Polynomial approximations to continuous functions and stochastic compositions

Takis Konstantopoulos * Linglong Yuan † Michael A. Zazanis ‡

Abstract

This paper presents a stochastic approach to theorems concerning the behavior of iterations of the Bernstein operator B_n taking a continuous function $f \in C[0,1]$ to a degree-n polynomial when the number of iterations k tends to infinity and n is kept fixed or when n tends to infinity as well. In the first instance, the underlying stochastic process is the so-called Wright-Fisher model, whereas, in the second instance, the underlying stochastic process is the Wright-Fisher diffusion. Both processes are probably the most basic ones in mathematical genetics. By using Markov chain theory and stochastic compositions, we explain probabilistically a theorem due to Kelisky and Rivlin, and by using stochastic calculus we compute a formula for the application of B_n a number of times $k = k(n)$ to a polynomial f when $k(n)/n$ tends to a constant.

1 Introduction

About 100 years ago, Bernstein [3] introduced a concrete sequence of polynomials approximating a continuous function on a compact interval. That polynomials are dense in the set of continuous functions was shown by Weierstrass [25], but Bernstein was the first to give a concrete method, one that has withstood the test of time. We refer to [20] for a history of approximation theory, including *inter alia* historical references to Weierstrass' life and work and to the subsequent work of Bernstein. Bernstein's approach was probabilistic and is nowadays included in numerous textbooks on probability theory, see, e.g., [21, p. 54] or [4, Theorem 6.2].

Several years after Bernstein's work, the nowadays known as Wright-Fisher stochastic model was introduced and proved to be a founding one for the area of quantitative genetics. The work was done in the context of Mendelian genetics by Ronald A. Fisher [11, 10] and Sewall Wright [26].

This paper aims to explain the relation between the Wright-Fisher model and the Bernstein operator B_n, that takes a function $f \in C[0,1]$ and outputs a degree-n approximating polynomial. Bernstein's original proof was probabilistic. It is thus natural to expect that subsequent properties of B_n can also be explained via probability theory. In doing so, we shed new light to what happens when we apply the Bernstein operator B_n a large number of times k to a function f. In fact, things become particularly interesting when k and n converge simultaneously to ∞. This convergence can be explained by means of the original Wright-Fisher model as well as a continuous-time approximation to it known as Wright-Fisher diffusion.

*takiskonst@gmail.com; Department of Mathematics, Uppsala University, SE-751 06 Uppsala, Sweden; the work of this author was supported by Swedish Research Council grant 2013-4688

†yuanlinglongcn@gmail.com; Department of Mathematics, Uppsala University, SE-751 06 Uppsala, Sweden

‡zazanis@aueb.gr; Department of Statistics, 76 Patission St., Athens University of Economics, Athens 104 34, Greece

Our paper was inspired by the Monthly paper of Abel and Ivan [1] that gives a short proof of the Kelisky and Rivlin theorem [17] regarding the limit of the iterates of B_n when n is fixed. We asked what is the underlying stochastic phenomenon that explains this convergence and found that it is the composition of independent copies of the empirical distribution function of n i.i.d. uniform random variables. The composition turns out to be precisely the Wright-Fisher model. Being a Markov chain with absorbing states, 0 and 1, its distributional limit is a random variable that takes values in $\{0, 1\}$; whence the Kelisky and Rivlin theorem [17].

Composing stochastic processes is in line with the first author's current research interests [6]. Indeed, such compositions often turn out to have interesting, nontrivial, limits [5]. Stochastic compositions become particularly interesting when they explain some natural mathematical or physical principles. This is what we do, in a particular case, in this paper. Besides giving fresh proofs to some phenomena, stochastic compositions help find what questions to ask as well.

We will specifically provide probabilistic proofs for a number of results associated to the Bernstein operator (1). First, we briefly recall Bernstein's probabilistic proof (Theorem 1) that says that $B_n f$ converges uniformly to f as the degree n converges to infinity. Second, we look at iterates B_n^k of B_n, meaning that we compose B_n k times with itself and give a probabilistic proof of the Kelisky and Rivlin theorem stating that $B_n^k f$ converges to $B_1 f$ as the number of iterations k tends to infinity (Theorem 2). Third, we exhibit, probabilistically, a geometric rate of convergence to the Kelisky and Rivlin theorem (Proposition 1). Fourth, we examine the limit of $B_n^k f$ when both n and k converge to infinity in a way that k/n converges to a constant (Theorem 3) and show that probability theory gives us a way to prove and set up computation methods for the limit for "simple" functions f such as polynomials (Proposition 2). A crucial step is the so-called Voronovskaya's theorem (Theorem 4) which gives a rate of convergence to Bernstein's theorem but also provides the reason why the Wright-Fisher model converges to the Wright-Fisher diffusion; this is explained in Section 5.

Regarding notation, we let $C[0, 1]$ be the set of continuous functions $f : [0, 1] \to \mathbb{R}$, and $C^2[0, 1]$ the set of functions having a continuous second derivative f'', including the boundary points, so $f''(0)$ (respectively, $f''(1)$) is interpreted as derivative from the right (respectively, left). For a bounded function $f : [0, 1] \to \mathbb{R}$, we denote by $\|f\|$ the quantity $\sup_{0 \le x \le 1} |f(x)|$.

2 Recalling Bernstein's theorem

The Bernstein operator B_n maps any function $f : [0, 1] \to \mathbb{R}$ into the polynomial

$$B_n f(x) := \sum_{j=0}^{n} \binom{n}{j} x^j (1 - x)^{n-j} f\left(\frac{j}{n}\right). \tag{1}$$

We are mostly interested in viewing B_n as an operator on $C[0, 1]$. Bernstein's theorem is:

Theorem 1 (Bernstein, 1912). *If $f \in C[0, 1]$ then $B_n f$ converges uniformly to f:*

$$\lim_{n\to\infty} \max_{0\le x\le 1} |B_n f(x) - f(x)| = 0.$$

The proof of this theorem is elementary if probability theory is used and goes like this. Let $X_1, \ldots, X_n$ be independent Bernoulli random variables with $\mathbb{P}(X_i = 1) = x$, $\mathbb{P}(X_i = 0) = 1 - x$

for some $0 \le x \le 1$. If S_n denotes the number of variables with value 1 then S_n has a binomial distribution:

$$\mathbb{P}(S_n = j) = \binom{n}{j} x^j (1-x)^{n-j}, \quad j = 0, 1, \dots n. \tag{2}$$

Therefore

$$\mathbb{E} f(S_n/n) = \sum_{j=0}^{n} f(j/n) \mathbb{P}(S_n = j) = B_n f(x). \tag{3}$$

Now let

$$m(\varepsilon) := \max_{|x-y| \le \varepsilon} |f(x) - f(y)|, \quad 0 < \varepsilon < 1.$$

Since f is continuous on the compact set $[0, 1]$ it is also uniformly continuous and so $m(\varepsilon) \to 0$ as $\varepsilon \to 0$. Let A be the event that $|S_n/n - x| \le \varepsilon$ and $\mathbf{1}_A$ the indicator of A (a function that is 1 on A and 0 on its complement). We then write

$$\mathbb{E}|f(S_n/n) - f(x)| = \mathbb{E}\left(|f(S_n/n) - f(x)|\mathbf{1}_A\right) + \mathbb{E}\left(|f(S_n/n) - f(x)|\mathbf{1}_{A^c}\right) \le m(\varepsilon) + 2\|f\|\, \mathbb{P}(A^c).$$

By Chebyshev's inequality,

$$\mathbb{P}(A^c) = \mathbb{P}(|S_n - nx| \ge n\varepsilon|) \le (n\varepsilon)^{-2}\, \mathbb{E}(S_n - nx)^2 = (n\varepsilon)^{-2}\, nx(1-x) \le \frac{1}{4}\varepsilon^{-2} n^{-1}.$$

Therefore,

$$\mathbb{E}|f(S_n/n) - f(x)| \le m(\varepsilon) + \frac{\|f\|}{2\varepsilon^2 n}.$$

Letting $n \to \infty$ the last term goes to 0 and letting $\varepsilon \to 0$ the first term vanishes too, thus establishing the theorem.

***Remark* 1.** A variant of Bernstein's theorem due to Marc Kac [14] gives better estimate if f is Lipschitz or, more generally, Hölder continuous. Indeed, if f satisfies $|f(x) - f(y)| \le C|x - y|^\alpha$ for some $0 < \alpha \le 1$ then,

$$\mathbb{E}|f(S_n/n) - f(x)| \le C\, \mathbb{E}|S_n/n - x|^\alpha \le C\left(\mathbb{E}|S_n/n - x|^2\right)^{\alpha/2} = C(n^{-1}x(1-x))^{\alpha/2} \le \frac{C\, 2^{-\alpha}}{n^{\alpha/2}},$$

where the first inequality used the Hölder continuity of f, while the second used Jensen's inequality twice; indeed, if Z is a positive random variable then $\mathbb{E}Z^\beta \le (\mathbb{E}Z)^\beta$, by the concavity of the function $z \mapsto z^\beta$ when $0 < \beta < 1$.

