Approximation of Cumulative Distribution Functions by Bernstein Phase-Type Distributions*

A. Horváth[a], I. Horváth[b,c], M. Paolieri[d], M. Telek[b,c], Enrico Vicario[e]

[a] *Department of Computer Science, University of Turin, Italy*
[b] *Department of Networked Systems and Services, Budapest University of Technology and Economics, Hungary*
[c] *HUN-REN-BME Information Systems Research Group, Hungary*
[d] *Department of Electrical and Computer Engineering, University of Southern California, USA*
[e] *Department of Information Engineering, University of Florence, Italy*

Abstract

The inclusion of generally distributed random variables in stochastic models is often tackled by choosing a parametric family of distributions and applying fitting algorithms to find appropriate parameters. A recent paper proposed the approximation of probability density functions (PDFs) by Bernstein exponentials, which are obtained from Bernstein polynomials by a change of variable and result in a particular case of acyclic phase-type distributions. In this paper, we show that this approximation can also be applied to cumulative distribution functions (CDFs), which enjoys advantageous properties and achieves similar accuracy; by focusing on CDFs, we propose an approach to obtain stochastically ordered approximations. The use of a scaling parameter in the approximation is also presented, evaluating its effect on approximation accuracy.

Keywords: Bernstein Polynomials, Phase-Type Distributions, Markov Chains, Analytic Approximation

1. Introduction

Continuous-time models of stochastic systems frequently need to include random variables with general (i.e., non-exponential) probability distributions, to represent properties enforced by design (e.g., periodic releases or deterministic timeouts) or by contract (e.g., service times guaranteed along the development process or by some agreed service level objective), or to fit observed data or learned parameters. A standard approach is to select a parametric family of probability distributions and to apply fitting algorithms to find appropriate parameter values to approximate the observed random variables. An ideal family of distributions should be sufficiently general to result in accurate approximations, but it should also support simple fitting procedures and allow efficient analysis (or simulation) of the resulting system model.

The family of phase-type (PH) distributions [1], defined as the time to absorption in Markov chains, is broadly used to approximate general random variables. By varying the number of phases in the Markov chain, this family allows a tradeoff between accuracy of the approximation and analysis cost of the resulting model. PH parameter fitting methods include maximum likelihood methods [2, 3], moment matching [4, 5], tail behavior matching [6, 7], or both [8]. Stochastic models with PH type distributions result in underlying Markov chains with regular structures, which can be analyzed efficiently through matrix analytic methods [9, 10].

In [11], an approach was proposed to approximate probability density functions (PDFs) using *Bernstein exponentials* (BEs), i.e., linear combinations of Bernstein polynomials (BP) [12, 13] where the support $[0, 1]$ is mapped to $[0, \infty)$ through a change of variable. This approach results in a subclass of acyclic PH

*I. Horváth and M. Telek were supported by the OTKA K-138208 project of the Hungarian Scientific Research Fund. A. Horváth and E. Vicario were supported by the Italian National Recovery and Resilience Plan (NRRP), partnership on "Telecommunications of the Future" (PE0000001 - program "RESTART", project Net4Future).

distributions, which we refer to as Bernstein phase type (BPH) distributions, that preserve shape properties of approximated density functions and enjoy derivation simplicity as its parameters can be derived in closed form, while allowing efficient model analysis.

In this paper, we study the properties of BPH approximations for cumulative distribution functions (CDFs) rather than PDFs. In fact, BPH approximations guarantee uniform convergence and preserve local shape properties, notably including non-negativity, but they do not preserve integral measure and thus require normalization to obtain valid PDFs with unitary measure. Conversely, we show that, when BPH approximations are applied to CDFs, or to complementary CDFs (CCDFs), the resulting functions are valid cumulative distributions and preserve the important properties of the original CDFs including monotonicity, upper and lower bounds, and exact limit values at 0 and $+\infty$. In particular, we focus on stochastic order for models designed for the evaluation of safe guarantees of system metrics (e.g., quality of service) [14, 15]. We present an approach to obtain BPH approximations that guarantee smaller and greater stochastic order, and characterize the required tail conditions and minimum degree of the BPH approximation. This paper extends our preliminary work [16] by introducing a scaling parameter in BPH approximations and by evaluating its effect on the approximation accuracy. We also present additional experiments showing that BPH approximations based on CDFs achieve similar accuracy to those based on PDFs (while providing technical advantages) and we illustrate their limitations in the approximation of distributions with high coefficient of variation (as common for other general purpose PH fitting approaches).

The paper is organized as follows. In Section 2, we recall background information on Bernstein polynomials, Bernstein exponentials, and PH distributions. In Section 3, we present the properties of BPH approximations of CDFs, while in Section 4 we propose an approach to obtain stochastically ordered BPH approximations. In Section 5, we evaluate our approach numerically to highlight advantages and limitations. In Section 6, we compare the BPH approximations of CDFs with BPH approximations of PDFs. In Section 7, we explore the selection of the scaling parameter in the mapping of the support applied by BEs. Conclusions are drawn in Section 8.

2. Background

2.1. Bernstein Polynomials

For any order $n \in \mathbb{N}$, the Bernstein operator B_n maps a function $G : [0, 1] \to \mathbb{R}$ onto a polynomial defined as [12]:

$$B_n(G; y) := \sum_{i=0}^{n} G\left(\frac{i}{n}\right) \binom{n}{i} y^i (1-y)^{n-i}. \tag{1}$$

The Bernstein operator is linear, i.e., $B_n(\lambda_1 G_1 + \lambda_2 G_2; y) = \lambda_1 B_n(G_1; y) + \lambda_2 B_n(G_2; y)$, and it represents polynomials up to degree n exactly, i.e., $B_n(y^k; y) = y^k$ for any $0 \le k \le n$. $B_n(G; y)$ also preserves many properties of G, which motivated its investigation and wide application as a tool for approximation.

Boundary Conditions, Bounds, Monotonicity. $B_n(G; y)$ is exactly equal to $G(y)$ at the endpoints of the domain $[0, 1]$ and preserves upper and lower bounds, i.e., $G(0) = B_n(G; 0)$, $G(1) = B_n(G; 1)$, and $\forall y \in [0, 1], m \le G(y) \le M \implies \forall y \in [0, 1], m \le B_n(G; y) \le M$. Moreover, if G is monotonic increasing (or decreasing) over $[0, 1]$, so is $B_n(G; y)$. By combination of these properties, if $G(y)$ is a CDF (or a CCDF) with support $[0, 1]$, so is $B_n(G, y)$ for any $n \in \mathbb{N}$, i.e., the Bernstein operator B_n maps distributions to valid distributions.

Uniform Convergence. For any continuous function G, the Bernstein operator ensures asymptotic convergence to 0 of the error $|G(y) - B_n(G; y)|$ when $n \to \infty$, uniformly over the entire support $[0, 1]$:

$$\forall \epsilon > 0, \exists \bar{n} \in \mathbb{N} \text{ such that } n > \bar{n} \implies \forall y \in [0, 1], |G(y) - B_n(G; y)| < \epsilon. \tag{2}$$

For further related results and explicit bounds we refer to [13].

2.2. Bernstein Exponentials

The BE operator extends the Bernstein operator to the class of bounded functions with infinite support $[0,\infty)$ through the change of variables $y = e^{-x}$ (i.e., $x = -\log(y)$) which maps the support $[0,1]$ onto $[0,\infty)$.

According to this, for any order $n \in \mathbb{N}$, the BE operator maps a function $F : [0,\infty) \to \mathbb{R}$ onto an exponential mixture of the form:

$$BE_n(F;x) := \sum_{i=0}^{n} F\left(\log\left(\frac{n}{i}\right)\right) \binom{n}{i} e^{-ix}(1-e^{-x})^{n-i}. \tag{3}$$

By design, the BE operator inherits various shape preservation properties of the Bernstein operator. Since the change of variables $y = e^{-x}$ is continuous and strictly monotonic, $BE_n(F;x)$ is exactly equal to $F(x)$ for $x = 0$ and $x \to \infty$, it preserves the bounds of F, and if $F(x)$ is monotonic, so is $BE_n(F;x)$. Moreover, $BE_n(F;x)$ converges to $F(x)$ uniformly over $[0,\infty)$ as $n \to \infty$.

2.3. Probability Distributions

Given a probability space $(\Omega, \mathcal{F}, P)$ and a random variable $X : \Omega \to \mathbb{R}$, its CDF and CCDF are defined, respectively, as $F(x) := P(X \le x)$ and $\bar{F}(x) := P(X > x) = 1 - F(X)$ for all $x \in \mathbb{R}$; if $F(x)$ is continuous, we say that X is a continuous random variable and the PDF of X is defined as $f(x) := \frac{d}{dx}F(x)$. Given a nonnegative random variable $X : \Omega \to [0,\infty)$, we say that the function $F : [0,\infty) \to [0,1]$ is a *defective CDF* for X if $\lim_{x\to\infty} F(x) < 1$ and we say that it has *probability mass at* 0 if $F(0) > 0$.

2.4. Phase-Type Distributions

A degree n continuous-time PH distribution is given by the time to absorption in a continuous-time Markov chain (CTMC) with n transient states and one absorbing state. A PH distribution that can be represented by a CTMC without cycles is called acyclic PH.

