar{X}_n) = \sum_{k=0}^\infty u\Big(\frac{k}{n}\Big) \mathbb{P}_\theta\Big(\sum_{i=1}^n X_i = k\Big) = \sum_{k=0}^\infty u\Big(\frac{k}{n}\Big) \frac{(n\theta)^k}{k!} e^{-n\theta} \to u(\theta)$$

uniformly on compact sets.

□

Moment problem. Consider a random variable $X \in [0,1]$ and let $\mu_k = \mathbb{E}X^k$ be its moments. Given a sequence $(c_0, c_1, c_2, \ldots)$ let us define a sequence of increments by $\Delta c_k = c_{k+1} - c_k$. Then

$$-\Delta\mu_k = \mu_k - \mu_{k+1} = \mathbb{E}(X^k - X^{k+1}) = \mathbb{E}X^k(1-X),$$

$$(-\Delta)(-\Delta\mu_k) = (-1)^2\Delta^2\mu_k = \mathbb{E}X^k(1-X) - \mathbb{E}X^{k+1}(1-X) = \mathbb{E}X^k(1-X)^2$$

and by induction

$$(-1)^r\Delta^r\mu_k = \mathbb{E}X^k(1-X)^r.$$

Clearly, $(-1)^r\Delta^r\mu_k \geq 0$ since $X \in [0,1]$. If u is a continuous function on $[0,1]$ and B_n is its corresponding Bernstein polynomial then

$$\mathbb{E}B_n(X) = \sum_{k=0}^{n} u\Big(\frac{k}{n}\Big)\binom{n}{k}\mathbb{E}X^k(1-X)^{n-k} = \sum_{k=0}^{n} u\Big(\frac{k}{n}\Big)\binom{n}{k}(-1)^{n-k}\Delta^{n-k}\mu_k.$$

Since $B_n(X)$ converges uniformly to $u(X)$, $\mathbb{E}B_n(X)$ converges to $\mathbb{E}u(X)$. Let us define

$$p_k^{(n)} = \binom{n}{k}(-1)^{n-k}\Delta^{n-k}\mu_k \geq 0, \ \sum_{k=0}^{n} p_k^{(n)} = 1 \text{ (take } u = 1).$$

We can think of $p_k^{(n)}$ as the distribution of a r.v. $X^{(n)}$ such that

$$\mathbb{P}\Big(X^{(n)} = \frac{k}{n}\Big) = p_k^{(n)}. \tag{5.0.1}$$

We showed that

$$\mathbb{E}B_n(X) = \mathbb{E}u\big(X^{(n)}\big) \to \mathbb{E}u(X)$$

for any continuous function u. We will later see that by definition this means that $X^{(n)}$ converges to X in distribution. Given the moments of a r.v. X, this construction allows us to approximate the distribution of X and expectation of $u(X)$.

□

Next, given a sequence (μ_k), when is it the sequence of moments of some $[0,1]$ valued r.v. X? By the above, it is necessary that

$$\mu_k \geq 0, \mu_0 = 1 \text{ and } (-1)^r\Delta^r\mu_k \geq 0 \text{ for all } k, r. \tag{5.0.2}$$

It turns out that this is also sufficient.

Theorem 7 *(Hausdorff) There exists a r.v. $X \in [0,1]$ such that $\mu_k = \mathbb{E}X^k$ iff (5.0.2) holds.*

Proof. The idea of the proof is as follows. If μ_k are the moments of the distribution of some r.v. X, then the discrete distributions defined in (5.0.1) should approximate it. Therefore, our goal will be to show that condition (5.0.2) ensures that $(p_k^{(n)})$ is indeed a distribution and then show that the moments of (5.0.1) converge to μ_k. As a result, any limit of these distributions will be a candidate for the distribution of X.