***Remark* 2.** Hölder continuous functions with small $\alpha < 1$ are "rough" functions. The previous remark tells us that we may not have a good rate of convergence for these functions. On the other hand, if f is smooth can we expect a good rate of convergence? A simple calculation with $f(x) = x^2$ shows that $B_n f(x) = \mathbb{E}(S_n/n)^2 = (\mathbb{E}(S_n/n))^2 + \mathrm{var}(S_n/n) = x^2 + \frac{1}{n}x(1-x)$. Excluding the trivial case $f(x) = ax + b$ (the only functions f for which $B_n f = f$), can the rate of convergence be better than $1/n$ for some smooth function f? No, and this is due to Voronovskaya's theorem (Theorem 4 in Section 7).

***Remark* 3** (Some properties of the Bernstein operator)**.**
(i) B_n is an increasing operator: If $f \le g$ then $B_n f \le B_n g$. (Proof: $B_n f$ is an expectation; see (3).)
(ii) If f is a convex function then $B_n f \ge f$. Indeed, $B_n f(x) = \mathbb{E}f(S_n/n) \ge f(\mathbb{E}(S_n/n))$, by Jensen's inequality, and, clearly, $\mathbb{E}(S_n/n) = x$.
(iii) If f is a convex function than $B_n f$ is a also convex. See Lemma 2 in Section 3 for a proof.

3 Iterating Bernstein operators

Let $B_n^2 := B_n \circ B_n$ be the composition of B_n with itself and, similarly, let $B_n^k := B_n \circ \cdots \circ B_n$ (k times). Abel and Ivan [1] give a short proof of the following.

Theorem 2 (Kelisky and Rivlin, 1967). *For fixed $n \in \mathbb{N}$, and any function $f : [0,1] \to \mathbb{R}$,*

$$\lim_{k\to\infty} \max_{0\le x\le 1} |B_n^k f(x) - f(0) - (f(1) - f(0))x| = 0.$$

***Remark* 4.** Note that this says that $B_n^k f(x) \to B_1 f(x)$, as $k \to \infty$, uniformly in $x \in [0,1]$. If f is convex by Remark 3(ii), $B_n f \ge f$. By Remark 3(i), $B_n^k f(x)$ is increasing in k and hence the limit of Theorem 2 is actually achieved by an increasing sequence of convex functions (Remark 3(iii)).

A probabilistic proof for Theorem 2. To prepare the ground, we construct the earlier Bernoulli random variables in a different way. We take $U_1, \dots, U_n$ to be independent random variables, all uniformly distributed on the interval $[0,1]$, and their *empirical distribution function*

$$G_n(x) := \frac{1}{n}\sum_{j=1}^{n} \mathbf{1}_{U_j \le x} = \frac{1}{n} S_n(x), \quad 0 \le x \le 1. \tag{4}$$

We shall think of G_n as a random function. Note that, for each x, $S_n(x)$ has the binomial distribution (2). The advantage of the current representation is that x, instead of being a parameter of the probability distribution, is now an explicit parameter of the new random object $G_n(x)$. We are allowed to (and we will) pick a sequence of independent copies $G_n^1, G_n^2, \dots$ of G_n. For a positive integer k let

$$H_n^k := G_n^k \circ G_n^{k-1} \circ \cdots \circ G_n^1 \tag{5}$$

be the composition of the first k random functions. So H_n^k is itself a random function. By using the independence and the definition of B_n we have that (See also Section A1 in the Appendix)

$$\mathbb{E} f(H_n^k(x)) = B_n^k f(x), \quad 0 \le x \le 1, \tag{6}$$

for any function f. Hence the limit over $k \to \infty$ of the right-hand side is the expectation of the limit of the random variable $f(H_n^k(x))$, if this limit exists. (We make use of the fact that (6) is a finite sum!) To see that this is the case, we fix n and x and consider the sequence

$$H_n^k(x), \quad k = 1, 2, \dots$$

with values in

$$\mathbb{I}_n := \left\{0, \frac{1}{n}, \dots, \frac{n-1}{n}, 1\right\}.$$

We observe that this has the Markov property,[1] namely, $H_n^{k+1}(x)$ is independent of $H_n^1(x), \dots, H_n^{k-1}(x)$, conditional on $H_n^k(x)$. By (2), the one-step transition probability of this Markov chain is

$$p\left(\frac{i}{n}, \frac{j}{n}\right) := \mathbb{P}\left(H_n^{k+1}(x) = \frac{j}{n} \mid H_n^k(x) = \frac{i}{n}\right) = \binom{n}{j}\left(\frac{i}{n}\right)^j \left(1 - \frac{i}{n}\right)^{n-j}. \tag{7}$$

[1] Admittedly, it is a bit unconventional to use an upper index for the time parameter of a Markov chain but, in our case, we keep it this way because it appears naturally in the composition operation.

Since $p(0,0) = p(1,1) = 1$, states 0 and 1 are absorbing, whereas for any $x \in \mathbb{I}_n \setminus \{0,1\}$, we have $p(x,y) > 0$ for all $y \in \mathbb{I}_n$. Define the absorption time

$$T(x) := \inf\left\{k \in \mathbb{N} : \, H_n^k(x) \in \{0,1\}\right\}.$$

Elementary Markov chain theory [12, Ch. 11] tells us that

Lemma 1. *For all x, $\mathbb{P}(T(x) < \infty) = 1$.*

Therefore, with probability 1, we have that

$$H_n^k(x) = H_n^{T(x)}(x) =: W(x), \quad \text{for all but finitely many } k.$$

Hence $f(H_n^k(x)) = f(W(x))$ for all bur finitely many k and so

$$\lim_{k\to\infty} \mathbb{E}f(H_n^k(x)) = \mathbb{E}f(W(x)).$$

But the random variable $W(x)$ takes two values: 0 and 1. Notice that $\mathbb{E}H_n^k(x) = x$ for all k. Hence $\mathbb{E}W(x) = x$. But $\mathbb{E}W(x) = 1 \times \mathbb{P}(W(x) = 1) + 0 \times \mathbb{P}(W(x) = 0)$. Thus $\mathbb{P}(W(x) = 1) = x$, and $\mathbb{P}(W(x) = 0) = 1 - x$. Hence

$$\lim_{k\to\infty} \mathbb{E}f(H_n^k(x)) = f(0)(1-x) + f(1)x. \tag{8}$$

This proves the announced limit of Theorem 2 but without the uniform convergence. However, since all polynomials of the sequence are of degree at most n, and n is a fixed number, convergence for each x implies convergence of the coefficients of the polynomials. □

To prove a rate of convergence in Theorem 2 we need to show what was announced in Remark 3(iii). Recall that $G_n(x) = S_n(x)/n$.

Lemma 2 (Convexity preservation)**.** *If f is convex then so is $B_n f$.*

Proof. Let $f : [0,1] \to \mathbb{R}$ and $\varepsilon > 0$. We shall prove that

$$\frac{d}{dx} B_n f(x) = n\mathbb{E}\left[f\left(\frac{S_{n-1}(x)+1}{n}\right) - f\left(\frac{S_{n-1}(x)}{n}\right)\right] \tag{9}$$

This can be done by direct computation using (2). Alternatively, we can give a probabilistic argument. Consider $B_n f(x+\varepsilon) - B_n f(x) = \mathbb{E}[f(G_n(x+\varepsilon)) - f(G_n(x))]$ and compute first-order terms in ε. By (4), $f(G_n(x+\varepsilon)) - f(G_n(x))$ is nonzero if and only if at least one of the U_i's falls in the interval $(x, x+\varepsilon]$. The probability that 2 or more variables fall in this interval is $o(\varepsilon)$, as $\varepsilon \to 0$. Hence, if F_ε is the event that *exactly* one of the variables falls in this interval, then

$$B_n f(x+\varepsilon) - B_n f(x) = \sum_{k=0}^{n-1} [f((k+1)/n) - f(k/n)]\, \mathbb{P}(S_n(x) = k, F_\varepsilon) + o(\varepsilon). \tag{10}$$