A graphical example of PH distribution is provided in Figure 1 where the absorbing state is colored in gray. The initial probability vector and the infinitesimal generator of the CTMC are $q = \begin{pmatrix} 0.4 & 0 & 0.6 & 0 \end{pmatrix}$ and

$$Q = \begin{pmatrix} -5.2 & 3 & 2.2 & 0 \\ 1.2 & -2.5 & 0.5 & 0.8 \\ 4 & 2.3 & -7.55 & 1.25 \\ 0 & 0 & 0 & 0 \end{pmatrix}.$$

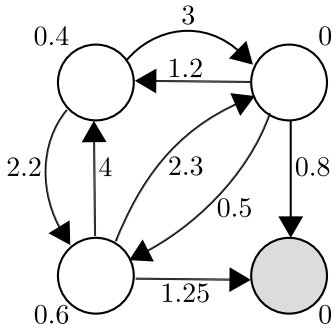

Figure 1: A degree 3 PH distribution

Note that the last entry of q (in this case zero) and the last column of Q can be calculated from the rest (the row corresponding to the absorbing state is filled with zeros). Accordingly, the most widely used representation of a PH distribution includes only the parts of the initial probability vector and of the infinitesimal generator that correspond to transient states. For the above example, the representation is the vector-matrix pair

$$a = \begin{pmatrix} 0.4 & 0 & 0.6 \end{pmatrix}, \quad A = \begin{pmatrix} -5.2 & 3 & 2.2 \\ 1.2 & -2.5 & 0.5 \\ 4 & 2.3 & -7.55 \end{pmatrix}.$$

Given a vector-matrix pair (a, A), the corresponding PH distribution will be denoted by $\mathrm{PH}(a, A)$. The PDF, the CDF, and the CCDF of $\mathrm{PH}(a, A)$ will be denoted and can be calculated as

$$f_{a,A}(x) = ae^{xA}(-A\mathbf{1}),\ F_{a,A}(x) = 1 - ae^{xA}\mathbf{1},\ \text{and}\ \bar{F}_{a,A}(x) = ae^{xA}\mathbf{1}$$

where $\mathbf{1}$ denotes the column vector of ones.

In [11] it was shown and illustrated numerically through several examples that normalized BE approximation of a PDF results in a PH distribution. [11] provides also a more detailed description of the characteristics of BP and BE.

3. Approximation of Cumulative Density Functions by Bernstein Phase-Type Distributions

The degree n BE approximation of a given CDF $F(x)$ with support $[0,\infty)$ is

$$\hat{F}_n(x) = \sum_{i=0}^{n} F\left(\log \frac{n}{i}\right) \cdot T_{n,i}(x) \tag{4}$$

where

$$T_{n,i}(x) := \binom{n}{i} e^{-ix}(1-e^{-x})^{n-i}\,.$$

The division by zero in case of $i = 0$ is resolved by considering the limiting value of $F(x)$ as x tends to infinity, i.e., $F\left(\log \frac{n}{0}\right) = \lim_{x\to\infty} F(x)$ which is equal to 1 if the CDF is not defective (we will denote this limit also simply by $F(\infty)$). At the other end, for $i = n$ we have $F(\log(n/n)) = F(0)$ which is 0 if there is no probability mass at 0 in the distribution.

The same BE can be obtained based on the CCDF.

Proposition 1. *Let $\bar{F}(x)$ be the CCDF of a given CDF $F(x)$, i.e., $\bar{F}(x) = 1 - F(x)$. The distribution obtained by the degree n BE approximation of $F(x)$, given by $\hat{F}_n(x)$ in (4), is equal to the distribution derived from the degree n BE approximation of $\bar{F}(x)$, i.e., $\hat{F}_n(x) = 1 - \hat{\bar{F}}_n(x)$.*

Proof. The degree n BE approximation of $\bar{F}(x)$ is

$$\begin{aligned}
\hat{\bar{F}}_n(x) &= \sum_{i=0}^{n} \bar{F}\left(\log \frac{n}{i}\right) \cdot \binom{n}{i} e^{-ix}(1-e^{-x})^{n-i} \\
&= \sum_{i=0}^{n} \left(1 - F\left(\log \frac{n}{i}\right)\right) \cdot \binom{n}{i} e^{-ix}(1-e^{-x})^{n-i} \\
&= \sum_{i=0}^{n} \binom{n}{i} e^{-ix}(1-e^{-x})^{n-i} - \hat{F}_n(x) = \left(e^{-x} + (1-e^{-x})\right)^n - \hat{F}_n(x) = 1 - \hat{F}_n(x)
\end{aligned} \tag{5}$$

from which $\hat{F}_n(x) = 1 - \hat{\bar{F}}_n(x)$ directly follows. □

The following theorem shows that the approximation given in (4) corresponds to an acyclic PH distribution.

Theorem 1. *When $F(x)$ is a CDF with support $[0,\infty)$, i.e., $\lim_{x\to\infty} F(x) = 1$, then $F_{a,A}(x) = \hat{F}_n(x)$, where*

$$a = \begin{pmatrix} a_1 & \dots & a_n \end{pmatrix} \text{ with } a_i = F\left(\log \frac{n}{i-1}\right) - F\left(\log \frac{n}{i}\right), \tag{6}$$

and

$$A = \begin{pmatrix} -1 & 1 & 0 & & \dots & \\ 0 & -2 & 2 & 0 & & \dots \\ & & & \ddots & & \\ & \dots & & 0 & -(n-1) & n-1 \\ & & \dots & & 0 & -n \end{pmatrix}. \tag{7}$$

I.e., the CDF of PH(a, A) is equal to the approximation in (4).

The graphical representation of PH(a, A) is shown in Figure 2 (where the role of $F(0)$ and $F(\infty)$ can be explicitly seen).

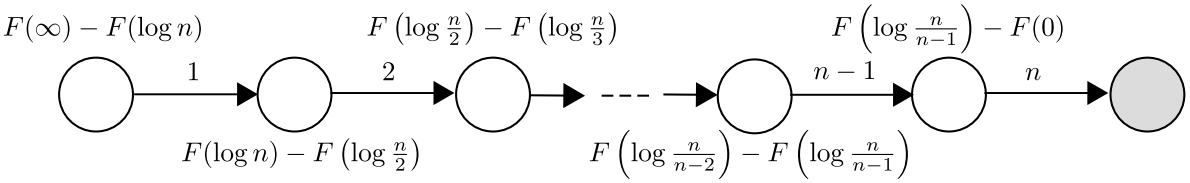

Figure 2: Bernstein PH approximation of a CDF $F(x)$

Proof. The Laplace transform of the PDF of PH(a, A) is

$$f^*_{a,A}(s) = \sum_{i=1}^{n} \left(F\left(\log \frac{n}{i-1}\right) - F\left(\log \frac{n}{i}\right)\right) \prod_{j=i}^{n} \frac{j}{j+s} \tag{8}$$

where the product $\prod_{j=i}^{n} \frac{j}{j+s}$ corresponds to the convolution of $n-i+1$ exponential random variables with rates $i, i+1, ..., n$.

$\hat{F}_n(x)$ is the weighted sum of terms in the form

$$T_{n,i}(x) = \binom{n}{i} e^{-ix}(1-e^{-x})^{n-i} \tag{9}$$

with derivative

$$t_{n,i}(x) = T'_{n,i}(x) = (n-i)\binom{n}{i} e^{-(i+1)x}(1-e^{-x})^{n-i-1} - i\binom{n}{i} e^{-ix}(1-e^{-x})^{n-i}$$

whose Laplace transform is

$$t^*_{n,i}(s) = \int_0^\infty e^{-sx} t_{n,i}(x)dx = \prod_{j=i+1}^{n} \frac{j}{j+s} - \prod_{j=i}^{n} \frac{j}{j+s}\,.$$

Accordingly, the Laplace transform of $\hat{f}(x) = \hat{F}'_n(x)$ is

$$\begin{aligned} \hat{f}^*(s) = \int_0^\infty e^{-sx}\hat{f}(x)dx &= \sum_{i=0}^{n} F\left(\log \frac{n}{i}\right) \left(\prod_{j=i+1}^{n} \frac{j}{j+s} - \prod_{j=i}^{n} \frac{j}{j+s} \right) \\ &= \sum_{i=1}^{n} \left(F\left(\log \frac{n}{i-1}\right) - F\left(\log \frac{n}{i}\right)\right) \prod_{j=i}^{n} \frac{j}{j+s} \end{aligned} \tag{10}$$

from which, by comparing (8) and (10), we have $f^*_{a,A}(s) = \hat{f}^*(s)$. Additionally, using $1 - F_{a,A}(0) = \sum_{i=1}^{n} a_i = F(\infty) - F(0) = 1 - F(0)$ and $F(0) = \hat{F}_n(0)$, $F(\infty) = \hat{F}_n(\infty)$, as discussed after (3), we also have $F_{a,A}(x) = \hat{F}_n(x)$. □

The time domain equivalent of (10) is provided in the following proposition.

Proposition 2. *If $F(0) = 0$, then*

$$\hat{F}_n(x) = \sum_{i=1}^{n} \left(F\left(\log \frac{n}{i-1}\right) - F\left(\log \frac{n}{i}\right)\right) \sum_{j=0}^{i-1} T_{n,j}(x)\,. \tag{11}$$

Proof. From (4) and (9), we have

$$\begin{aligned}
\hat{F}_n(x) &= \sum_{i=0}^{n} F\left(\log\frac{n}{i}\right)\cdot T_{n,i}(x) = \sum_{i=0}^{n} F\left(\log\frac{n}{i}\right)\cdot\left(\sum_{j=0}^{i} T_{n,j}(x) - \sum_{j=0}^{i-1} T_{n,j}(x)\right)\\
&= \sum_{i=0}^{n} F\left(\log\frac{n}{i}\right)\cdot\sum_{j=0}^{i} T_{n,j}(x) - \sum_{i=1}^{n} F\left(\log\frac{n}{i}\right)\sum_{j=0}^{i-1} T_{n,j}(x)\\
&= \sum_{i=0}^{n} F\left(\log\frac{n}{i}\right)\cdot\sum_{j=0}^{i} T_{n,j}(x) - \sum_{i=0}^{n-1} F\left(\log\frac{n}{i+1}\right)\sum_{j=0}^{i} T_{n,j}(x)\\
&= \underbrace{F\left(0\right)}_{0}\cdot\underbrace{\sum_{j=0}^{n} T_{n,j}(x)}_{1} + \sum_{i=0}^{n-1}\left(F\left(\log\frac{n}{i}\right) - F\left(\log\frac{n}{i+1}\right)\right)\sum_{j=0}^{i} T_{n,j}(x)\\
&= \sum_{i=1}^{n}\left(F\left(\log\frac{n}{i-1}\right) - F\left(\log\frac{n}{i}\right)\right)\sum_{j=0}^{i-1} T_{n,j}(x)
\end{aligned}$$

□

From Proposition 2 it also follows directly that

$$\frac{d^k}{dx^k}\sum_{j=0}^{i-1} T_{n,j}(x)|_{x=0} = 0 \qquad \text{for} \quad 0 \le k \le n-i \tag{12}$$

by considering Fig. 2 when the last $n-i$ nodes have 0 initial probability, and, as a further consequence, we also have

$$\frac{d^j}{dx^j} T_{n,i}(x)|_{x=0} = 0 \qquad \text{for} \quad 0 \le j \le n-i-1\,. \tag{13}$$

According to Theorem 1, if $F(x)$ is non-decreasing, $F(0) = 0$ and $F(\infty) = 1$, then its $F_{a,A}(x) = \hat{F}_n(x)$ approximation based on (4) is such that $a_i > 0$ for $i = 1, ..., n$, and $\sum_{i=1}^{n} a_i = 1$.

