Density and Distribution Function estimation through iterates of fractional Bernstein Operators
Claude Manté

To cite this version:
Density and Distribution Function estimation through iterates of fractional Bernstein Operators

Claude Manté, Aix-Marseille Université, Université du Sud Toulon-Var, CNRS/INSU, IRD, MIO, UM 110, Campus de Luminy Marseille, France, claude.mante@mio.osupytheas.fr

Abstract. We describe a method for distribution function and density estimation with Bernstein polynomials. We take advantage of results about the eigenstructure of the Bernstein operator to refine the Sevy’s convergence acceleration method, based on iterates of this operator; the original Sevy’s algorithm is improved by introducing fractional operators. The proposed algorithm has better convergence properties than the classical one; the price to pay is a controllable loss of the shape-preserving properties of the Bernstein approximation (monotonicity and positivity in the Density Estimation setting). The method is tested on simulated data.

Keywords. Density Estimation, Bernstein operator, root of operators, Bernstein polynomials, Lagrange polynomials

1 Introduction

Bernstein simultaneously introduced in 1912 the polynomials and the operator that bear his name in a famous paper [2]. But, as Farouki [8] noticed, this approximation has been seldom used, due to its slow convergence. For instance, to approach \( f(t) = t^2 \) on the unit interval with a maximal error of \( 10^{-4} \), we need a polynomial of degree 2500 [8]! Nevertheless, this operator (denoted \( B_n \)) has attractive shape-preserving properties: if \( f \) is positive (or monotone, or convex), its image \( B_n[f] \) is so (see [5] for further properties). Consequently, the structure of a distribution function (d.f.) is preserved by \( B_n \); this point strongly motivated the use of the Bernstein approximation in Density Estimation [19, 1, 3, 12, 13, 14].

Notations

We will work in the Banach space \( C[0,1] \) of continuous functions on \([0,1]\), equipped with the norm \( \| f \| := \max_{x \in [0,1]} |f(x)| \). The subspace of polynomials of degree \( \leq n \) will be denoted \( \mathcal{P}_n \), and
The Bernstein operator \( B_n : C[0, 1] \to C[0, 1] \) is defined by [4, 15, 18]:
\[
B_n[f](x) := \sum_{j=0}^{n} w_{n,j}(x) f \left( \frac{j}{n} \right)
\]
with \( w_{n,j}(x) := \binom{n}{j} x^j (1-x)^{n-j} \); its range \( \mathcal{R}(B_n) \subseteq \mathfrak{P}_n \). Cooper and Waldron [4] gave its spectral decomposition, which can be also written into the form hereunder [16].

**Theorem 2.1.** The Bernstein operator can be represented in the diagonal form
\[
B_n[f] = \sum_{j=0}^{n} \lambda_j^{[n]} \pi_j^{[n]} \otimes \pi_j^*[n] (\mathcal{L}_n[f])
\]
where \( f \in C[0,1], \lambda_j^{[n]} = \frac{n!}{(n-j)!j!} \in [0,1] \) and \( \pi_j^{[n]} \in \mathfrak{P}_n \) are its eigenvalues and eigenvectors, \( \pi_j^*[n] \) is the dual vector of \( \pi_j^{[n]} \), and \( u \otimes v^*[w] := u \langle v^*, w \rangle \).

We will need the Lagrange interpolation operator (equispaced case) \( \mathcal{L}_n : C[0,1] \to C[0,1] \) defined by:
\[
\mathcal{L}_n[f](x) := \sum_{j=0}^{n} \ell_{n,j}(x) f \left( \frac{j}{n} \right),
\]
with
\[
\ell_{n,j}(x) := \prod_{k=0, k \neq j}^{n} \frac{n x - k}{j - k}.
\]