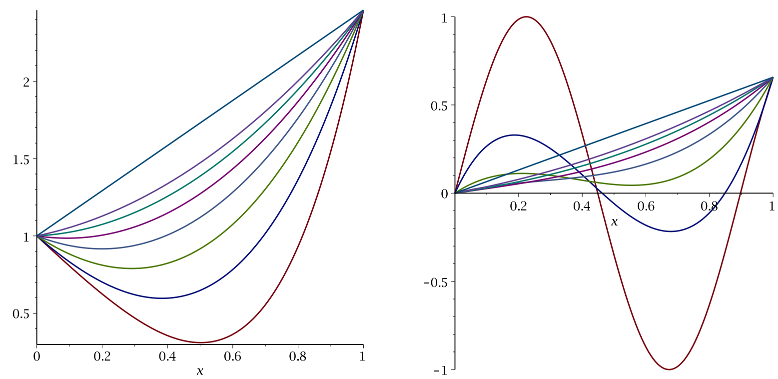
If we let $F_{j,\varepsilon}$ be the event that only U_j is in $(x, x+\varepsilon)$, then $\mathbb{P}(S_n(x) = k, F_{j,\varepsilon})$ is independent of j, so $\mathbb{P}(S_n(x) = k, F_\varepsilon) = n\mathbb{P}(S_n(x) = k, F_{n,\varepsilon}) = n\varepsilon\, \mathbb{P}(S_{n-1}(x) = S_{n-1}(x+\varepsilon) = k) = n\varepsilon\, \mathbb{P}(S_{n-1}(x) = k)\, (1 - \varepsilon/(1-x))^{n-1-k} = (n\varepsilon + o(\varepsilon))\, \mathbb{P}(S_{n-1}(x) = k)$. So (10) becomes

$$B_n f(x+\varepsilon) - B_n f(x) = n\varepsilon\mathbb{E}\left[f\left(\frac{S_{n-1}(x)+1}{n}\right) - f\left(\frac{S_{n-1}(x)}{n}\right)\right] + o(\varepsilon),$$

and, upon dividing by ε and letting $\varepsilon \to 0$, we obtain (9). Applying the same formula (9) once more (no further work is needed) we obtain

$$\frac{d^2}{dx^2} B_n f(x) = n(n-1)\mathbb{E}\left[f\left(\frac{S_{n-2}(x)+2}{n}\right) - 2f\left(\frac{S_{n-2}(x)+1}{n}\right) + f\left(\frac{S_{n-2}(x)}{n}\right)\right].$$

Bring in now the assumption that f is convex, whence $f(y+(2/n)) - 2f(y+(1/n)) + f(y) \geq 0$, for all $0 \leq y \leq 1-(2/n)$, and deduce that $(B_n f)''(x) \geq 0$ for all $0 \leq x \leq 1$. So $B_n f$ is a convex function. □



Convergence of the iterates of $B_n^k f$ as $k \to \infty$
for a convex f (left) and a nonconvex one (right).

We can now exhibit a rate of convergence.

Proposition 1. *For all $0 \leq x \leq 1$, $k, n \in \mathbb{N}$,*

$$|B_n^k f(x) - B_1 f(x)| \leq 2\|f\|\, \beta(k,x), \tag{11}$$

where

$$\beta(k,x) := \mathbb{P}(H_n^k(x) \notin \{0,1\}). \tag{12}$$

Moreover,

$$\beta(k,x) \leq \beta(k,1/2) \tag{13}$$

and

$$\beta(k,x) \leq n\left(1-\frac{1}{n}\right)^{k-1} x\,(1-x). \tag{14}$$

Proof. We have, for all positive integers ℓ and k,

$$\begin{aligned} B_n^k f(x) - B_n^{k+\ell} f(x) &= \mathbb{E}\left[f(H_n^{k+\ell}(x)) - f(H_n^k(x))\right] \\ &= \mathbb{E}\left[f(G_n^{k+\ell}\circ\cdots\circ G_n^{k+1}(H_n^k(x))) - f(H_n^k(x))\right] \\ &= \sum_{y\in\mathbb{I}_n\setminus\{0,1\}} \mathbb{P}(H_n^k(x)=y)\,\mathbb{E}\left[f(G_n^{k+\ell}\circ\cdots\circ G_n^{k+1}(y)) - f(y)\right], \end{aligned}$$

whence

$$\left|B_n^k f(x) - B_n^{k+\ell} f(x)\right| \le 2\|f\|\,\beta(k,x).$$

Letting $\ell \to \infty$ and using Theorem 2 yields (11). To prove (13), we notice that $\beta(k,x) = \mathbb{E}\varphi(H_n^k(x))$, where $\varphi(x) = 0$ if $x \in \{0,1\}$ and 1 otherwise, and, by (6), $\beta(k,x) = B_n^k\varphi(x)$. Note that $\beta(1,x) = B_n\varphi(x) = 1 - x^n - (1-x)^n$ is a concave function. By (6), $\beta(k,x) = B_n^k\varphi(x) = B_n^{k-1}B_n\varphi(x)$ which is also concave by Lemma 2. Hence

$$\frac{1}{2}\beta(k,x) + \frac{1}{2}\beta(k,1-x) \le \beta(k,1/2).$$

Since, by symmetry, $\beta(k,x) = \beta(k,1-x)$, inequality (13) follows. For the final inequality (14), notice that

$$\beta(k,x) = 1 - \mathbb{E}\left[(H_n^{k-1}(x))^n + (1 - H_n^{k-1}(x))^n\right] \le n\,\mathbb{E}\left[H_n^{k-1}(x)\left(1 - H_n^{k-1}(x)\right)\right],$$

where we used the inequality $1 - t^n - (1-t)^n \le nt(1-t)$, for all $0 \le t \le 1$. Therefore,

$$\beta(k,x) \le n\,\mathbb{E}\left[H_n^{k-1}(x)\left(1 - H_n^{k-1}(x)\right)\right] =: n\,\gamma(k-1,x).$$

Using

$$\mathbb{E}\left[G_n(x)\,(1 - G_n(x))\right] = \left(1 - \frac{1}{n}\right) x\,(1-x),$$

we obtain the recursion

$$\gamma(k-1,x) = \left(1 - \frac{1}{n}\right)\gamma(k-2,x).$$

Taking into account that $\gamma(0,x) = x(1-x)$ we find that $\gamma(k-1,x) = \left(1 - \frac{1}{n}\right)^{k-1} x\,(1-x)$. □

***Remark* 5.** Combining (11) and (14) we get

$$|B_n^k f(x) - B_1 f(x)| \le 2\|f\|\,n\left(1 - \frac{1}{n}\right)^{k-1} x(1-x).$$

This should be compared with [1, Eq. (4)] that says that $|B_n^k f(x) - B_1 f(x)| \le M(f,n)\ \left(1 - \frac{1}{n}\right)^{k-1} x(1-x)$, for some constant $M(f,n)$ which has not been computed in [1], whereas we have an explicit constant $2\|f\|n$. Now, the factor n is probably wasteful and this comes from the fact that the inequality $1 - t^n - (1-t)^n \le nt(1-t)$ is not good when n is large. We only used it because of the simplicity of the right-hand side that enabled us to compute $\gamma(k,x)$ very easily. We have a better inequality, namely (11), but to make it explicit one needs to compute $\mathbb{P}(H_n^k(x) = 0)$.

4 Interlude: population genetics and the Wright-Fisher model

We now take a closer look at the Markov chain described by the sequence $(H_n^k(x), k \in \mathbb{N})$ for fixed n. We repeat formula (7):

$$\mathbb{P}\left(nH_n^{k+1}(x) = j \mid nH_n^k(x) = i\right) = \binom{n}{j}\left(\frac{i}{n}\right)^j\left(1 - \frac{i}{n}\right)^{n-j}. \tag{7'}$$

We recognize that it describes the simplest stochastic model for reproduction in population genetics that goes as follows. There is a population of N individuals each of which carries 2 genes. Genes come in 2 variants, I and II, say, Thus, an individual may have 2 genes of type I both, or of type II both, or one of each. Hence there are $n = 2N$ genes in total. We observe the population at successive generations and assume that generations are non-overlapping. Suppose that the k-th generation consists of i genes of type I and $n - i$ of type II. In Figure 1 below, type I genes are yellow, and type II are red. To specify generation $k + 1$, we let each gene of generation $k + 1$ select

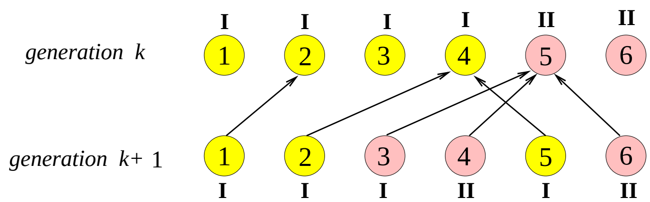


Figure 1: *An equivalent way to form generation $k + 1$ from generation k is by letting φ be a random element of the set $[n]^{[n]}$, the set of all mappings from $[n] := \{1, \dots, n\}$ into itself and by drawing an arrow with starting vertex i end ending vertex $\varphi(i)$. If $\varphi(i)$ is of type* I *(or* II*) then i becomes* I *(or* II*) too.*

a single "parent" at random from the genes of the previous generation. The gene adopts the type of its parent. The parent selection is independent across genes. The probability that a specific gene selects a parent of type I is i/n. Since we have n independent trials, the probability that generation $k + 1$ will contain j genes of type I is given by the right-hand side of formula (7′). If we start the process at generation 0 with genes of type I being chosen, independently, with probability x each, then the number of alleles at the k-th generation has the distribution of $nH_n^k(x)$.

This stochastic model we just described is known as the Wright-Fisher model, and is fundamental in mathematical biology for populations of fixed size. The model is very far from reality, but has nevertheless been extensively studied and used.

Early on, Wright and Fisher observed that performing exact computations with this model is hard. They devised a continuous approximation observing that the probability $\mathbb{P}(H_n^k(x) \le y)$ as a function of x, y and k can, when n is large, be approximated by a smooth function of x and y. (See, e.g., Kimura [18] and the recent paper by Tran, Hofrichter, and Jost [23].) Rather than approximating this probability, we follow modern methods of stochastic analysis in order to approximate the discrete stochastic process $(H_n^k(x), k \in \mathbb{N})$ by a continuous-time continuous-space stochastic process that is nowadays known as Wright-Fisher diffusion.