In a BE approximation the coefficient of the term $T_{n,n}(x)$ is equal to the value of the approximation at zero. Vice versa, the coefficient of the term $T_{n,0}$ is equal to the value of the approximation as $x \to \infty$. This implies the following proposition.

Proposition 3. *Given a CDF $F(x)$ that corresponds to a distribution that is with mass at zero ($F(0) > 0$) and/or defective ($\lim_{x\to\infty} F(x) < 1$), the approximation*

$$\hat{F}_n(x) = 0\cdot T_{n,n}(x) + \sum_{i=1}^{n-1} F\left(\log\frac{n}{i}\right) T_{n,i}(x) + 1\cdot T_{n,0}(x) \tag{14}$$

corresponds to a non-defective distribution without mass at zero. The same can be achieved by approximating the CCDF $\bar{F}(x)$ in the form

$$\hat{\bar{F}}_n(x) = 0\cdot T_{n,0}(x) + \sum_{i=1}^{n-1} \bar{F}\left(\log\frac{n}{i}\right) T_{n,i}(x) + 1\cdot T_{n,n}(x)\,. \tag{15}$$

Note that every PH distribution constructed through a BE approximation has the same infinitesimal generator A given in (7). For this reason, given a vector $a = (a_1\ ...\ a_n)$ the distribution PH(a, A) will be referred to as BPH(a). The PDF, the CDF, and the CCDF of a BPH(a) will be denoted by $f_a(x), F_a(x)$ and $\bar{F}_a(x)$, respectively.

4. Stochastically Smaller and Larger Approximation

In this section, we study the possibility to create BPH distributions that guarantee stochastic order with respect to the distribution we aim to approximate. If $\bar{F}_X(x)$ and $\bar{F}_Y(x)$ are the CCDF of X and Y, then X is stochastically smaller than Y (equivalently, Y is stochastically larger than X) if and only if

$$P(X > z) = \bar{F}_X(z) \leq \bar{F}_Y(z) = P(Y > z), \forall z \geq 0\,. \tag{16}$$

In the sequel we will use the notation

$$\bar{F}_{+\epsilon}(x) = \min(\bar{F}(x) + \epsilon, 1) \;\text{ and }\; \bar{F}_{-\epsilon}(x) = \max(\bar{F}(x) - \epsilon, 0)\,, \tag{17}$$

among which $\bar{F}_{+\epsilon}(x)$ is useful to obtain larger BPH approximations while $\bar{F}_{-\epsilon}(x)$ to obtain smaller ones. In case of $\epsilon > 0$, the distributions corresponding to the CCDFs in (17) are either with mass at zero or are defective. As shown by Proposition 3, we can still easily obtain approximations of them that correspond to non-defective distributions without mass at zero using (14) or (15).

The main result is the following theorem. Intuitively, the theorem guarantees that a stochastically larger (smaller) BPH approximation can be obtained for some sufficiently large degree $\hat{n}$ when: (1) the tail of the CCDF decays faster than an exponential with rate 1 (slower than an exponential with arbitrarily small rate); (2) at 0, the CCDF is not flat (has a finite derivative). Note that this section uses CCDF approximations (equivalent to CDF approximations) for technical convenience.

Theorem 2. *(a) Let $\bar{F}(x)$ be a continuous CCDF with the following property:*

- *$\bar{F}(x) \leq Ce^{-x}$ for some $0 < C < \infty$;*
- *there exists a finite minimal n_d for which $\frac{\mathrm{d}^{n_d}}{\mathrm{d}x^{n_d}}\bar{F}(x)|_{x=0} \neq 0$ (i.e., $\frac{\mathrm{d}^{k}}{\mathrm{d}x^{k}}\bar{F}(x)|_{x=0} = 0$ for $k < n_d$).*

Then for any $0 < \epsilon < 0.5$ there exists an $\hat{n}$ such that for any $n > \hat{n}$, the function

$$\hat{\bar{F}}_{+\epsilon,n}(x) = \sum_{i=1}^{n} \bar{F}_{+\epsilon}\left(\log \frac{n}{i}\right) \cdot T_{n,i}(x) \tag{18}$$

is the CCDF of a BPH distribution that is stochastically larger than $\bar{F}(x)$.

(b) Let $\bar{F}(x)$ be a continuous CCDF with the following properties:

- *$\bar{F}(x) > ce^{-ax}$ for some $0 < c < \infty$ and $a < \infty$;*
- *$\frac{\mathrm{d}}{\mathrm{d}x}\bar{F}(x)|_{x=0}$ is finite.*

Then for any $0 < \epsilon < 0.5$ there exists an $\hat{n}$ such that for any $n > \hat{n}$, the function

$$\hat{\bar{F}}_{-\epsilon,n}(x) = \sum_{i=1}^{n-1} \bar{F}_{-\epsilon}\left(\log \frac{n}{i}\right) \cdot T_{n,i}(x) + T_{n,n}(x) \tag{19}$$

is the CCDF of a BPH distribution that is stochastically smaller than $\bar{F}(x)$.

Even though we are mainly interested in stochastic order results in the BPH setup, the results are actually meaningful (and possibly useful) for the domain of the original Bernstein approximation over the interval $[0, 1]$. We start with the classic setup over $[0, 1]$, and switch to the exponential domain and prove Theorem 2 after.

Using notation in accordance with Section 2.1, we assume $G(y)$ to be a continuous, increasing function on $[0, 1]$ with $G(0) = 0$, $G(1) = 1$.

We denote

$$G_{+\epsilon}(y) = \min(G(y) + \epsilon, 1) = \begin{cases} G(y) + \epsilon & y < \hat{y}, \\ 1 & y \geq \hat{y}, \end{cases} \tag{20}$$

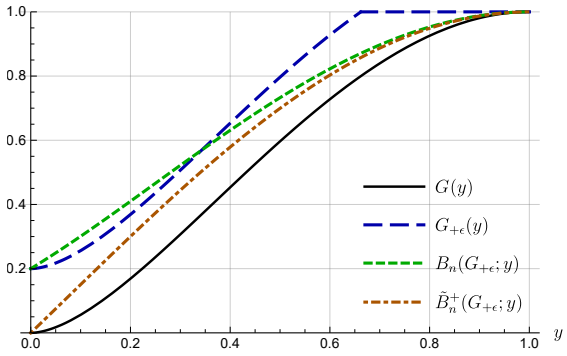

Figure 3: Example approximations for $G(y) = y^2(1 - 2\log(y))$, $n = 4$, $\epsilon = 0.2$ and $\hat{y} = 0.66$.

with $\hat{y} := \min\{y : G(y) \geq 1 - \epsilon\}$.

We also define the operator

$$\tilde{B}_n^+(G_{+\epsilon}; y) := \sum_{i=1}^{n} G_{+\epsilon}\left(\frac{i}{n}\right)\binom{n}{i} y^i (1-y)^{n-i}. \tag{21}$$

We note that $\tilde{B}_n^+(G_{+\epsilon}; y) \neq B_n(G_{+\epsilon}; y)$ as the $i = 0$ term is missing from the sum in (21); accordingly,

$$\tilde{B}_n^+(G_{+\epsilon}; y) = B_n(G_{+\epsilon}; y) - \epsilon(1-y)^n. \tag{22}$$

Example. *Fig. 3 shows the various operators applied to the function*

$$G(y) = y^2(1 - 2\log(y))$$

using parameters $n = 4$ and $\epsilon = 0.2$. (This function is actually the CCDF of the Erlang(2) distribution with mean 1 reverted back to $[0, 1]$ domain from exponential domain.)

The properties of $G(y)$, $G_{+\epsilon}(y)$, $B_n(G_{+\epsilon}; y)$, and $\tilde{B}_n^+(G_{+\epsilon}; y)$ are as follows:

- $G(y)$ is increasing with $G(0) = 0, G(1) = 1$;
- $G_{+\epsilon}(y)$ is increasing with $G_{+\epsilon}(0) = \epsilon, G_{+\epsilon}(1) = 1$, and this function is constant 1 on interval $[\hat{y}, 1]$;
- $B_n(G_{+\epsilon}; y)$ is the Bernstein approximation of $G_{+\epsilon}$, so it is increasing, $B_n(G_{+\epsilon}; 0) = \epsilon$, $B_n(G_{+\epsilon}; 1) = 1$, but this function is not constant over any interval;
- $\tilde{B}_n^+(G_{+\epsilon}; y)$ is increasing, $\tilde{B}_n^+(G_{+\epsilon}; 0) = 0$, $\tilde{B}_n^+(G_{+\epsilon}; 1) = 1$, and it is strictly smaller than $B_n(G_{+\epsilon}; y)$ over $[0, 1)$.