Three bases of \( \mathfrak{P}_n \) will be needed:
1. the Bernstein’s basis \( W_n := \{ w_{n,j}(x), 0 \leq j \leq n \} \)
2. the Lagrange’s basis \( L_n := \{ \ell_{n,j}(x), 0 \leq j \leq n \} \)
3. the eigenvectors of \( B_n, \Pi_{[n]} := \{ \pi_j^{[n]}(x), 0 \leq j \leq n \} \).
Let us denote $LW_{[n]}$ the transformation matrix associated with the bases $L_n$ and $W_n$, whose $j^{th}$ column consists in the coordinates of $w_{n,j}$ in the basis $L_n$. The following results can be easily demonstrated [16]:

**Lemma 2.2.** $Mat\left(\overset{o}{B}_n; L_n, W_n\right) = I_n$ and $Mat\left(\overset{o}{B}_n; W_n, W_n\right) = LW_{[n]}$.

Thank to this lemma, we obtain for any $k \geq 2$ a first matrix representation of $B_n^k$ from the diagram:

$$B_n^k : C [0, 1] \xrightarrow{\mathcal{L}_n} (\mathfrak{P}_n, L_n) \xrightarrow{I_n} (\mathfrak{P}_n, W_n) \xrightarrow{LW_{[n]}^{k-1}} (\mathfrak{P}_n, W_n).$$

(1)

Besides, Theorem 2.1 gives an alternative representation of this operator:

$$B_n^k : C [0, 1] \xrightarrow{\mathcal{L}_n} (\mathfrak{P}_n, L_n) \xrightarrow{\Pi_{[n]}} (\mathfrak{P}_n, \Pi_{[n]}) \xrightarrow{\Lambda_{[n]}} (\mathfrak{P}_n, \Pi_{[n]}) \xrightarrow{\Pi_{W_{[n]}}} (\mathfrak{P}_n, W_n)$$

where $\Lambda_{[n]}$ is the diagonal matrix associated with the vector

$$\left(1, 1, 1 - 1/n, (3n - 2)/n^2, \ldots, n!/n^n\right)$$

of eigenvalues of $B_n$, and $\Pi_{[n]}$ and $\Pi_{W_{[n]}}$ are transformation matrices associated with the three bases.

### 3 Sevy’s sequences for d.f. and density approximation

We saw that in the elementary case $f(t) = t^2$, the speed of convergence of $B_n [f]$ towards $f$ is only $O \left(\frac{1}{n}\right)$ [8]; the situation is worse in the special case of d.f.s, since it can be proven [15] that one should rather expect $O \left(\frac{1}{\sqrt{n}}\right)$. To get a sequence of approximations converging faster than $B_n$, Sevy [17] proposed to supersede $B_n$ by the iterated operator

$$\mathcal{J}_n^I := \left(1 - (1 - B_n)^I\right)$$

and proved the following result.

**Theorem 3.1.** ([18], see also [4]) For $n \geq 1$ fixed, and any function $F$ defined on $[0,1]$, we have:

$$\left\|\mathcal{J}_n^I [F] - \mathcal{L}_n [F]\right\| \underset{I \to \infty}{\longrightarrow} 0.$$  

Such a sequence build a bridge between $\mathcal{J}_n^I [F] = B_n [F]$ and $\mathcal{L}_n [F]$. It is worth noting that $\mathcal{L}_n [F]$ interpolates the data but can be very bumpy and that in the equispaced case, the interpolation errors are maximal ([6, Ch. 2]; [11, Ch. 5]). Suppose now $F$ is a d.f.; $B_n [F]$ is also a d.f., but in general $\mathcal{L}_n [F]$ will not share the same characteristics. Thus, it is natural to try to determine some optimal number of iterations $I^* \geq 1$ in order that $\mathcal{J}_n^{I^*} [F]$ has the structure of a d.f., while $\mathcal{J}_n^{I^*+1} [F]$ has not. In other words, the density approximation $\tilde{f}_n^{(I^*)} (x) := \frac{d}{dx} \mathcal{J}_n^{I^*} [F] (x)$ should be *bona fide*, i.e. should belong to $\mathcal{F}^+ \cap \mathcal{F}^1$, while $\tilde{f}_n^{(I^*+1)} (x) \notin \mathcal{F}^+ \cap \mathcal{F}^1$ (see [15]).
4 Interpolating Sevy’s sequences (see [16])