5 The Wright-Fisher diffusion

Our eventual goal is to understand what happens when we consider the limit of $B_n^k f$, when both k and n tend to infinity. From the Bernstein and the Kelisky-Rivlin theorems we should expect that the order at which limits over k and n are taken matters. It turns out that the only way to obtain a limit is when the ratio k/n tends to a constant, say, t. This is intimately connected to the Wright-Fisher diffusion that we introduce next. We assume that the reader has some knowledge of stochastic calculus, including the Itô formula and stochastic differential equations driven by a Brownian motion at the basic level of Øksendal [19] or at the more advanced level of Bass [2].

We first explain why we expect that the Markov chain $H_n^k(x)$, $k \in \mathbb{Z}_+$ has a limit, in a certain sense, as $n \to \infty$. Our explanation here will be informal. We shall give rigorous proofs of only what we need in the following sections.

The first thing we do is to compute the expected variance of the increment of the chain, and examine whether it converges to zero and at which rate: see (42), Theorem 5, Section A3 in the Appendix. The rate of convergence gives us the right time scale. Our case at hand is particularly simple because we have an exact formula:

$$\mathbb{E}\left[(H_n^{k+1}(x) - H_n^k(x))^2 \mid H_n^k(x) = y\right] = \mathbb{E}\left[(G_n(y) - y)^2\right] = \frac{1}{n} y(1-y). \tag{15}$$

This suggests that the right time scale at which we should run the Markov chain is such that the time steps are of size $1/n$. In other words, consider the points

$$(0, x),\ (1/n, H_n^1(x)),\ (2/n, H_n^2(x)),\ \ldots \tag{16}$$

and draw the random curve

$$t \mapsto H_n^{[nt]}(x)$$

(where $[nt]$ is the integer part of nt) as in Figure 2. This is at the right time scale.

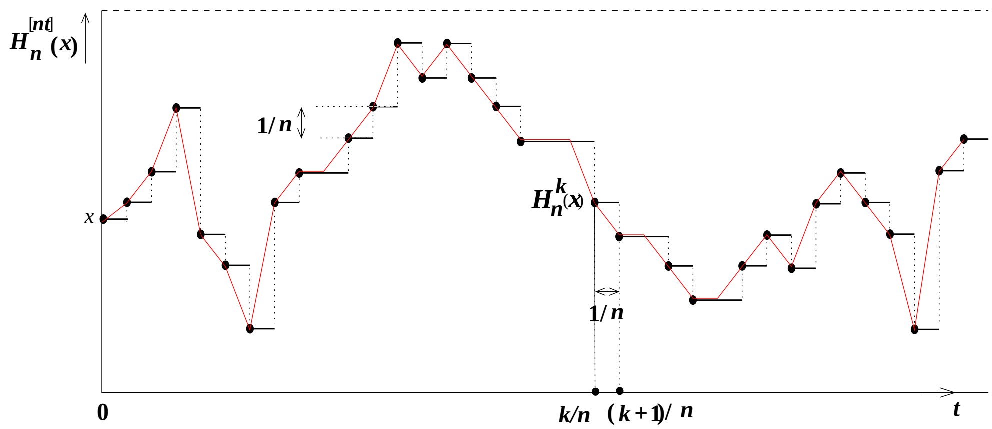


Figure 2: *A continuous-time curve from the discrete-time Markov chain*

The second thing we do (see (43), Theorem 5, Section A3) is to compute the expected change of the Markov chain. In our case, this is elementary:

$$\mathbb{E}\left[H_n^{k+1}(x) - H_n^k(x) \mid H_n^k(x) = y\right] = \mathbb{E}\left[G_n(y) - y\right] = 0. \tag{17}$$

The functions $\sigma(y)^2 = y(1-y)$ and $b(y) = 0$ obtained in (15) and (17) suggest that the limit of the random curve $(H_n^{[nt]}(x),\, t \geq 0)$ should be a diffusion process $X_t(x)$, $t \geq 0$, satisfying the stochastic differential equation

$$\begin{aligned} dX_t(x) &= \sigma(X_t(x))dW_t + b(X_t(x))dt \\ &= \sqrt{X_t(x)\,(1 - X_t(x))}\,dW_t, \end{aligned} \tag{18}$$

with initial condition $X_0(x) = x$, where W_t, $t \geq 0$, is a standard Brownian motion.

It is actually possible to prove that $(H_n^{[nt]}, t \geq 0)$ converges *weakly* to $(X_t(x), t \geq 0)$, but this requires an additional estimate on the size of the increments of the Markov chain that is, in our case, provided by the following inequality: for any $\varepsilon > 0$,

$$\mathbb{P}(|H_n^{k+1}(x) - H_n^k(x)| > \varepsilon \mid H_n^k(x) = y) = \mathbb{P}(|G_n(y) - y| > \varepsilon) \leq 2e^{-\frac{1}{2}\varepsilon^2 n}. \tag{19}$$

To see this, apply Hoeffding's inequality (see (40) Section A2 in the Appendix).

We thus have that (15), (17), (19) are the conditions (42), (43) and (44) of Theorem 5, Section A3. In addition, it can be shown that the stochastic differential equation (18) admits a unique strong solution for any initial condition x. This is, e.g., a consequence of the Yamada-Watanabe theorem [2, Theorem 24.4]. Hence, by Theorem 5, the sequence of continuous random curves $(H_n^{[nt]}(x), t \geq 0)$ converges weakly to the continuous random function $(X_t, t \geq 0)$.

One particular conclusion of weak convergence is that $\mathbb{E}f(H_n^{[nt]}(x)) \to \mathbb{E}f(X_t(x))$ for any $f \in C[0, 1]$ or, equivalently, that

Theorem 3 (joint limits theorem)**.** *For any $f \in C[0, 1]$ and any $t \geq 0$,*

$$\lim_{n\to\infty} B_n^{[nt]} f(x) = \mathbb{E}f(X_t(x)), \quad \textit{uniformly in } x. \tag{20}$$

Since understanding the theorem of Stroock and Varadhan requires advanced machinery, we shall prove Theorem 3 directly. The proof is deferred until Section 8. The nice thing with this theorem is that we have a way to compute the limit by means of stochastic calculus, the tools of which we shall assume as known.

Let f be a twice-continuously differentiable function. Then, Itô's formula ([2, Ch. 11], [19, Ch. 4]) says that

$$f(X_t) = f(X_0) + \int_0^t f'(X_s)dX_s + \frac{1}{2}\int_0^t f''(X_s)X_s(1 - X_s)ds. \tag{21}$$

If we let

$$\mathcal{L}f := \frac{1}{2}x(1 - x)\frac{d^2 f}{dx^2}, \tag{22}$$

take expectations in (21), and differentiate with respect to t, we obtain

$$\frac{\partial}{\partial t}\mathbb{E}f(X_t(x)) = \mathbb{E}(\mathcal{L}f)(X_t(x)), \tag{23}$$

the so-called forward equation of the diffusion. Now let, for all $s \geq 0$,

$$\mathsf{P}_s g(x) := \mathbb{E}g(X_s(x)),$$

noticing that $\mathsf{P}_s g$ is defined for all bounded and measurable g and that $\mathbb{P}_0 g(x) = g(x)$. If g is such that $\mathsf{P}_s g \in C^2$ then we can set $f = \mathsf{P}_s g$ in (23). Now, $\mathbb{E}(\mathsf{P}_s g)(X_t(x)) = \mathbb{E}g(X_{t+s}(x))$, because of the Markov property of X_t, $t \geq 0$, and so (23) becomes

$$\frac{\partial}{\partial t}\mathbb{E}g(X_{t+s}(x)) = \mathbb{E}(\mathcal{L}\mathsf{P}_s g)(X_t(x)).$$

Letting $t \to 0$, we arrive at the backward equation

$$\frac{\partial}{\partial s}\mathbb{E}g(X_s(x)) = \mathcal{L}\mathsf{P}_s g(x), \tag{24}$$

which is valid if $\mathsf{P}_s g$ is twice continuously differentiable. The class of functions g such that both g and $\mathsf{P}_s g$ are in C^2 is nontrivial in our case. It contains, at least polynomials. This is what we show next.

6 Moments of the Wright-Fisher diffusion

It turns out that in order to prove Theorem 3 we need to compute $\mathbb{E}f(X_t(x))$ when f is a polynomial.

Proposition 2. *For a positive integer r, the following holds for the Wright-Fisher diffusion:*

$$\mathbb{E}X_t(x)^r = \sum_{i=1}^r b_{i,r}(t)\, x^i,$$

where

$$b_{i,r}(t) = \sum_{j=i}^r \frac{A_{i,r}}{B_{i,j,r}}\, e^{-\alpha_j t}, \tag{25}$$

$$\alpha_j = \frac{1}{2}j(j-1), \quad A_{i,r} = \prod_{k=i+1}^r \alpha_k, \quad B_{i,j,r} = \prod_{k=i,k\neq j}^r (\alpha_k - \alpha_j), \quad 1 \le i \le j \le r \tag{26}$$

(where, as usual, a product over an empty set equals 1).