Theorem 3. *Assuming that*

- *$G(y)$ is a continuous, increasing function on $[0, 1]$ with $G(0) = 0$, $G(1) = 1$, and*
- *$G(y) \leq Cy$ for some $C < \infty$, and*
- *there exists an $n_d \geq 1$ such that $\frac{\mathrm{d}^{n_d}}{\mathrm{d}y^{n_d}} G(y)\Big|_{y=1} \neq 0$*

then for any $0 < \epsilon < 0.5$ there exists an $\hat{n}$ such that for all $n > \hat{n}$,

$$G(y) \leq \tilde{B}_n^+(G_{+\epsilon}; y) \leq G(y) + 2\epsilon \qquad \forall y \in [0, 1]\,. \tag{23}$$

The proof of Theorem 3 is provided in Appendix A.

Remark 1. *The conditions of Theorem 3 are also necessary in the following sense:*

- *Around 0:* $\tilde{B}_n^+(G_{+\epsilon}; y)$ *is a polynomial for any choice of* n *and* ϵ*, so its derivative at 0 is finite, and it can dominate functions with* $G(y) \leq Cy$ *for some finite* C*. We also note that the condition* $G(y) \leq Cy$ *follows from* $G'(0) < \infty$*, which is a practical way to check* $G(y) \leq Cy$ *for most functions.*
- *Around 1: If* $G(y)$ *was such that* $\frac{d^n}{dy^n} G(y)|_{y \nearrow 1} = 0$ *for all* $n > 0$ *then* $\tilde{B}_n^+(G_{+\epsilon}; y) \geq G(y)$ *could not hold around* $y = 1$*, since* $\tilde{B}_n^+(G_{+\epsilon}; y)$ *is a polynomial with some positive derivative at* $y \nearrow 1$*.*

There is an obvious counterpart of Theorem 3 for a lower bound. Introduce the notations

$$G_{-\epsilon}(y) = \max(G(y) - \epsilon, 0) = \begin{cases} 0 & y \leq \check{y}, \\ G(y) - \epsilon & y > \check{y}, \end{cases} \tag{24}$$

with $\check{y} := \max\{y : G(y) \leq \epsilon\}$ and the operator

$$\tilde{B}_n^-(G_{-\epsilon}; y) = \sum_{i=0}^{n-1} G_{-\epsilon}\left(\frac{i}{n}\right) \binom{n}{i} y^i (1-y)^{n-i} + y^n. \tag{25}$$

We note that $\epsilon < 0.5$ ensures $\check{y} \leq \hat{y}$.

Theorem 4. *Assuming that*

- $G(y)$ *is a continuous, increasing function on* $[0, 1]$ *with* $G(0) = 0$, $G(1) = 1$*, and*
- $1 - G(1 - y) < Cy$ *for some* $C < \infty$*, and*
- *there exists a finite minimal* n_d *such that* $\frac{\mathrm{d}^{n_d}}{\mathrm{d}y^{n_d}} G(y)\Big|_{y=0} \neq 0$

then for any $0 < \epsilon < 0.5$ *there exists an* $\hat{n}$ *such that for all* $n > \hat{n}$*,*

$$G(y) \geq \tilde{B}_n^-(G_{-\epsilon}; y) \geq G(y) - 2\epsilon \qquad \forall y \in [0, 1]\,. \tag{26}$$

Remark 2. *Similarly to Remark 1,* $1 - G(1 - y) < Cy$ *follows directly from* $G'(1) < \infty$*.*

Proof. (Theorem 4) To prove (26), we use the operator

$$J(G; y) = 1 - G(1 - y). \tag{27}$$

J is an involution (that is, $J(JG) = G$), and it preserves the properties $G(0) = 0, G(1) = 1$, monotonicity and continuity.

To check how J transforms the approximation (25) into (21), we first expand $H(y) = 1 - G(1 - y)$ in two steps:

$$U(z) = 1 - G(z), \qquad H(y) = U(1 - y),$$

and write

$$\begin{aligned} 1 - G_{-\epsilon}\left(\frac{i}{n}\right) &= 1 - \max\left(G\left(\frac{i}{n}\right) - \epsilon, 0\right) = \min\left(1 - G\left(\frac{i}{n}\right) + \epsilon, 1\right) \\ &= \min\left(U\left(\frac{i}{n}\right) + \epsilon, 1\right) = \min\left(H\left(1 - \frac{i}{n}\right) + \epsilon, 1\right) = H_{+\epsilon}\left(1 - \frac{i}{n}\right). \end{aligned} \tag{28}$$

Next we rewrite (25) as

$$\tilde{B}_n^-(G_{-\epsilon}; z) = \sum_{i=0}^{n-1} G_{-\epsilon}\left(\frac{i}{n}\right) \binom{n}{i} z^i (1-z)^{n-i} + z^n = \sum_{i=0}^{n} G_{-\epsilon}\left(\frac{i}{n}\right) \binom{n}{i} z^i (1-z)^{n-i} + \epsilon z^n.$$

Then, using $\sum_{i=0}^{n}\binom{n}{i}z^i(1-z)^{n-i}=1$, we obtain

$$\begin{aligned}
1-\tilde{B}_n^-(G_{-\epsilon};z) &= \sum_{i=0}^{n}\left(1-G_{-\epsilon}\left(\frac{i}{n}\right)\right)\binom{n}{i}z^i(1-z)^{n-i}-\epsilon z^n \\
&\overset{(28)}{=} \sum_{i=0}^{n}H_{+\epsilon}\left(1-\frac{i}{n}\right)\binom{n}{i}z^i(1-z)^{n-i}-\epsilon z^n \\
&\overset{z=1-y}{=} \sum_{i=0}^{n}H_{+\epsilon}\left(1-\frac{i}{n}\right)\binom{n}{i}(1-y)^i y^{n-i}-\epsilon(1-y)^n \\
&\overset{j=n-i}{=} \sum_{j=0}^{n}H_{+\epsilon}\left(\frac{j}{n}\right)\binom{n}{j}y^j(1-y)^{n-j}-\epsilon(1-y)^n \\
&= \sum_{j=1}^{n}H_{+\epsilon}\left(\frac{j}{n}\right)\binom{n}{j}y^j(1-y)^{n-j}=\tilde{B}_n^+(H_{+\epsilon};y).
\end{aligned}$$

Checking how J transforms the conditions and conclusion of Theorem 4 into the conditions and conclusion of Theorem 3 follows the same pattern. □

The symmetry used in the proof of Theorem 4 is due to the fact that the theorem is presented with Bernstein polynomials in the $[0,1]$ domain. Now we switch to the BE in the $[0,\infty)$ domain, providing the proof for Theorem 2.

Proof. (Theorem 2) For both parts, we apply the transformation $\bar{F}(x)=G(e^{-x})$, that is,

$$y=e^{-x} \qquad \Longleftrightarrow \qquad x=-\log(y)\,.$$

This transformation maps the interval $x\in[0,\infty)$ to $y\in[1,0]$; we examine how it transforms the functions and conditions in part (a) and part (b) of Theorem 2 respectively.

(a) The function $\hat{\bar{F}}_{+\epsilon,n}(x)$ in (18) is mapped to $\tilde{B}_n^+(G_{+\epsilon};y)$ (see (21)).

The conditions of part (a) of Theorem 2 are also mapped directly to the conditions of Theorem 3:

- the condition $G(y)\le Cy$ for some $C<\infty$ is equivalent to $\bar{F}(x)\le Ce^{-x}$ with the same C;
- since $\frac{\mathrm{d}}{\mathrm{d}x}e^{-x}\Big|_{x=0}=1$, the mapping $x\to e^{-x}$ preserves the derivative at $x=0$, so

$$\frac{\mathrm{d}^{n_d}}{\mathrm{d}x^{n_d}}\bar{F}(x)\Big|_{x=0}\neq 0 \qquad \Longleftrightarrow \qquad \frac{\mathrm{d}^{n_d}}{\mathrm{d}y^{n_d}}G(y)\Big|_{y=1}\neq 0.$$

The stochastic ordering conclusion also directly follows, since transforming the variable domain does not affect inequalities for the values of the functions involved.

(b) Applying the same transformation $y=e^{-x}$ and $\bar{F}(x)=G(e^{-x})$, we now get that the function $\hat{\bar{F}}_{-\epsilon,n}(x)$ in (19) is mapped to $\tilde{B}_n^-(G_{-\epsilon};y)$ (see (25)).

The conditions of part (b) of Theorem 2 are also mapped directly to the conditions of Theorem 4:

- $\frac{\mathrm{d}}{\mathrm{d}x}\bar{F}(x)|_{x=0}\le C$ is equivalent to $G'(1)\le C$, which follows directly from $1-G(1-y)<Cy$;
- $\bar{F}(x)>ce^{-ax}$ for some $0<c<\infty$ and $a<\infty$ is equivalent to $G(y)>cy^a$, which means that in the Taylor series expansion of G around $y=0$, no more than the first $\lceil a-1\rceil$ derivatives can be 0.

The stochastic ordering conclusion also directly follows, similarly to part (a).

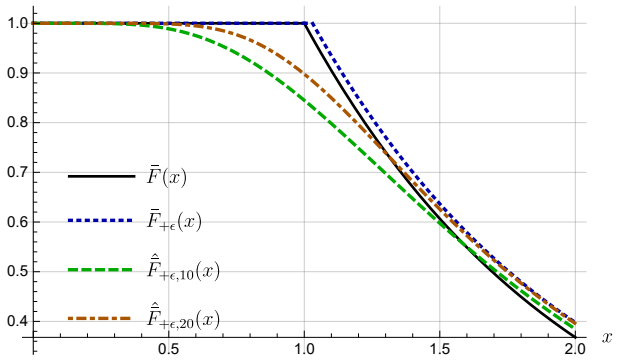

Figure 4: BPH approximations of a shifted exponential distribution: there does not exist a larger BPH distribution.