To refine Sevy’s sequences, we build for $K \geq 2$ the $K^{th}$ “root” of the operator $G_n := (1 - B_n)$ involved in Formula 3. Because $B_n$ only preserves $\mathcal{P}_1$, the eigenvalues of $G_n$ belong to $[0, 1]$. Thus, thanks to classical results about convergent series of operators (see [10] for instance), one may consistently define the fractional operator

$$G_n^{(1/K)} := \exp\left(\frac{1}{K} \log (G_n)\right).$$

One can easily verify the following lemma.

**Lemma 4.1.** $\forall I \geq 1$,

$$I_n^I = (1 - (1 - B_n)^I) = \left(1 - \left(1 - \tilde{B}_n\right)^I\right) \circ \mathcal{L}_n.$$

Consequently, we can proceed as if $f \in \mathcal{R}(\mathcal{L}_n)$ and don’t have to worry about the “Lagrange residual” $f - \mathcal{L}_n[f]$. Since $\mathcal{P}_1$ is preserved by $B_n$ and because of Lemma 4.1, $\mathfrak{I}_n^k (f) = \mathcal{L}_1 [f] + \mathfrak{I}_n^k (\mathcal{L}_n[f] - \mathcal{L}_1 [f])$, and we can set the definition of K-fractional Sevy’s sequences.

**Definition 4.2.** Let $K \geq 2$, and $f \in C[0, 1]$. The K-fractional Sevy’s sequence of approximations of $f$ is:

$$\mathfrak{I}^j_{n;K} [f] := \mathcal{L}_1 [f] + \left(1 - G_n^{(j/K)}\right) (\mathcal{L}_n[f] - \mathcal{L}_1 [f]) , j \geq 1.$$

Such a sequence interpolates the original one, since $\mathfrak{I}^j_{n;K} [f] = \mathfrak{I}^j_n (f)$. Its matrix representation stems from diagram 2.

**Lemma 4.3.** $\text{Mat} \left(\mathfrak{I}^j_{n;K}; L_n, W_n\right) = \Pi W_n \circ \Lambda^{(j/K)} W_n \circ L \Pi W_n$, where $\Lambda^{(j/K)} W_n$ is the diagonal matrix associated with the vector

$$\left(1, 1, 1 - \left(\frac{1}{n}\right)^{(j/K)}, 1 - \left(\frac{3n - 2}{n^2}\right)^{(j/K)}, \cdots, 1 - \left(\frac{n!}{n^n}\right)^{(j/K)}\right).$$

5 Numerical issues

Because of Lemmas 2.2 and 4.1, building a classical Sevy’s sequence amounts to compute powers of the transformation matrix $L W_n$ (see diagram 1). The condition number of this matrix in the $\ell^2$-norm is [7]:

$$\frac{\|L W_n\|_2}{\|L W_n - 1\|_2} = \frac{\lambda^{(n)}_n}{\lambda^{(n)}_1} \approx \frac{n^n}{n^2 \sqrt{2\pi n}}$$

(assymptotically - see [9]). Thus, $L W_n$ is ill-conditioned, and one must expect to meet numerical problems when $n$ is big. The situation is potentially worse for fractional sequences, since Lemma 4.3 shows that the matrix of the restricted operator depends on both the ill-conditioned transformation matrices $L \Pi W_n$ and $W_n$ (see Figure 1).
But the point for us is merely to control numerical errors in computing $\mathcal{J}_{n,K}^j[f]$! Notice that on the one hand $\text{Mat} \left( B_n; L_n, W_n \right) = I_n$ (Lemma 2.2), while on the other hand

$$\text{Mat} \left( B_n; L_n, W_n \right) = \Pi W_n \circ \Lambda_n \circ L \Pi_n$$