Proof. Write $X_t = X_t(x)$ to save space. By Itô's formula (21) applied to $f(x) = x^r$,

$$X_t^r = x^r + r\int_0^t X_s^{r-1} dX_s + \frac{1}{2}r(r-1)\int_0^t X_s^{r-2} X_s(1-X_s)ds.$$

Since the first integral is (as a function of t) a martingale starting from 0 its expectation is 0. Thus, if we let

$$m_r(t,x) := \mathbb{E}X_t(x)^r,$$

we have

$$m_r(t,x) = x^r + \alpha_r \int_0^t (m_{r-1}(s,x) - m_r(s,x))ds.$$

Thus, $m_1(t,x) = x$, as expected, and

$$\frac{\partial}{\partial t}m_r(t,x) = \alpha_r(m_{r-1}(t,x) - m_r(t,x)), \quad r = 2, 3, \ldots$$

Defining the Laplace transform

$$\widehat{m}_r(s,x) = \int_0^\infty e^{-st} m_r(t,x)dt,$$

and using integration by parts to see that $\int_0^t e^{-st} \frac{\partial}{\partial t} m_r(t,x) dt = s\widehat{m}_r(s,x) - m_r(0,x) = s\,\widehat{m}_r(s,x) - x^r$ we have

$$s\,\widehat{m}_r(s,x) - x^r = \alpha_r(\widehat{m}_{r-1}(s,x) - \widehat{m}_r(s,x)).$$

Iterating this easy recursion yields

$$\widehat{m}_r(s,x) = \sum_{i=1}^{r} \frac{\alpha_r}{s+\alpha_r} \cdots \frac{\alpha_{i+1}}{s+\alpha_{i+1}} \frac{x^i}{s+\alpha_i} = \sum_{i=1}^{r} A_{i,r} \sum_{j=i}^{r} \frac{1/B_{i,j,r}}{s+\alpha_j} x^i,$$

where the second equality was obtained by partial fraction expansion (and the notation is as in (26)). Since the inverse Laplace transform of $1/(s+\alpha_j)$ is $e^{-\alpha_j t}$, the claim follows. □

***Remark* 6.** Formula (25) was proved by Kelisky and Rivlin [17, Eq. (3.13)] and Karlin and Ziegler [16, Eq. (1.13)] by entirely different methods. (the latter paper contains a typo in the formula). Eq. (3.13) of [17] reads:

$$b_{i,r}(t) = \frac{i}{r}\binom{r}{i}^2 \sum_{j=i}^{r} \frac{(-1)^{i+j}\binom{r-i}{j-i}^2}{\binom{2j-2}{j-i}\binom{j+r-1}{r-j}} e^{-\frac{1}{2}j(j-1)t}. \tag{27}$$

Comparing this with (25) and (26) we obtain

$$\frac{\prod_{k=i+1}^{r}\binom{k}{2}}{\prod_{k=i,k\neq j}^{r}\left[\binom{k}{2}-\binom{j}{2}\right]} = \frac{i}{r}\binom{r}{i}^2 \frac{(-1)^{i+j}\binom{r-i}{j-i}^2}{\binom{2j-2}{j-i}\binom{j+r-1}{r-j}},$$

valid for any integers i, j, r with $1 \le i \le j \le r$. This equality can be verified directly by simple algebra.

7 Convergence rate to Bernstein's theorem: Voronovskaya's theorem

An important result in the theory of approximation of continuous functions is Voronovskaya's theorem [24]. It is the simplest example of *saturation*, namely that, for certain operators, convergence cannot be too fast even for very smooth functions. See DeVore and Lorentz[7, Theorem 3.1]. Voronovskaya's theorem gives a rate of convergence to Bernstein's theorem. From a probabilistic point of view, the theorem is nothing else but the convergence of the generator of the discrete Markov chain to the generator of the Wright-Fisher diffusion. We shall not use anything from the theory of generators, but we shall give an independent probabilistic proof below for C^2 functions f, including a slightly improved form under the assumption that f'' is Lipschitz. In this case, its Lipschitz constant is

$$\mathrm{Lip}(f'') := \sup_{x\neq y} |f''(x) - f''(y)|/|x-y|.$$

Recall that $\mathcal{L}f$ is defined by (22).

Theorem 4 (Voronovskaya, 1932). *For any $f \in C^2[0,1]$,*

$$\lim_{n\to\infty} \max_{x\in[0,1]} |n(B_n f(x) - f(x)) - \mathcal{L}f(x)| = 0. \tag{28}$$

If moreover f'' is Lipschitz then, for any $n \in \mathbb{N}$,

$$\max_{x\in[0,1]} |n(B_n f(x) - f(x)) - \mathcal{L}f(x)| \le \frac{\mathrm{Lip}(f'')}{16 \cdot 3^{1/4}} n^{-1/2} \tag{29}$$

Proof. Using Taylor's theorem with the remainder in integral form,

$$f(G_n(x)) - f(x) = f'(x)(G_n(x) - x) + \int_x^{G_n(x)} (G_n(x) - t) f''(t) dt.$$

Since $\int_x^{G_n(x)} (G_n(x) - t) dt = \frac{1}{2}(G_n(x) - x)^2$, we have, from (15), $\mathbb{E} \int_x^{G_n(x)} (G_n(x) - t) dt = \frac{1}{2} \frac{x(1-x)}{n}$. Therefore,

$$n(\mathbb{E}f(G_n(x)) - f(x)) - \frac{1}{2} x(1-x) f''(x) = n\,\mathbb{E} \int_x^{G_n(x)} (G_n(t) - t)\,(f''(t) - f''(x)) dt =: n\,\mathbb{E}J_n(x).$$

We estimate $\mathbb{E}J_n(x)$ by splitting the expectation as

$$\mathbb{E}J_n(x) = \mathbb{E}[J_n(x); |G_n(x) - x| \le \delta] + \mathbb{E}[J_n(x); |G_n(x) - x| > \delta],$$

where $\delta > 0$ is chosen by the uniform continuity of f'': for $\varepsilon > 0$ let δ be such that $|f''(x) - f''(y)| \le \varepsilon$ whenever $|x - y| \le \delta$. Thus, $|J_n(x)| \le \varepsilon |\int_x^{G_n(x)} (G_n(x) - t) dt|$ and so

$$|\mathbb{E}[J_n(x); |G_n(x) - x| \le \delta]| \le \varepsilon \mathbb{E}\left[\left| \int_x^{G_n(x)} (G_n(x) - t) dt \right| ; \, |G_n(x) - x| \le \delta \right] \le \varepsilon \frac{x(1-x)}{n} \le \frac{\varepsilon}{4n}.$$

On the other hand, with $\|f''\| = \max_{0\le x\le 1} |f''(x)|$, we have $|J_n(x)| \le 2\|f''\| |\int_x^{G_n(x)} (G_n(t) - t) dt| \le 2\|f''\|$, so

$$|\mathbb{E}[J_n(x); |G_n(x) - x| > \delta]| \le 2\|f''\| \mathbb{P}(|G_n(x) - x| > \delta) \le 4\|f''\| e^{-\delta n^2/2},$$

by (19). Hence

$$|n(\mathbb{E}f(G_n(x)) - f(x)) - \frac{1}{2} x(1-x) f''(x)| \le \frac{\varepsilon}{4} + 4\|f''\| n e^{-\delta n^2/2}.$$

Letting $n \to \infty$, and since $\varepsilon > 0$ is arbitrary, we obtain the first claim (28).
§Assume next that f'' is Lipschitz with Lipschitz constant $\mathrm{Lip}(f'')$. Then

$$\begin{aligned}
|n(\mathbb{E}f(G_n(x)) - f(x)) - \frac{1}{2} x(1-x) f''(x)| &\le n\mathbb{E} \int_{x\wedge G_n(x)}^{x\vee G_n(x)} |G_n(x) - t|\, |f''(t) - f''(x)| dt \\
&\le n\,\mathrm{Lip}(f'')\,\mathbb{E} \int_{x\wedge G_n(x)}^{x\vee G_n(x)} |G_n(x) - t|\, |t - x| dt \\
&= \frac{1}{6} n\,\mathrm{Lip}(f'')\,\mathbb{E}|G_n(x) - x|^3 \\
&\le \frac{1}{6} n\,\mathrm{Lip}(f'')\,(\mathbb{E}(G_n(x) - x)^4)^{3/4} \qquad (30)
\end{aligned}$$

We have $\mathbb{E}(G_n(x)-x)^4 = n^{-4}\mathbb{E}(S_n-nx)^4$, where S_n is a binomial random variable–see (2). We can easily find (or look in a textbook) that

$$\mathbb{E}(S_n-nx)^4 = nx(1-x)(1-6x+6x^2+3nx-3nx^2) =: \mu(n,x).$$

Since $\mu(n,x)=\mu(n,1-x)$ and, for $n\geq 2$, the function $\mu(n,x)$ is concave in x, it follows that $\mu(n,x)\leq\mu(n,1/2)=n(2n-1)/16\leq 3n^2/16$. On the other hand, for $n=1$, $\mu(1,x)\leq 3/16$ for all x. Thus, the last term of (30) is upper bounded by

$$\frac{1}{6}n\,\mathrm{Lip}(f'')\left(\frac{\mu(n,x)}{n^4}\right)^{3/4}\leq\frac{\mathrm{Lip}(f'')}{6}n^{-1/2}(3/16)^{3/4}=\frac{\mathrm{Lip}(f'')}{16\cdot 3^{1/4}}n^{-1/2}.$$

□

Remark 7. Probabilistically, the operator B_n-I, where I is the identity operator, maps a function $f\in C[0,1]$ to the function $y\mapsto\mathbb{E}f((G_n(y))-f(y)=\mathbb{E}\big[f(H_n^{k+1}(x))-f(H_n^k(x))|H_n^k(x)=y\big]$ which is the expected change of the function f (under the action of the chain) per unit time, conditional on the current state being equal to y. Since the natural time scale is counted in time units that are multiples of $1/n$, we can interpret $n(B_n-I)f$ as the expected change of f per unit of real time. Thus, its "limit" $\mathcal{L}$ should play a similar role for the diffusion. And, indeed, it does, but we shall not use this here. For further information on diffusions and their generators see, e.g., Karlin and Taylor [15].