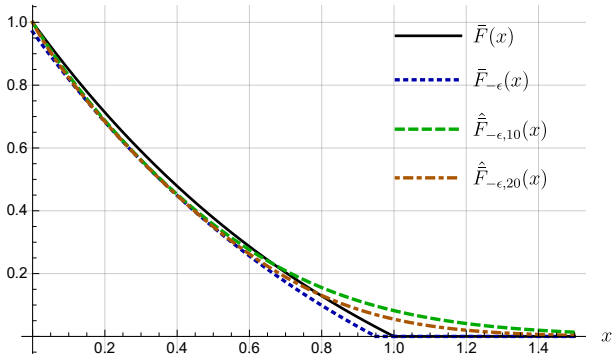

Figure 5: BPH approximations of a truncated exponential distribution: there does not exist a smaller BPH distribution.

□

The assumptions on the derivatives in Theorem 2 are necessary, since for any $\bar{F}$ BPH CCDF, there exists a finite n_d for which $\frac{d^{n_d}}{dx^{n_d}}\bar{F}(x)|_{x=0} \neq 0$, so no BPH CCDF can dominate a function whose every derivatives at $x = 0$ are 0, such as $1 - e^{-\frac{1}{x^2}}$. Similarly, every $\bar{F}$ BPH CCDF has a finite derivative at 0, so no BPH CCDF can be dominated by a function like $\hat{F}(x) = 1 - \sqrt{x}$ that has an infinite derivative at 0.

The CCDFs of bounded random variables also violate the conditions of Theorem 2 and cannot be dominated either; we state this as a separate corollary.

Corollary 1. *Consider a CCDF $\bar{F}(x)$. Then*

$$\exists x > 0,\ \bar{F}(x) = 1 \implies \nexists\ BPH(a),\ \forall x \geq 0,\ \bar{F}(x) \leq \bar{F}_a(x)$$

and

$$\exists x > 0,\ \bar{F}(x) = 0 \implies \nexists\ BPH(a),\ \forall x \geq 0,\ \bar{F}(x) \geq \bar{F}_a(x),$$

Relevant examples of bounded random variables include shifted and truncated exponentials:

$$\bar{F}_{se}(x) = \begin{cases} 1 & x \leq 1, \\ e^{-(x-1)} & x > 1, \end{cases} \quad \bar{F}_{te}(x) = \begin{cases} \dfrac{e^{1-x} - 1}{e - 1} & x \leq 1, \\ 0 & x > 1. \end{cases} \tag{29}$$

Figures 4 and 5 illustrate the issues when trying to find stochastically larger approximations of shifted exponentials and stochastically smaller approximations of truncated exponentials, respectively.

5. Numerical Investigations

In this section we use several distributions to numerically investigate BPH approximations obtained based on the CCDF (which, as shown in Proposition 1, is equivalent to using the CDF). Let us list these distributions and discuss their characteristics.

5.1. Considered Distributions

The order k Erlang distribution with mean equal to 1 is with CCDF

$$\bar{F}(x) = \sum_{i=0}^{k-1} e^{-kx}(kx)^i/i!\ , \tag{30}$$

and we will refer to it as Erlang(k).

While an Erlang distribution is itself a PH distribution, it provides a straightforward way to analyze several crucial characteristics of BPH approximations for the following reasons. The order k can be used to control both the behavior at zero and as x tends to infinity. The larger k, the longer the CCDF remains close to one (the first $k-1$ derivatives of the CCDF are zero at $x=0$), and the more phases we need to construct a stochastically larger BPH. At $x \to \infty$, the larger k, the faster the CCDF decays (at rate $e^{-kx}x^{k-1}$) and the more phases we need to construct a stochastically smaller BPH. Moreover, the Erlang distribution is known to have the smallest possible squared coefficient of variation (SCV) among PH distributions of a given order k independent of the mean [17], namely $1/k$. This allows us to study easily also the impact of the SCV on the goodness of the approximation.

Erlang distributions satisfy the conditions of both part (a) and (b) of Theorem 2 with any order and any mean, which guarantees that there exists an n such that $\hat{\bar{F}}_{+\epsilon,n}(x)$, defined in (18), ($\hat{\bar{F}}_{-\epsilon,n}(x)$ defined in (19)) defines a random variable that is stochastically larger (smaller) than the one defined by $\bar{F}(x)$.

The Weibull distribution with scale λ and shape k, which we will denote by Weibull(λ, k), has CCDF

$$\bar{F}(x) = e^{-(x/\lambda)^k},$$

and we have to distinguish two cases:

- with $k > 1$ the conditions of part (a) of Theorem 2 are satisfied (i.e., the existence of a stochastically larger BPH is guaranteed) but not those of part (b) because $\bar{F}(x)$ decays at $x \to \infty$ faster than exponential (i.e., there does not exist a stochastically smaller BPH);
- with $k < 1$ the conditions of part (a) are not satisfied because $\bar{F}(x)$ decays at $x \to \infty$ slower than exponential, and neither those of part (b) are met because the derivative at $x = 0$ is infinite.

(With $k = 1$ the Weibull distribution is identical to the exponential distribution.)

For what concerns the (shifted) Pareto distribution with CCDF

$$\bar{F}(x) = \frac{1}{(x+1)^\alpha} \text{ with } \alpha > 0,$$

which we will denote by Pareto(α), conditions of part (a) of Theorem 2 are not satisfied because the decay of the CCDF is slower than exponential as $x \to \infty$ (i.e., there does not exist a larger BPH distribution) but those of part (b) are met (i.e., there exist smaller BPH distributions).

Furthermore, we will use also uniform distributions specifying their support.

5.2. Distribution Approximations

We approximate $\bar{F}(x)$ itself and its increased and decreased variants, i.e., $\bar{F}_{+\epsilon}(x)$ and $\bar{F}_{-\epsilon}(x)$ as given in (17) as well, in order to obtain stochastically larger and smaller BPH distributions. For a given n and ϵ, checking whether $\hat{\bar{F}}_{+\epsilon,n}(x) \geq \bar{F}(x)$ for every $x \geq 0$ ($\hat{\bar{F}}_{-\epsilon,n}(x) \leq \bar{F}(x)$ for every $x \geq 0$) is not straightforward. We analyzed the difference between $\bar{F}(x)$ and the approximations numerically over the main body of the functions. If $\min_x \hat{\bar{F}}_{+\epsilon,n}(x) - \bar{F}(x) = 0$ then $\hat{\bar{F}}_{+\epsilon,n}(x)$ defines a random variable that is stochastically larger than the one of $\bar{F}(x)$; vice versa, if $\max_x \hat{\bar{F}}_{-\epsilon,n}(x) - \bar{F}(x) = 0$ then $\hat{\bar{F}}_{-\epsilon,n}(x)$ defines a random variable that is stochastically smaller than the one of $\bar{F}(x)$. (Note that $\bar{F}(0) = \hat{\bar{F}}_{+\epsilon,n}(0) = \hat{\bar{F}}_{-\epsilon,n}(0) = 1$ is guaranteed by the approximation, which implies that $\min_x \hat{\bar{F}}_{+\epsilon,n}(x) - \bar{F}(x) \leq 0$ and $\max_x \hat{\bar{F}}_{-\epsilon,n}(x) - \bar{F}(x) \geq 0$.)

We start with Erlang distributions for which, as mentioned above, the existence of both smaller and larger BPH distributions is guaranteed.

Figure 6 shows the CCDFs and PDFs resulting from the approximation of the Erlang(2) CCDF with $n = 40$ and $\epsilon = 0.1$. Approximating $\bar{F}(x)$ itself via (3) gives a good approximation but does not guarantee stochastic order. Approximating $\bar{F}_{+\epsilon}(x)$ and $\bar{F}_{-\epsilon}(x)$ provides a larger and a smaller distribution, respectively, but the resulting CCDFs are far from the original due to the relatively large values of ϵ. For a CCDF $\bar{F}(x)$ we

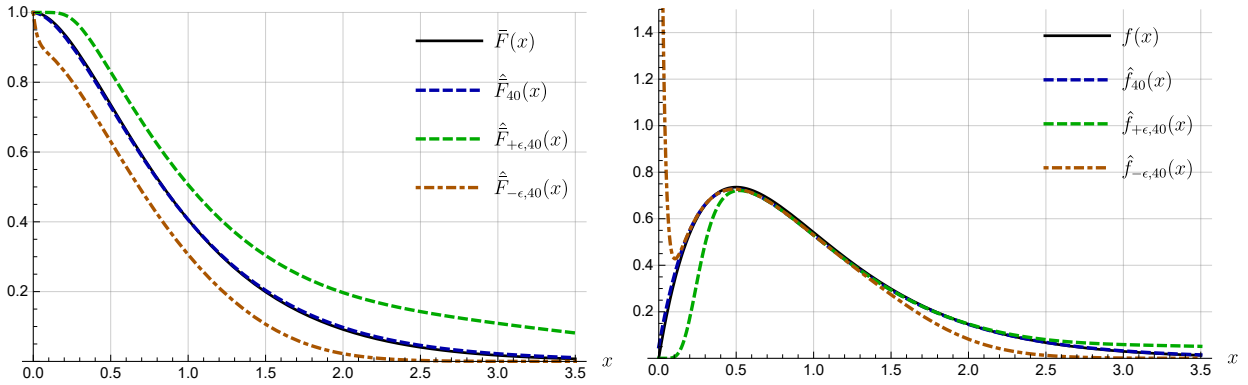

Figure 6: Approximating the Erlang(2) CCDF with $n = 40, \epsilon = 0.1$: on the left the resulting CCDFs, on the right the corresponding PDFs.