First of all, let us express μ_k in terms of $(p_k^{(n)})$. Since $\Delta\mu_k = \mu_{k+1} - \mu_k$ we have the following inversion formula:

$$\begin{aligned}\mu_k &= \mu_{k+1} - \Delta\mu_k = (\mu_{k+2} - \Delta\mu_{k+1}) + (-\Delta\mu_{k+1} + \Delta^2\mu_k) \\ &= \mu_{k+2} - 2\Delta\mu_{k+1} + \Delta^2\mu_k = \sum_{j=0}^{r}\binom{r}{j}(-1)^{r-j}\Delta^{r-j}\mu_{k+j},\end{aligned}$$

by induction. Take $r = n-k$. Then

$$\mu_k = \sum_{j=0}^{n-k} \frac{\binom{n-k}{j}}{\binom{n}{k+j}} \binom{n}{k+j} (-1)^{n-(k+j)} \Delta^{n-(k+j)} \mu_{k+j} = \sum_{j=0}^{n-k} \frac{\binom{n-k}{j}}{\binom{n}{k+j}} p^{(n)}_{k+j}.$$

We have

$$\frac{\binom{n-k}{j}}{\binom{n}{k+j}} = \frac{(n-k)!}{j!(n-k-j)!} \frac{(k+j)!(n-k-j)!}{n!} = \frac{\binom{k+j}{k}}{\binom{n}{k}}$$

so that

$$\mu_k = \sum_{j=0}^{n-k} \frac{\binom{k+j}{k}}{\binom{n}{k}} p^{(n)}_{k+j} = \sum_{m=k}^{n} \frac{\binom{m}{k}}{\binom{n}{k}} p^{(n)}_m.$$

By (5.0.2), $p^{(n)}_m \geq 0$ and $\sum_{m\leq n} p^{(n)}_m = \mu_0 = 1$ so we can consider a r.v. $X^{(n)}$ such that

$$\mathbb{P}\Big(X^{(n)} = \frac{m}{n}\Big) = p^{(n)}_m \text{ for } 0 \leq m \leq n.$$

We have

$$\begin{aligned} \mu_k &= \sum_{m=k}^{n} \frac{\binom{m}{k}}{\binom{n}{k}} p^{(n)}_m = \sum_{m=k}^{n} \frac{m(m-1)\cdots(m-k+1)}{n(n-1)\cdots(n-k+1)} p^{(n)}_m = \sum_{m=k}^{n} \frac{\frac{m}{n}(\frac{m}{n}-\frac{1}{n})\cdots(\frac{m}{n}-\frac{k+1}{n})}{1(1-\frac{1}{n})\cdots(1-\frac{k+1}{n})} p^{(n)}_m \\ &\overset{n\to\infty}{\approx} \sum_{m=0}^{n} \Big(\frac{m}{n}\Big)^k p^{(k)}_m = \mathbb{E}\Big(X^{(n)}\Big)^k \overset{n\to\infty}{\longrightarrow} \mu_k. \end{aligned}$$

Any continuous function u can be approximated by (for example, Bernstein) polynomials so the limit $\lim_{n\to\infty} \mathbb{E}u\big(X^{(n)}\big)$ exists. By selection theorem that we will prove later in the course, one can choose a subsequence $X^{(n_i)}$ that converges to some r.v. X in distribution and, as a result,

$$\mathbb{E}\Big(X^{(n_i)}\Big)^k \to \mathbb{E}X^k = \mu_k,$$

which means that μ_k are the moments of X.

□

de Finetti's theorem. Consider an *exchangeable* sequence $X_1, X_2, \ldots, X_n, \ldots$ of Bernoulli random variables which means that for any $n \geq 1$ the probability

$$\mathbb{P}(X_1 = x_1, ..., X_n = x_n)$$

depends only on $x_1 + \ldots + x_n$, i.e. it does not depend on the order of 1's or 0's. Another way to say this is that for any $n \geq 1$ and any permutation π of $1, \ldots, n$ the distribution of $(X_{\pi(1)}, \ldots, X_{\pi(n)})$ does not depend on π. Then the following holds.