8 Joint limits

The goal of this section is a probabilistic proof of Theorem 3. First notice that it suffices to prove (20) for polynomial functions. Indeed, if f is continuous on $[0,1]$ and $\varepsilon>0$, there is a polynomial $h=B_kf$ such that $\|h-f\|\leq\varepsilon$ (by Bernstein's Theorem 2). But then

$$\|B_n^{[nt]}f-f\|\leq\|B_n^{[nt]}f-B_n^{[nt]}h\|+\|\mathsf{P}_tf-\mathsf{P}_th\|+\|B_n^{[nt]}h-\mathsf{P}_th\|\leq 2\varepsilon+\|B_n^{[nt]}h-\mathsf{P}_th\|,$$

where we used the fact that both $B_n^{[nt]}$ and P_t are defined via expectations, and so $\|B_n^{[nt]}f-B_n^{[nt]}h\|=\|B_n^{[nt]}(f-h)\|\leq\|f-h\|$, and, similarly, $\|\mathsf{P}_tf-\mathsf{P}_th\|=\|\mathsf{P}_t(f-h)\|\leq\|f-h\|$. Therefore, if $\|B_n^{[nt]}h-\mathsf{P}_th\|\to 0$ for polynomial h then $\|B_n^{[nt]}f-f\|\to 0$ for any continuous f. Equivalently, we need to prove that

$$\lim_{n\to\infty}\mathbb{E}h(H_n^{[nt]}(x))=\mathbb{E}h(X_t(x)),\ \text{uniformly in } x. \tag{31}$$

Notice that $H_n^{[nt]}(x)$, $t\geq 0$, is not a Markov process. However if $\Phi(t)$, $t\geq 0$, is a standard Poisson process with rate 1, starting from $\Phi(0)=0$, and independent of everything else, then

$$X_t^{(n)}(x):=H_n^{\Phi(nt)}(x),\quad t\geq 0,$$

is a Markov process for each n. Moreover, for all $f\in C^2[0,1]$,

$$\left|\mathbb{E}f(H_n^{\Phi(nt)}(x))-\mathbb{E}f(H_n^{[nt]}(x))\right|\to 0,\quad\text{uniformly in } x. \tag{32}$$

To see this, let $G_n^1, G_n^2, \dots$ be i.i.d. copies of G_n, as in (5), and write the triangle inequality

$$\left|\mathbb{E}f(G_n^2 \circ G_n^1(x)) - f(x)\right| \le \left|\mathbb{E}[f(G_n^2 \circ G_n^1(x)) - f(G_n^1(x))]\right| + \left|\mathbb{E}[f(G_n^1(x)) - f(x)]\right|. \tag{33}$$

Since $\left|\mathbb{E}[f(G_n^2(y)) - f(y)]\right| \le \|B_n f - f\|$ for all y, and since G^1 is independent of G^2, we have that the first term of the right side of (33) is $\le \|B_n f - f\|$, and so

$$\left|\mathbb{E}f(G_n^2 \circ G_n^1(x)) - f(x)\right| \le 2\|B_n f - f\|, \quad \text{for all } 0 \le x \le 1.$$

By the same argument, for $k < \ell$,

$$\left|\mathbb{E}f(G_n^\ell \circ \cdots \circ G_n^{k+1}(y)) - f(y)\right| \le (\ell - k)\|B_n f - f\|, \quad \text{for all } 0 \le y \le 1.$$

Since $H_n^k(x)$ is independent of $G_n^\ell \circ \cdots \circ G_n^{k+1}$, we can replace the y in the last display by $H_n^k(x)$ and obtain

$$\left|\mathbb{E}[f(H_n^\ell(x)) - f(H_n^k(x))]\right| \le (\ell - k)\|B_n f - f\|, \quad \text{for all } 0 \le x \le 1.$$

Using the fact that the Poisson process Φ is independent of everything else, we obtain

$$\left|\mathbb{E}\{f(H_n^{\Phi(nt)}(x)) - f(H_n^{[nt]}(x))\}\right| \le \mathbb{E}\{|\Phi(nt) - [nt]|\}\, \|B_n f - f\|, \quad \text{for all } 0 \le x \le 1. \tag{34}$$

But $\mathbb{E}\{|\Phi(nt) - [nt]|\} \le \mathbb{E}\{|\Phi(nt) - nt|\} + 1 \le \sqrt{\mathbb{E}\{|\Phi(nt) - nt|\}^2} + 1 = \sqrt{nt} + 1$, while, from Voronovskaya's theorem, $n\|B_n f - f\| \to \|\mathcal{L}f\|$. Therefore the right-hand side of (34) converges to 0 as $n \to \infty$, and this proves (32).

Therefore (31) will follow from

$$\lim_{n\to\infty} \mathbb{E}h(X_t^{(n)}(x)) = \mathbb{E}h(X_t(x)), \quad \text{uniformly in } x, \tag{35}$$

for all polynomial h.

For each $s \ge 0$, define a random curve $Y_t^{(s)}$ that follows $X^{(n)}$ up to time s and then switches to X. More precisely, define

$$Y_t^{(s)} := \begin{cases} X_t^{(n)}(x), & 0 \le t \le s \\ X_{t-s}(X_s^{(n)}(x)), & t \ge s \end{cases}. \tag{36}$$

See Figure 3. It is here assumed that the Brownian motion W driving the defining equation (18) for the Wright-Fisher diffusion is independent of all random variables used for the construction of $X^{(n)}$. Since, for any given initial state, the solution to (18) is unique, we may replace the initial state by a random variable independent of the Brownian motion, and this is what we did in the last formula. Thus, if we prove that $\mathbb{E}h(Y_t^{(s)})$ is differentiable with respect to s, we shall have, for $t \ge s$,

$$\mathbb{E}h(X_t(x)) - \mathbb{E}h(X_t^{(n)}(x)) = \mathbb{E}h(Y_t^{(0)}) - \mathbb{E}h(Y_t^{(t)}) = \int_0^t \frac{\partial}{\partial s}\mathbb{E}h(X_s(X_{t-s}^{(n)}(x)))\, ds. \tag{37}$$

To show that the last derivative exists as well as estimate it, we estimate $\frac{\partial}{\partial s}\mathbb{E}h(X_s(X_t^{(n)}(x)))$ and $\frac{\partial}{\partial t}\mathbb{E}h(X_s(X_t^{(n)}(x)))$.

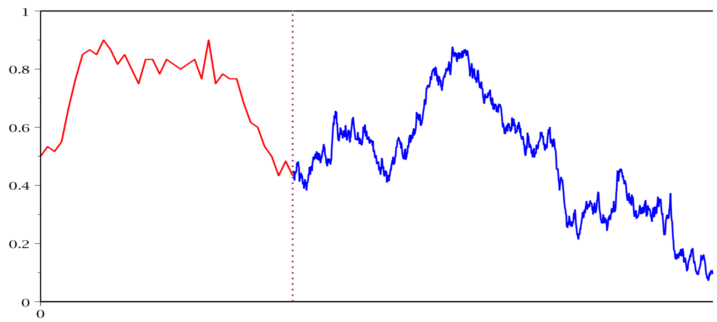


Figure 3: *A trajectory of (joined-by-straight-lines version of) the Markov chain* $X^{(n)}$ *followed by the diffusion* X.