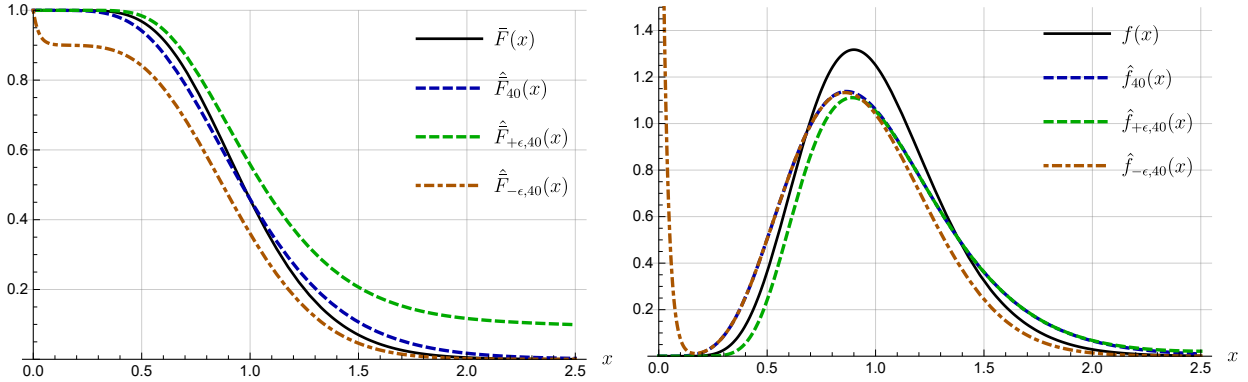

Figure 7: Approximating the Erlang(10) CCDF with $n = 40, \epsilon = 0.1$: on the left the resulting CCDFs, on the right the corresponding PDFs.

denote the corresponding PDF as $f(x) = -\bar{F}'(x)$. A particular consequence of using $\bar{F}_{-\epsilon}(x)$ can be observed for the PDF $\hat{f}_{-\epsilon,40}(x)$ at zero where we have $\hat{f}_{-\epsilon,40}(0) = 4.05$. This is due to the fact that the derivative of $\hat{\bar{F}}_{-\epsilon,40}(x)$ is negative at zero. The larger ϵ, the larger $\hat{f}_{-\epsilon,40}(0)$. The SCV in this case is relatively large, $1/2$, and hence, approximating $\hat{F}(x)$ provides a CCDF and a corresponding PDF that follows closely the original CCDF and PDF.

Figure 7 shows analogous experiments for Erlang(10). Similar to the Erlang(2) case, approximating $\bar{F}_{+\epsilon}(x)$ and $\bar{F}_{-\epsilon}(x)$ provides a larger and a smaller distribution, respectively, and we have a peak at zero in the PDF of the smaller distribution. The main difference with respect to Erlang(2) is the much lower SCV (1/10 for Erlang(10)). Accordingly, $\hat{f}_{40}(x)$ is unable to capture the "narrow" shape of $f(x)$. The rigid structure of the BPH distribution (fixed intensities and distributed initial probabilities, see Figure 2) is not appropriate to obtain low SCV with an order 40 BPH distribution.

In Figure 8 we show the difference between the original CCDF and the approximating CCDFs for the two experiments considered so far, i.e., $n = 40, \epsilon = 0.1$ with Erlang(2) and Erlang(10). For the approximating $\bar{F}_n(x)$, close to zero the resulting CCDF is smaller than the original, and later it becomes larger. According to our experience this is the typical situation when approximating a CCDF that has one or more derivatives equal to zero at zero.

As expected, for smaller values of ϵ, it can be necessary to increase the number of phases in order to obtain a stochastically larger (or smaller) approximation. In Figure 9, using Erlang(10) and $\epsilon = 0.02$, we show the difference between the approximating CCDFs and the original one with $n = 40$ and $n = 160$. With 40 phases, there is no stochastic order between the Erlang and the approximations. Indeed, both

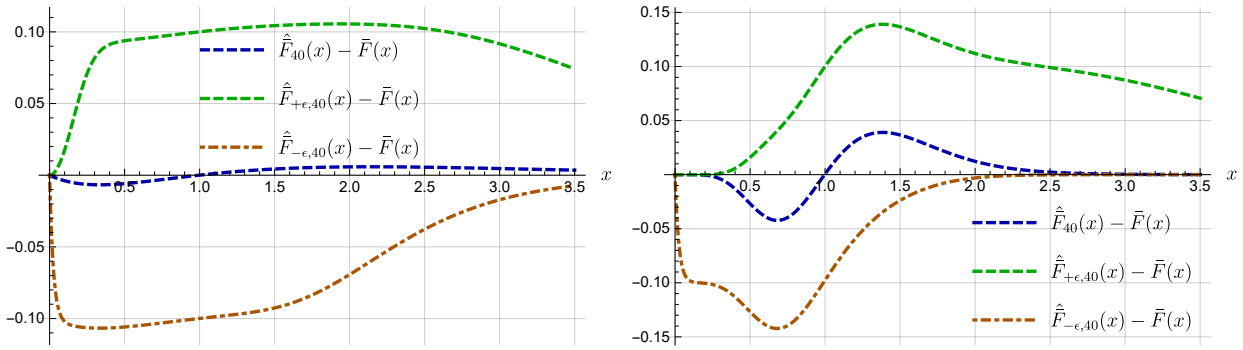

Figure 8: Difference in the CCDFs approximating the Erlang CCDF with $n = 40, \epsilon = 0.1$, Erlang(2) on the left and Erlang(10) on the right.

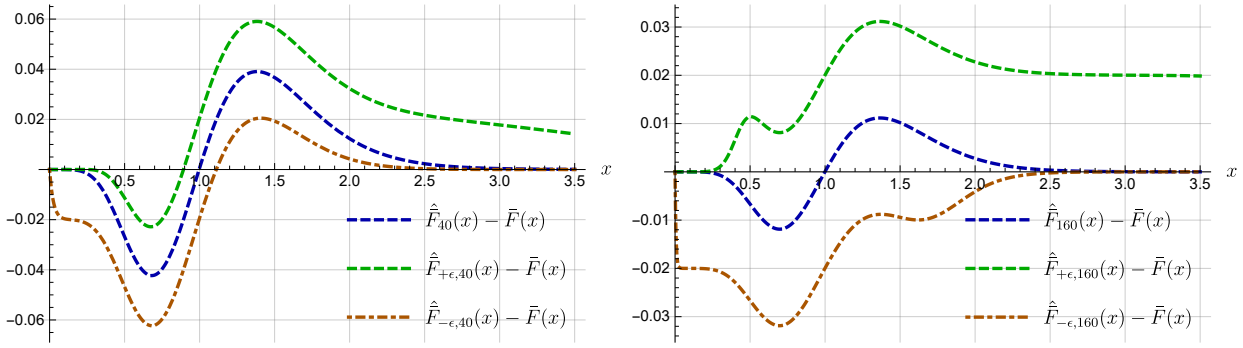

Figure 9: Difference in the CCDFs approximating the Erlang(10) CCDF with $\epsilon = 0.02$; with $n = 40$ on the left and with $n = 160$ on the right.

$\hat{\bar{F}}_{+\epsilon,40}(x) - \bar{F}(x)$ and $\hat{\bar{F}}_{-\epsilon,40}(x) - \bar{F}(x)$ cross the x axis. For $n = 160$, stochastic order is guaranteed.

We report also some results for several orders of the BE approximation: $n = 5, 10, 20, 40, 80$ and 160. We investigate the minimum and the maximum of $\hat{\bar{F}}_n(x) - \bar{F}(x)$, $\hat{\bar{F}}_{+\epsilon,n}(x) - \bar{F}(x)$ and $\hat{\bar{F}}_{-\epsilon,n}(x) - \bar{F}(x)$. The results are shown in Figure 10 using Erlang(2) and Erlang(10) with $\epsilon = 0.05$. The minimum and maximum of $\hat{\bar{F}}_n(x) - \bar{F}(x)$ show a symmetric behavior with respect to the x axis. For the stochastically larger and smaller approximations, $\min_x \hat{\bar{F}}_{+\epsilon,n}(x) - \bar{F}(x)$ is symmetric to $\max_x \hat{\bar{F}}_{-\epsilon,n}(x) - \bar{F}(x)$ and $\min_x \hat{\bar{F}}_{-\epsilon,n}(x) - \bar{F}(x)$ is symmetric to $\max_x \hat{\bar{F}}_{+\epsilon,n}(x) - \bar{F}(x)$. A larger (smaller) distribution is guaranteed once n is increased so that $\min_x \hat{\bar{F}}_{+\epsilon,n}(x) - \bar{F}(x)$ ($\max_x \hat{\bar{F}}_{-\epsilon,n}(x) - \bar{F}(x)$) reaches the x axis.

We turn our attention now to a distribution for which neither larger nor smaller BPH approximations exist, namely, the Weibull$(0.5, 0.5)$ distribution. Figure 11 shows the approximations of the CCDF itself for various values of n. Close to 0 the approximations are not able to follow the fast decay of the Weibull$(0.5, 0.5)$ CCDF (its derivative at 0 is infinite) while toward the tail the approximations decrease faster than the tail of the Weibull$(0.5, 0.5)$ distribution because it is heavier than exponential. Increasing n, as expected, results in approximations that are closer to the Weibull$(0.5, 0.5)$ CCDF both around 0 and toward the tail.