(diagram 2). Consequently, the matrix norms

$$\begin{align*}
\left\| \Pi W_n \circ \Lambda_n \circ L \Pi_n - I_n \right\|_1 \\
\left\| \Pi W_n \circ \Lambda_n \circ L \Pi_n - I_n \right\|_\infty
\end{align*}$$

are convenient indicators of loss of numerical accuracy imputable to the ill-conditioning of the transformation matrices. Since the only easy-to-handle basis is the power basis, the transformation matrices $PL_n$, $P \Pi_n$, and $PW_n$ are straightforwardly computed, and we can write:

$$\begin{align*}
\Pi W_n &= P \Pi_n^{-1} \circ PW_n \\
L \Pi_n &= PL_n \circ P \Pi_n
\end{align*}$$

(formally). But we can derive from Figure 1 that these inverse matrices cannot be computed with sufficient accuracy in general. Thus, it’s necessary to supersede in (6) the inverse matrices by the Moore-Penrose generalized inverses $P \Pi_n^+$ and $PL_n^+$. This gives rise to the regularized operators:

$$\begin{align*}
\overline{\Pi W}_n := P \Pi_n^+ \circ PW_n \\
\overline{L \Pi}_n := PL_n^+ \circ P \Pi_n.
\end{align*}$$

On Figure 2, we plotted the logarithm of the second indicator of Formula (5), for $n$ ranging from 1 to 35 (a similar graph can be obtained for the first indicator). Two cases must be distinguished on this plot: the “symbolic” one, where polynomial eigenfunctions were computed from the recurrence formula given by [4], and the “numerical” one, where they were computed by polynomial interpolation of the eigenvectors of $L \Pi_n$, giving rise to the alternative basis $\hat{\Pi}_n := \{ \hat{\pi}_j^n(x), 0 \leq j \leq n \}$. Of course, we should have $\hat{\pi}_j^n = \pm \pi_j^n \forall 0 \leq j \leq n$ if there were
Density and d.f. estimation through iterated Bernstein Operators

Figure 2. The sequences \( \{ \log \left( \left\| \Pi W_{\lfloor n \rfloor} \circ \Lambda_{\lfloor n \rfloor}^T \circ \tilde{L} \Pi_{\lfloor n \rfloor} - I_n \right\|_\infty \right) \}, 1 \leq n \leq 35 \} \) and
\( \{ \log \left( \left\| \Pi W_{\lfloor n \rfloor} \circ \Lambda_{\lfloor n \rfloor}^T \circ \tilde{L} \Pi_{\lfloor n \rfloor} - I_n \right\|_\infty \right) \}, 1 \leq n \leq 35 \} ; \Lambda_{\lfloor n \rfloor}^T \) is the diagonal matrix obtained by setting to zero each eigenvalue \(< 10^{-12}\).

not different roundoff errors on both sides, imputable to different algorithms! That is why we took into account the numerical rank of \( B_n \), discarding from the computation of Formula (5) eigenvectors associated with eigenvalues smaller than \( 10^{-12} \) (see Figure 2 and its legend).

It is worth noting that the computational cost in the symbolic case is considerable: it took about 6600 seconds to produce the symbolic part of Figure 2, while the numeric part was obtained in 80 seconds.

6 Application to density and d.f. estimation

Suppose \( F \) is some differentiable d.f. associated with a random variable \( X \) defined on \([0, 1]\), and that \( S_N := \{ X_1, \cdots, X_N \} \) is a \( N \)-sample of \( X \), giving rise to the empirical d.f. \( F_N(x) \). Babu et al. [1] proposed to estimate \( F \) by a Bernstein polynomial \( \tilde{F}_{N,m} \) of degree \( m \):

\[
\tilde{F}_{N,m}(x) := \sum_{k=0}^{m} F_N \left( \frac{k}{m} \right) w_{m,k}(x) = B_m[F_N] \tag{8}
\]

with \( m \leq m_0 := \lceil N/\log(N) \rceil \). The proposed method consists in superseding \( B_{m_0}[F_N] \) by some \( B_{m^*,K}[F_N] \), where \( m^* \leq m_0 \) and \( I^* \geq K \) (fixed) are convenient values of the degree of the estimator and of the number of iterations in Definition 4.2.