Since h is a polynomial, it follows that $\mathsf{P}_s h(x) = \mathbb{E}h(X_s(x))$ is also a polynomial function of x (see Proposition 2) and so the forward equation (24) holds:

$$\frac{\partial}{\partial s}\mathbb{E}h(X_s(x)) = \mathcal{L}\mathsf{P}_s h(x).$$

Therefore, for any random variable Y with values in $[0,1]$ and independent of X_s, we have

$$\frac{\partial}{\partial s}\mathbb{E}\{h(X_s(Y))|Y\} = \mathcal{L}\mathsf{P}_s h(Y).$$

Taking expectations of both sides and interchanging differentiation and expectation (by the DCT) we have

$$\frac{\partial}{\partial s}\mathbb{E}h(X_s(Y)) = \mathbb{E}\mathcal{L}\mathsf{P}_s h(Y).$$

Setting $Y = X_t^{(n)}(x)$, we have

$$\frac{\partial}{\partial s}\mathbb{E}h(X_s(X_t^{(n)}(x))) = \mathbb{E}\mathcal{L}(\mathsf{P}_s h)(X_t^{(n)}(x)). \tag{38}$$

Assume next that $f : [0,1] \to \mathbb{R}$ is any function. Using the identity

$$\mathbb{E}[f(H_n^{k+1}(x)) - f(H_n^k(x))|H_n^k(x) = y] = \mathbb{E}f(G_n(y)) - f(y) = B_n f(y) - f(y),$$

valid for any function f and any x, together with the fact that $\mathbb{P}(\Phi(t+h) - \Phi(t) = 1) = h + o(h)$, as $h \downarrow 0$, we arrive at

$$\frac{\partial}{\partial t}\mathbb{E}f(X_t^{(n)}(x)) = n\mathbb{E}[B_n f(X_t^{(n)}(x)) - f(X_t^{(n)}(x))].$$

Setting now $f = \mathsf{P}_s h$ and observing that $\mathbb{E}(\mathsf{P}_s h)(X_t^{(n)}(x)) = \mathbb{E}h(X_s(X_t^{(n)}(x)))$, we arrive at

$$\frac{\partial}{\partial t}\mathbb{E}f(X_s(X_t^{(n)}(x))) = n\mathbb{E}[B_n \mathsf{P}_s h(X_t^{(n)}(x)) - \mathsf{P}_s h(X_t^{(n)}(x))] \tag{39}$$

Combining (38) and (39) we have a formula for the derivative appearing in the last term of (37):

$$\frac{\partial}{\partial s}\mathbb{E}h(X_s(X_{t-s}^{(n)}(x))) = \mathbb{E}F(X_{t-s}^{(n)}(x)),$$

where

$$F(y) := \mathcal{L}(\mathsf{P}_s h)(y) - n[B_n \mathsf{P}_s h(y) - (\mathsf{P}_s h)(y)]$$

Assume now that $(\mathsf{P}_s h)''$ is Lipschitz with Lipschitz constant L_s. Then, by (29),

$$|F(y)| \le c_2 L_s n^{-1/2}, \quad 0 \le y \le 1,$$

where $c_2 = 1/16 \cdot 3^{1/4}$ and $L_s = \mathrm{Lip}((\mathsf{P}_s h)'')$, and so

$$\left|\mathbb{E}h(X_t^{(n)}(x)) - \mathbb{E}h(X_t(x))\right| \le \int_0^t \left|\mathbb{E}F(X_{t-s}^{(n)}(x)\right| ds \le c_2 n^{-1/2} \int_0^t L_s ds.$$

By the formula for $\mathsf{P}_s h$ when h is a polynomial (Proposition 2), it follows that the $\int_0^t L_s ds$ is a finite constant. Hence (35) has been proved. □

Corollary 1. *With $f_r(x) = x^r$,*

$$\lim_{n\to\infty} B_n^{[nt]} f_r(x) = \sum_{i=1}^r b_{i,r}(t)\, x^i,$$

where the $b_{i,r}(t)$ are given by (27).

This is Theorem 2 in Kelisky and Rivlin [17].

Corollary 2. *With $f_\theta(x) = e^{-\theta x}$, we have*

$$\lim_{n\to\infty} B_n^{[nt]} f_\theta(x) = \mathbb{E}e^{-\theta X_t(x)} =: H(t, x, \theta),$$

where $H(t, x, \theta)$ satisfies $H(0, x, \theta) = e^{-\theta x}$ and the PDE

$$\frac{\partial H}{\partial t} = -\frac{\theta^2}{2}\frac{\partial H}{\partial \theta} - \frac{\theta^2}{2}\frac{\partial^2 H}{\partial \theta^2}.$$

The solution to this PDE can be expressed in terms of modified Bessel functions. We shall not pursue this further here.

9 Further comments

We provided a fully stochastic explanation of the phenomenon of convergence of k iterates of Bernstein operators of degree n when n and k tend to infinity in different ways. This problem has received attention in the theory of approximations of continuous functions. We showed that the problem can be interpreted naturally via stochastic processes. In fact, these processes, the Wright-Fisher model and Wright-Fisher diffusion are very basic in probability theory and are well-understood.

There are a number of interesting directions that open up. The most crucial thing is that $B_n f(x)$ is the expectation of a random variable. We can construct different operators by using different random variables. See, e.g., Karlin and Ziegler [16, Eq. (1.5)] for an operator related to a Poisson random variable. Whereas Karlin and Ziegler study iterates of these operators, their approach is more analytical than probabilistic. By using approximations by stochastic differential equations, and taking advantage of the tools stochastic calculus, it is possible to derive convergence rates and other interesting results, including explicit formulas, such as the formula for $\lim_{n\to\infty} B_n^{[nt]} f_r(x)$ (Corollary 1 and formula (27)), obtained here by a simple application of the Itô formula.

In Section 4 we explained the most standard Wright-Fisher model where mutations are not allowed. If we assume that the probability that a gene of one type changing to another type also depends on the number of genes of each type, in a possibly nonlinear fashion, we obtain a more general model. Mathematically, this is captured by letting $h : [0,1] \to [0,1]$ be an appropriate function, and by considering the Markov chain obtained by iterating independent copies of the function $x \mapsto G_n(h(x))$, that is the Markov chain $G_n^k \circ h \circ \cdots \circ G_n^1 \circ h(x)$, $k \in \mathbb{N}$. For the case $h(x) = ax + b$ with a, b chosen so that $0 \le h(x) \le 1$, see Ethier and Norman [9].

Appendix

A1 Composing random maps

By a random map $G : T \to S$ we mean a random function from some probability space Ω into a subspace of S^T of functions from T to S. In rigorous probability theory, this means that all sets involved are equipped with σ-algebras in a way that $G(\omega) \in T^S$ is a measurable function of ω. As a concrete example, in this paper, we considered $S = T = [0,1]$ and $\Omega = [0,1]^n$. We equipped all sets with the Lebesgue measure. For $\omega = (\omega_1, \ldots, \omega_n) \in \Omega$, the map $U_i : \omega \mapsto \omega_i$ is a random variable with the uniform distribution. Moreover, $U_1, \ldots, U_n$ are independent. We defined $G(\omega)$ the function

$$G(\omega)(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\omega_i \le x}.$$

By taking a different Ω, we are able to construct two (or more) random maps G_1, G_2 that are independent. As usual in probability, we suppress the symbol ω from the definition of G. When we talk about the composition $G_2 \circ G_1$ of two random functions as above, we are talking about the composition with respect to x. That is, $G_2 \circ G_1(\omega)(x) := G_2(\omega)(G_1(\omega)(x))$. If f is a deterministic function we define the operator $f \mapsto T_i f$ by $T_i f(x) := \mathbb{E} f(G_i(x))$, $i = 1, 2$. We can then easily see that

$$\mathbb{E} f(G_2 \circ G_1(x)) = T_1 \circ T_2 f(x),$$

the point being that the order of composition outside the expectation is the reverse of the one inside. This is why, for instance, $\mathbb{E} f(G_n^k \circ \cdots \circ G_n^1(x)) = B_n \circ \cdots \circ B_n f(x)$ in eq. (6). Another example of a random map is the $x \mapsto X_s^{(n)}(x)$, where $X^{(n)}$ is the Markov chain constructed in Section 8. For fixed s and n, the initial state x is mapped into the state $X_s^{(n)}(x)$ at time s. And yet another random map is $x \mapsto X_t(x)$, where X_t, $t \ge 0$, is the Wright-Fisher diffusion. We composed these random maps in (36), after assuming that they are independent.

A2 Hoeffding's inequality

Let $X_1, \ldots, X_n$ be independent random variables with zero mean with values between $-c$ and c for some $c > 0$. Then, for $-c \le x \le c$,

$$\mathbb{P}(|S_n| > nx) \le 2e^{-nx^2/2c}. \tag{40}$$

This inequality, due to Hoeffding [13], is very well-known and can be found in many probability theory books. We just prove it below for completeness.