BPH distributions that are larger than the Weibull$(0.5, 0.5)$ distribution do not exists because of the tail behavior of the Weibull$(0.5, 0.5)$ distribution. However, it is possible to obtain BPH distributions whose CCDF is larger than that of the Weibull$(0.5, 0.5)$ up to a predefined limit by approximating the increased version of the Weibull$(0.5, 0.5)$ CCDF, i.e., $\bar{F}_{+\epsilon}(x)$. This is illustrated in Figure 12. The larger n, the longer $\hat{\bar{F}}_{+\epsilon,n}$ remains larger than $\bar{F}(x)$. On the left side of the figure, $\hat{\bar{F}}_{+\epsilon,10}(x)$ is larger than $\bar{F}(x)$ up to $x = 2.53$, $\hat{\bar{F}}_{+\epsilon,40}(x)$ up to $x = 4.22$ and $\hat{\bar{F}}_{+\epsilon,160}(x)$ up to $x = 5.96$. Similarly, the larger ϵ (i.e., the more we increase the CCDF during the approximation), the longer $\hat{\bar{F}}_{+\epsilon,n}$ remains larger than $\bar{F}(x)$. On the right side of

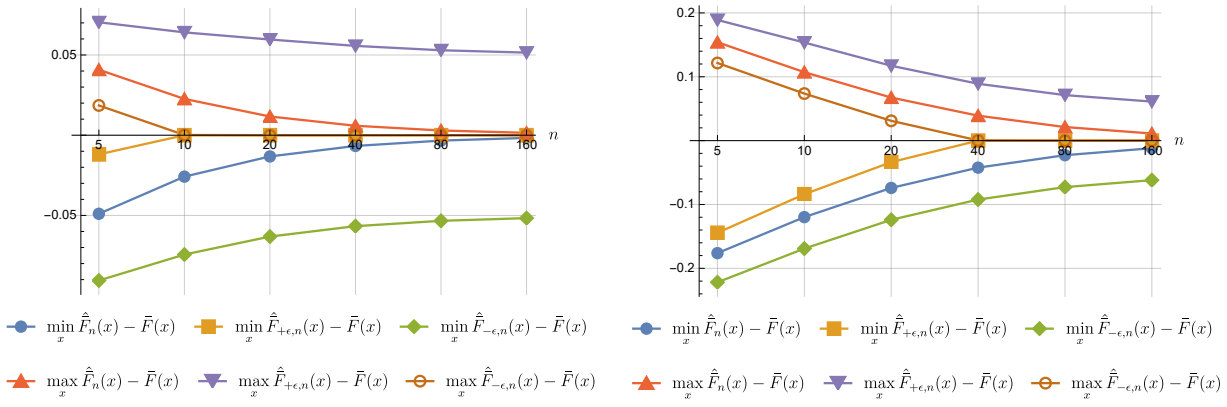

Figure 10: Minimum and the maximum of $\hat{\bar{F}}_n(x) - \bar{F}(x)$, $\hat{\bar{F}}_{+\epsilon,n}(x) - \bar{F}(x)$ and $\hat{\bar{F}}_{-\epsilon,n}(x) - \bar{F}(x)$ for various values of n with $\epsilon = 0.05$ and Erlang(2) (left) and Erlang(10) (right).

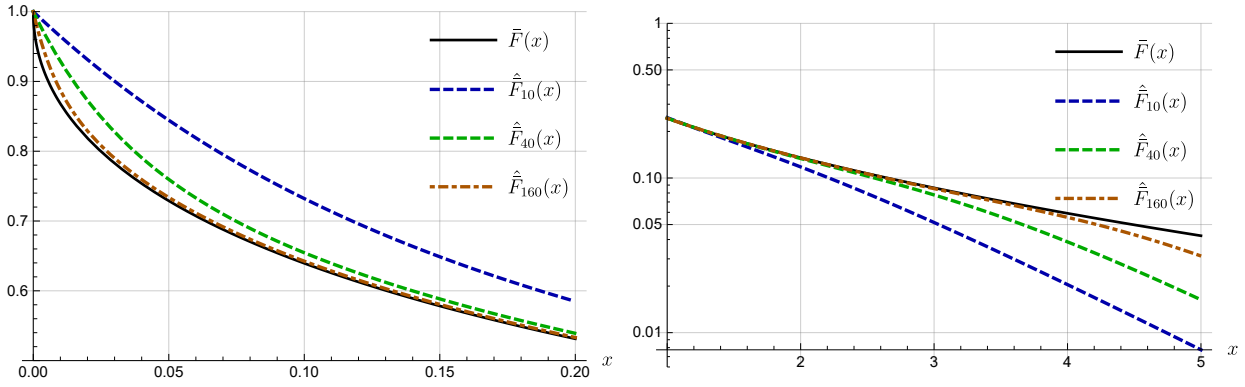

Figure 11: Approximations of the Weibull$(0.5, 0.5)$ CCDF with different values of n: the initial part of CCDFs on the left and their tail on the right.

the figure, $\hat{\bar{F}}_{+0.01,40}(x)$ is larger than $\bar{F}(x)$ up to $x = 3$, $\hat{\bar{F}}_{+0.05,40}(x)$ up to $x = 4.21$ and $\hat{\bar{F}}_{+0.1,40}(x)$ up to $x = 4.9$.

BPH distributions that are smaller than the Weibull$(0.5, 0.5)$ distribution do not exists because of the behavior of the Weibull$(0.5, 0.5)$ CCDF around 0. Nevertheless, by approximating the decreased version of the CCDF, $\bar{F}_{-\epsilon}$, it is possible to obtain BPH distributions with CCDF that is smaller than Weibull$(0.5, 0.5)$ CCDF for any x larger than a predefined value. Also in this case, increasing either n or ϵ improves the situation in the sense that the threshold value of x after which the CCDF of the BPH approximation is smaller becomes smaller. This is illustrated in Figure 13. On the left-hand side, varying n, we have $\hat{\bar{F}}_{-0.05,10}(x) < \bar{F}(x)$ for any x larger than 0.23, $\hat{\bar{F}}_{-0.05,40}(x) < \bar{F}(x)$ for $x > 0.037$ and $\hat{\bar{F}}_{-0.05,160}(x) < \bar{F}(x)$ for $x > 0.005$. On the right-hand side, varying ϵ, we have $\hat{\bar{F}}_{-0.01,40}(x) < \bar{F}(x)$ for any x larger than 0.15, $\hat{\bar{F}}_{-0.05,40}(x) < \bar{F}(x)$ for $x > 0.037$ and $\hat{\bar{F}}_{-0.1,40}(x) < \bar{F}(x)$ for $x > 0.02$.

While the Erlang(10) distribution, as mentioned already, is with low SCV, the Weibull$(0.5, 0.5)$ distribution has a high SCV, specifically, it is equal to 5. The high SCV is due to the heavy tail of the distribution, a feature that is notoriously hard to capture with general purpose PH fitting approaches (i.e., approaches that do not treat the tail of the distribution with special care). Indeed, the approximating BPH distributions are with an SCV that is far from that of the Weibull$(0.5, 0.5)$ even for large values of n. This is shown in Table 1. This weakness of the BPH approximations could be easily reduced to a large extent by combining the BPH approximation with the method proposed in [6] to fit heavy-tailed distributions. Such

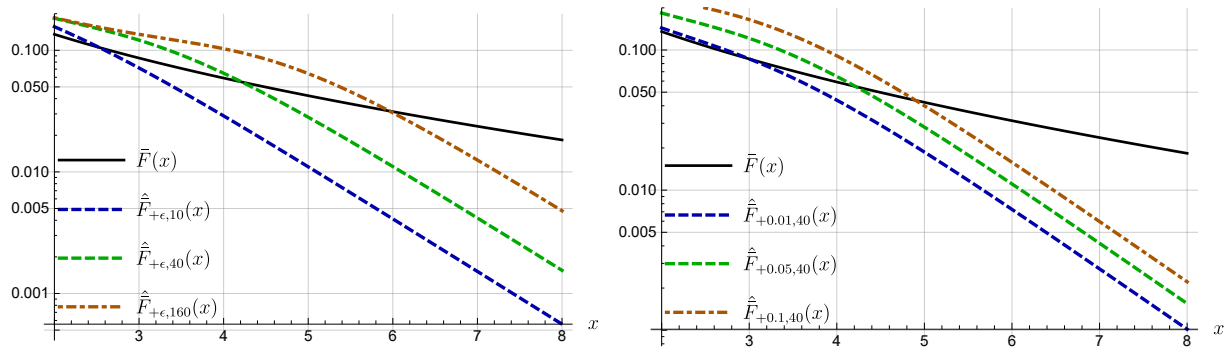

Figure 12: Approximations of the increased Weibull$(0.5, 0.5)$ CCDF with various values of n and $\epsilon = 0.05$ on the left and with various values of ϵ and $n = 40$ on the right.

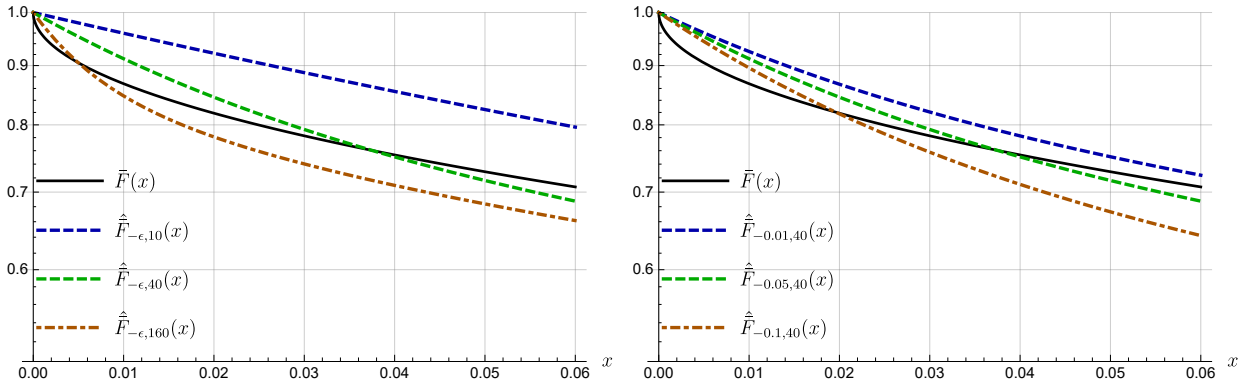

Figure 13: Approximations of the decreased Weibull$(0.5, 0.5)$ CCDF with various values of n and $\epsilon = 0.05$ on the left and with various values of ϵ and $n = 40$ on the right.

an approach, i.e., combining the method in [6] with a general purpose fitting procedure, was described and applied successfully in [8]. Note that, since the procedure in [6] is with very low computational complexity, its combination with BPH approximation would result in a method with execution time as low as that of pure BPH approximation.

n	10	50	100	500	1000	5000	10000	50000	100000
SCV	1.97	2.45	2.62	2.95	3.08	3.35	3.46	3.68	3.76

Table 1: SCV of various order of BPH approximations of the Weibull$(0.5, 0.5)$ distribution whose SCV is equal to 5.

5.3. Necessary Order to Obtain Smaller or Larger Distributions

Here we numerically investigate the minimal n as a function of ϵ, denoted by $n_+(\epsilon)$ and $n_-(\epsilon)$, that allows to obtain stochastically larger or smaller approximations, respectively.