Theorem 8 *(de Finetti) There exists a distribution F on $[0,1]$ such that*

$$p_k := \mathbb{P}(X_1 + \ldots + X_n = k) = \int_0^1 \binom{n}{k} x^k (1-x)^{n-k} dF(x).$$

This means that to generate such exchangeable sequence we can first pick $x \in [0,1]$ from distribution F and then generate a sequence of i.i.d Bernoulli random variables with probability of success x.
Proof. Let $\mu_0 = 1$ and for $k \geq 1$ define

$$\mu_k = \mathbb{P}(X_1 = 1, ..., X_k = 1). \tag{5.0.3}$$

We have

$$\begin{aligned}\mathbb{P}(X_1 = 1, ..., X_k = 1, X_{k+1} = 0) &= \mathbb{P}(X_1 = 1, ..., X_k = 1) \\ &- \mathbb{P}(X_1 = 1, ..., X_k = 1, X_{k+1} = 1) \\ &= \mu_k - \mu_{k+1} = -\Delta\mu_k.\end{aligned}$$

Next, using exchangeability

$$\begin{aligned}\mathbb{P}(X_1 = 1, ..., X_k = 1, X_{k+1} = 0, X_{k+2} = 0) &= \mathbb{P}(X_1 = 1, ..., X_k = 1, X_{k+1} = 0) \\ &- \mathbb{P}(X_1 = 1, ..., X_k = 1, X_{k+1} = 0, X_{k+2} = 1) \\ &= -\Delta\mu_k - (-\Delta\mu_{k+1}) = \Delta^2\mu_k.\end{aligned}$$

Similarly, by induction,

$$\mathbb{P}(X_1 = 1, ..., X_k = 1, X_{k+1} = 0, ..., X_n = 0) = (-1)^{n-k}\Delta^{n-k}\mu_k \geq 0.$$

By the Hausdorff theorem, $\mu_k = \mathbb{E}X^k$ for some r.v. $X \in [0, 1]$ and, therefore,

$$\begin{aligned}\mathbb{P}(X_1 = 1, ..., X_k = 1, X_{k+1} = 0, ..., X_n = 0) &= (-1)^{n-k}\Delta^{n-k}\mu_k \\ &= \mathbb{E}X^k(1 - X)^{n-k} = \int_0^1 x^k(1 - x)^{n-k}dF(x).\end{aligned}$$

Since, by exchangeability, changing the order of 1's and 0's does not affect the probability, we get

$$\mathbb{P}(X_1 + \ldots + X_n = k) = \int_0^1 \binom{n}{k} x^k(1 - x)^{n-k}dF(x).$$

□

Example. (Polya urn model). Suppose we have b blue and r red balls in the urn. We pick a ball

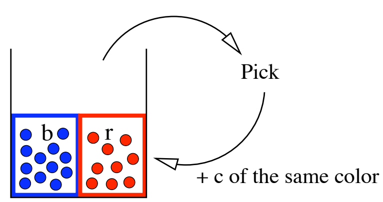

Figure 5.1: Polya urn model.

randomly and return it with c balls of the same color. Consider r.v.s

$$X_i = \begin{cases} 1 & \text{if the } i\text{th ball picked is blue} \\ 0 & \text{otherwise.}\end{cases}$$

X_i's are not independent but exchangeable. For example,

$$\mathbb{P}(bbr) = \frac{b}{b+r} \times \frac{b+c}{b+r+c} \times \frac{r}{b+r+2c}, \; \mathbb{P}(brb) = \frac{b}{b+r} \times \frac{r}{b+r+c} \times \frac{b+r}{b+r+2c}$$

are equal. To identify the distribution F in de Finetti's theorem, let us look at its moments μ_k in (5.0.3),

$$\mu_k = \mathbb{P}(\underbrace{b \ldots b}_{k \text{ times}}) = \frac{b}{b+r} \times \frac{b+c}{b+r+c} \times \cdots \times \frac{b+(k-1)c}{b+r+(k-1)c}.$$

One can recognize or easily check that μ_k are the moments of Beta(α, β) distribution with the density

$$\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

on $[0, 1]$ with parameters $\alpha = b/c, \beta = r/c$. By de Finetti's theorem, we can generate X_i's by first picking x from distribution $\text{Beta}\big(b/c, r/c\big)$ and then generating i.i.d. Bernoulli (X_i)'s with probability of success x. By strong law of large numbers, the proportion of blue balls in the first n repetitions will converge to this probability of success x, i.e. in the limit it will be random with Beta distribution. This example will come up once more when we talk about convergence of martingales.

□