Proof. Since $e^{\theta x}$ is a convex function of x for all $\theta > 0$ and so if we write X_i as a convex combination of $-c$ and c,

$$X_i = \frac{c - X_i}{2c}(-c) + \frac{X_i + c}{2c}c,$$

we obtain

$$e^{\theta X_i} \leq \frac{c - X_i}{2c}e^{-\theta c} + \frac{X_i + c}{2c}e^{\theta c},$$

Hence

$$\mathbb{E}e^{\theta X_i} = \frac{1}{2}e^{-\theta c} + \frac{1}{2}e^{\theta c} = \cosh(\theta c) \leq e^{\theta^2 c^2/2}.$$

Here we used the inequality $\cosh t \leq e^{t^2/2}$, valid for all real t. This implies that $\mathbb{E}e^{\theta S_n} = \prod_i \mathbb{E}e^{\theta X_i} \leq e^{n\theta^2 c^2/2}$. Hence, for any $\theta > 0$,

$$\mathbb{P}(S_n > nx) = \mathbb{P}(e^{\theta S_n} > e^{\theta n x}) \leq e^{-\theta n x}\mathbb{E}e^{\theta S_n} \leq e^{-\theta n x}e^{n\theta^2 c^2/2} = e^{\theta n(\theta c^2/2 - x)}.$$

The last exponent is minimized for $\theta = \theta^* = x/c^2$ and its minimum value is $\theta^* n(\theta^* c^2/2 - x) = -x^2 n/2c^2$. Hence $\mathbb{P}(S_n > nx) \leq e^{-nx^2/2c}$. Reversing the roles of X_i and $-X_i$, we have $\mathbb{P}(-S_n < -nx) \leq e^{-nx^2/2c}$ also. □

A3 Convergence of Markov chains to diffusions

Recall that sequence of real-valued random variables $Z := (Z_k, k = 0, 1, \ldots)$ is said to be a time-homogeneous Markov chain if $\mathbb{P}(Z_{k+1} \leq y \mid Z_k = x, Z_{k-1} \ldots, Z_0) = \mathbb{P}(Z_{k+1} \leq y \mid Z_k = x) = \mathbb{P}(Z_1 \leq y \mid Z_0 = x)$, for all k, x and y. Often, Markov chains depend on a parameter which, without loss of generality, we can take to be an integer n. Let $Z_n := (Z_{n,k}, k = 0, 1, \ldots)$, $n = 1, 2, \ldots$ be such a sequence. Frequently, it makes sense to consider this process at a time scale that depends on n. The problem then is to choose (if any) a sequence τ_n of positive real numbers such that $\lim_{n\to\infty} \tau_n = \infty$ and study instead the random function

$$Z_{n,[\tau_n t]}, \quad t \geq 0,$$

which, hopefully, may have a limit in a certain sense. This can be made continuous by linear interpolation. That is, consider the random function $Z_n(t)$, $t \geq 0$, defined by

$$Z_n(t) := Z_{n,[\tau_n t]} + (\tau_n t - [\tau_n t])(Z_{n,[\tau_n t]+1} - Z_{n,[\tau_n t]}). \tag{41}$$

We seek conditions under which the sequence of random continuous functions $(Z_n(t), t \geq 0)$ converges weakly to a random continuous function $(Z(t), t \geq 0)$.

To define weak convergence, we first define the notion of convergence in $C[0,\infty)$ by saying that a sequence of continuous functions f_n converges to a continuous function (write $f_n \to f$) if $\sup_{0\leq t\leq T} |f_n(t) - f(t)| \to 0$, as $n \to \infty$, for all $T \geq 0$. Then we say that $\varphi : C[0,\infty) \to \mathbb{R}$ is continuous if, for all continuous functions f, $\varphi(f_n) \to \varphi(f)$ whenever $f_n \to f$. Finally, we say that the sequence of random continuous functions Z_n converges weakly [8] to the random continuous function Z if $\mathbb{E}\varphi(Z_n) \to \mathbb{E}\varphi(Z)$, as $n \to \infty$, for all continuous $\varphi : C[0,\infty) \to \mathbb{R}$.

We now quote, without proof, a useful theorem that enables one to deduce weak convergence to a random continuous function that satisfies a stochastic differential equation. For the whole theory we refer to Stroock and Varadhan [22, Chapter 11].

Theorem 5. *Let, for each* $n \in \mathbb{N}$, $Z_{n,k}$, $k = 0, 1, \ldots$, *be a sequence of real random variables forming a time-homogeneous Markov chain. Assume there is a sequence* τ_n, *with* $\tau_n \to \infty$, *such that*

$$\tau_n \, \mathbb{E}[(Z_{n,k+1} - Z_{n,k})^2 \mid Z_{n,k}^{(n)} = x] \to \sigma^2(x), \tag{42}$$

uniformly over $|x| \le R$ *for all* $R > 0$, *for some continuous function* $\sigma^2(x)$. *For the same* τ_n, *we also assume that*

$$\tau_n \, \mathbb{E}[Z_{n,k+1} - Z_{n,k} \mid Z_{n,k} = x] \to b(x), \tag{43}$$

uniformly over $|x| \le R$ *for all* $R > 0$, *for some continuous function* $b(x)$. *Assume also that, for all* $R > 0$, *there are positive constants* c_1, c_2, p, *such that*

$$\tau_n \mathbb{P}(|Z_{n,k+1} - Z_{n,k}| > \varepsilon \mid Z_{n,k} = x) \le c_1 e^{-c_2 \varepsilon^p n}, \tag{44}$$

for all $\varepsilon > 0$ *and all* $|x| \le R$. *Finally, assume that there is* $x_0 \in \mathbb{R}$ *such that* $\mathbb{P}(|Z_{n,0} - x_0| > \varepsilon) \to 0$, *for all* $\varepsilon > 0$. *Then, as* $n \to \infty$, *the sequence of random continuous functions defined as in* (41) *converges weakly to the solution of the stochastic differential equation*

$$dZ(t) = b(Z(t))dt + \sigma(Z(t))dW_t, \quad Z_0 = x_0,$$

provided that this equation admits a unique strong solution.

This conditions in Theorem 5 are much stronger than those of [22, Theorem 11.2.3]. However, it is often the case that the conditions can be verified. This is indeed the case in this paper.

Acknowledgments. The authors would like to thank Andrew Heunis and Svante Janson for their comments on this paper.

References

[1] U. Abel, and M. Ivan. (2009). Over–iterates of Bernstein's operators: a short and elementary proof. *Amer. Math. Monthly* **116**, 535-538.

[2] Richard F. Bass. (2011). *Stochastic Processes.* Cambridge University Press.

[3] S. N. Bernstein (1912/13). Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités. *Commun. Soc. Math. Kharkow* **13**, 1-2.

[4] P. Billingsley (1995). *Probability and Measure*, 3d Ed. John Wiley, New York.

[5] Jérôme Casse, Jean-Franois Marckert (2015). Processes iterated ad libitum. arXiv:1504.06433 [math.PR]

[6] Nicolas Curien and Takis Konstantopoulos (2014). Iterating Brownian motions, *ad libitum. J. Theor. Probability* **27**, No. 2, 433-448.

[7] Ronald A. DeVore and George G. Lorentz (1993). *Constructive Approximation.* Springer-Verlag, Heidelberg.

[8] Stewart N. Ethier and Thomas G. Kurtz (1986). *Markov Processes: Characterization and Convergence.* Wiley.

[9] Stewart N. Ethier and M. Frank Norman (1977). Error estimate for the diffusion approximation of the Wright-Fisher model. *Proc. Natl. Acad. Sci. USA* **74**, No. 11, 5096-5098.

[10] Ronald A. Fisher (1930). *The Genetical Theory of Natural Selection.* Clarendon Press, Oxford.

[11] Ronald A. Fisher (1930). On the dominance ratio. *Proc. Roy. Soc. Edinburgh.* **42**, 321-341.

[12] Charles M. Grinstead and J. Laurie Snell (1997). *Introduction to Probability. American Math. Society*, Providence.

[13] Wassily Hoeffding (1963). Probability inequalities for sums of bounded random variables. *J. Amer. Stat. Assoc.* **58**, 13-30.

[14] Marc Kac (1937). Une remarque sur les pôlynomes de M. S. Bernstein. *Studia Math.* **7**, 49-51.

[15] Howard M. Taylor and Samuel Karlin (1981). *A second course in stochastic processes.* Academic Press, New York.

[16] S. Karlin and Z. Ziegler (1970). Iteration of positive approximation operators. *J. Approx. Th.* **3**, 310-339.

[17] R.P. Kelisky and T.J. Rivlin (1967). Iterates of Bernstein polynomials. *Pacific J. Math.* **21**, 511-520.

[18] Motoo Kimura (1964). *Diffusion Models in Population Genetics.* Methuen, London.

[19] Bernt Øksendal (2003). *Stochstic Differential Equationsa, 6th Ed.* Springer-Verlag, Berlin.

[20] Allan Pinkus (2000). Weierstrass and approximation theory. *J. Approx. Th.* **107**, 1-66.

[21] Albert N. Shiryayev (1984). *Probability.* Springer-Verlag, New York.

[22] Daniel W. Stroock and S.R.S. Varadhan (1979). *Multidimensional Diffusion Processes.* Springer-Verlag, Berlin.

[23] Tat Dat Tran, Julian Hofrichter and Jürgen Jost (2013). An introductin to the mathematical structure of the Wright-Fisher model of population genetics. *Theor. Biosci.* **132**, 73-82.

[24] E. Voronovskaya (1932). Determination de la forme asymptotique d'approximation des fonctions par les polynomes de M. Bernstein. *Dokl. A*, 79-85; Ch. 10:307.

[25] Karl Weierstrass (1885). Über die analytische Darstellbarkeit sogennanter willkürlicher Functionen [*sic*] einer reelen Veränderlichen. *Sitzungsberichte der Köninglich Preußischen Akad. der Wissensch. zu Berlin*, 663-639, 789-805.

[26] Sewall Wright (1931). Evolution in Mendelian populations. *Genetics* **16**, 97-151.