As an illustration, we displayed first on Figure 3 the results obtained with a sample of size 200 of \( \beta(3, 12) \), with \( K = 10 \). We found that \( I^* = 32 \) iterations of the fractional operator (4) simultaneously corresponded to a satisfactory fit of the e.d.f. and an approximately \( \text{bona fide} \) density estimation. Thus, in this case, the fractional number of iterations was \( r^* = 1 + \frac{22}{35} \). On this plot, we superimposed to the true d.f. three estimators: the Babu’s one, of degree \( m_0 = 38\),
Figure 3. Estimation of the $\beta(3,12)$ d.f. and density from a sample. Left panel: the true d.f. (orange), the Babu’s one (gray and dashed, of degree $m_0 = \lceil 200/\log(200) \rceil = 38$), the classical Bernstein estimator of degree $m = 34$ (gray), and the proposed one (black), of degree 34 too. Right panel: density estimators obtained by deriving the d.f.s estimated.

the Bernstein estimator of degree $m = 34$, and the iterated estimator (black), of degree 34 too. The density estimators are derivatives of these d.f.s.

In addition, we collected in Table 1 results from simulations carried on with 30 samples of size $N = 150$ ($\Rightarrow m_0 = 30$) of four Beta distributions. For sake of simplicity, we fixed $I^* = 20$ (see [16] for a theoretical justification). For each one of these samples and for each estimator (4 estimators of the d.f. and 3 estimators of the density, since the e.d.f. is not differentiable), the Integrated Squared Error (ISE) $\int (\hat{F}(x) - F(x))^2 \, dx$ and the $L^1$ error norm $\int |\hat{f}(x) - f(x)| \, dx$ were computed. Clearly, even in this suboptimal situation ($I^* = 20$), the proposed estimators outperformed classical ones, excepted in the very simple case $\beta(1,2)$ (uniform distribution).

Notice the honorable performances of the good old e.d.f.!

Table 1. Simulations results. First group of columns: the distribution simulated, and optimal value of $m$ (for further details, see [16]); second group: median of $10^3$ ISE of estimated distribution functions; third group: median of the $L^1$ error norms for estimated densities. Best result are in bold characters.

<table>
<thead>
<tr>
<th>Probability</th>
<th>$m^*$</th>
<th>c.d.f.</th>
<th>$B_{30}$</th>
<th>$B_{m^*}$</th>
<th>$\gamma_{30}^{20}$</th>
<th>$B'_{30}$</th>
<th>$B'_{m^*}$</th>
<th>$\gamma_{30}^{20}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta(1,2)$</td>
<td>16</td>
<td>0.497</td>
<td>0.415</td>
<td><strong>0.38</strong></td>
<td>0.569</td>
<td>0.1</td>
<td><strong>0.09</strong></td>
<td>0.108</td>
</tr>
<tr>
<td>$\beta(2,4)$</td>
<td>18</td>
<td>0.6</td>
<td>0.51</td>
<td>0.56</td>
<td><strong>0.368</strong></td>
<td>0.108</td>
<td>0.12</td>
<td><strong>0.099</strong></td>
</tr>
<tr>
<td>$\beta(3,12)$</td>
<td>25</td>
<td>0.32</td>
<td>0.783</td>
<td>0.908</td>
<td><strong>0.258</strong></td>
<td>0.197</td>
<td>0.207</td>
<td><strong>0.118</strong></td>
</tr>
<tr>
<td>$\beta(10,10)$</td>
<td>25</td>
<td>0.318</td>
<td>1.16</td>
<td>1.37</td>
<td><strong>0.289</strong></td>
<td>0.248</td>
<td>0.263</td>
<td><strong>0.153</strong></td>
</tr>
</tbody>
</table>

Bibliography


---

1In Section 6, the default iteration number has been erroneously set to 20. The correct value is indeed 2!
Density and d.f. estimation through iterated Bernstein Operators

and Inference, 105, 377-392.