Table 2 shows $n_+(\epsilon)$ and $n_-(\epsilon)$ for each choice of $\bar{F}$. When such $n_+(\epsilon)$ or $n_-(\epsilon)$ do not exist the table indicates $\nexists$ (this happens always in accordance with Theorem 2 as discussed in Section 5.1 distribution by distribution). It turns out that for all of the considered distributions, if $n_+(\epsilon)$ or $n_-(\epsilon)$ exist, then larger and smaller CCDFs are obtained for any $n \geq n_+(\epsilon)$ and $n \geq n_-(\epsilon)$, respectively, i.e.,

$$\hat{\bar{F}}_{+\epsilon,n}(x) \geq \bar{F}(x) \quad x \in [0, \infty), \ \forall n \geq n_+(\epsilon) \quad \text{and}$$
$$\hat{\bar{F}}_{-\epsilon,n}(x) \leq \bar{F}(x) \quad x \in [0, \infty), \ \forall n \geq n_-(\epsilon).$$

	Weibull(2, 2)		Weibull(0.5, 0.5)		Erlang(2)		Erlang(10)		Pareto(1)		Pareto(5)	
ϵ	$n_+(\epsilon)$	$n_-(\epsilon)$	$n_+(\epsilon)$	$n_-(\epsilon)$	$n_+(\epsilon)$	$n_-(\epsilon)$	$n_+(\epsilon)$	$n_-(\epsilon)$	$n_+(\epsilon)$	$n_-(\epsilon)$	$n_+(\epsilon)$	$n_-(\epsilon)$
0.1	7	$\nexists$	$\nexists$	$\nexists$	5	7	24	29	$\nexists$	1	$\nexists$	11
0.01	68	$\nexists$	$\nexists$	$\nexists$	27	28	192	180	$\nexists$	1	$\nexists$	74
0.001	687	$\nexists$	$\nexists$	$\nexists$	271	237	1976	1869	$\nexists$	1	$\nexists$	746

Table 2: Minimal order, $n_+(\epsilon)$ and $n_-(\epsilon)$, to obtain stochastically larger and smaller approximations, respectively, as function of ϵ.

Table 2 shows that apart from the exception (and possible corner case) Pareto(1), $n_+(\epsilon)$ and $n_-(\epsilon)$ typically increase linearly in $1/\epsilon$, with a constant factor depending on $\bar{F}$. For Pareto(1), $\hat{\bar{F}}_{-\epsilon,n}(x) \leq \bar{F}(x)$ holds already for $n = 1$, for any choice of ϵ examined.

5.4. Application of BPH Approximations in Queues

Finally, we apply some of the already studied approximations in M/G/1 queues. The service time distribution is Erlang(2) or Erlang(10) (see (30)). We use the same values of ϵ for approximations as in Table 2, namely 0.1, 0.01 and 0.001, and the minimal n that allows us to obtain stochastically larger and smaller approximate BPH service time distributions (this is also indicated in Table 2 as $n_+(\epsilon)$ and $n_-(\epsilon)$, respectively). The utilization of the queue is set to 0.7. The queue length distribution of the resulting M/BPH/1 queue can be calculated by the procedure provided in [11] that has linear complexity in the order n and hence allows us to use large values of n in the computation.

Since the BPH approximations guarantee stochastic order with respect to the original service time distribution, the CCDF of the queue length distribution and its upper and lower bounds for Erlang(2) are illustrated in Figure 14 on the left, while the difference between the bounds are plotted on the right. As expected, the bounds become tighter as $\epsilon \searrow 0$. The largest difference between the upper bound and the lower bound is 0.5202, 0.0885 and 0.01357, respectively, for $\epsilon = 0.1$, $\epsilon = 0.01$ and $\epsilon = 0.001$. For Erlang(10) the results are shown in Figure 15. In this case the largest difference between the upper bound and the lower bound is 0.6769, 0.09476 and 0.01388, respectively, for the same values of ϵ.

The figures indicate that for Erlang(2) and Erlang(10), the differences between the upper bound and the lower bound are rather similar in spite of the essentially different service time distribution (the SCV of the service time is 1/2 for Erlang(2) and 1/10 for Erlang(10)). We note that the figures hide an important aspect of the approximation which is highlighted in Table 2. Namely, in case of Erlang(10) the minimal order guaranteeing stochastic ordering is about seven times larger than in case of Erlang(2), however this increase does not lead to unfeasible computations due to the simplicity of the construction and the application of BPH approximations.

6. Comparison of PDF and CDF Approximations

BPH approximation was proposed in [11] for PDFs, while in this paper we study BPH approximations of CDFs. In this section we compare the PDF and CDF based BPH approximations.

The degree n BPH approximation of a given PDF $f(x)$ is

$$\hat{f}_n(x) = \frac{1}{c} \sum_{i=1}^{n} f\left(\log \frac{n}{i}\right) \binom{n}{i} e^{-ix}(1 - e^{-x})^{n-i}, \tag{31}$$

where $c = \sum_{i=1}^{n} \frac{f\left(\log \frac{n}{i}\right)}{i}$ is a normalizing constant ensuring that $\int_0^\infty \hat{f}_n(x)dx = 1$.

Figures 16, 17 and 18 plot the PDF and the CDF based approximations of the Uniform(0, 1), the Uniform(1, 2) and the Weibull(1, 2) distributions for $n = 50$. Visual inspection of these plots indicates approximations with similar precision. For a better numerical comparison we evaluated the error measures

$$\mathcal{D}_1^{\text{Pdf}} = \int_{x=0}^{\infty} \left|\hat{f}_n(x) - f(x)\right| dx, \quad \text{and} \quad \mathcal{D}_2^{\text{Pdf}} = \int_{x=0}^{\infty} \left(\hat{f}_n(x) - f(x)\right)^2 dx$$

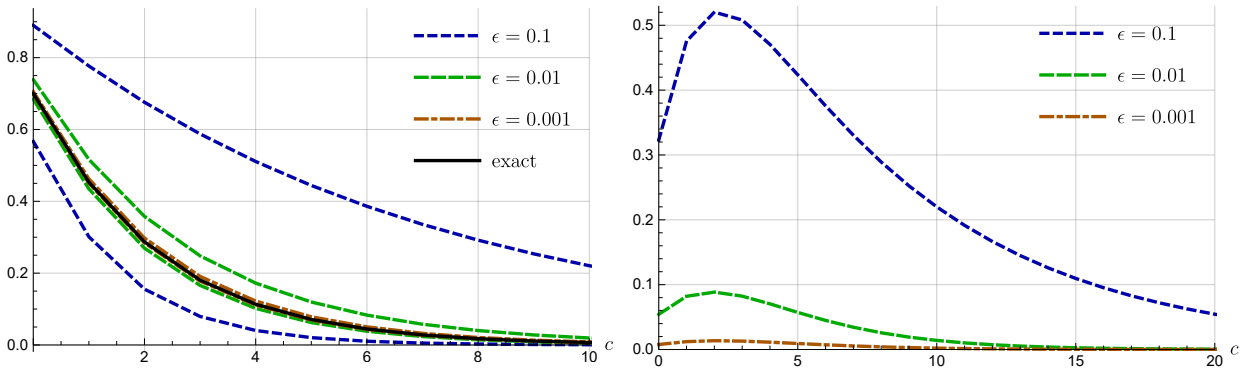

Figure 14: Bounds on the probability of more than c jobs present in the queue in case of Erlang(2) service time distribution (left) and the difference between the upper bound and the lower bound (right); obtained by larger and smaller BPH approximations with minimal order guaranteeing stochastic order for various values of ϵ. The plots are valid at integer values of c, and the discrete points are connected for better visibility.

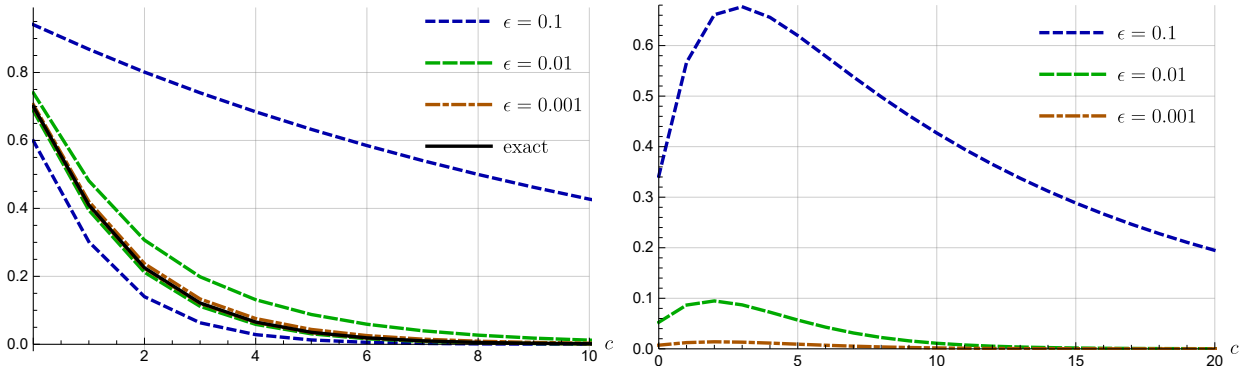

Figure 15: The same results as in Figure 14 with Erlang(10).

for the investigated distributions, where $\hat{f}_n(x)$ stands for the PDF of the BPH approximation obtained either based on the PDF or the CDF. The following table summarizes the results for order 25 and 50.

Distance	$\mathcal{D}_1^{\text{Pdf}}$				$\mathcal{D}_2^{\text{Pdf}}$			
Order of approx.	25		50		25		50	
BPH approx.	PDF based	CDF based	PDF based	CDF based	PDF based	CDF based	PDF based	CDF based
Uniform(0,1)	**0.2061**	0.2088	**0.15035**	0.15063	**0.06004**	0.06013	0.04327	**0.04302**
Uniform(1,2)	0.5928	**0.57139**	0.46055	**0.4058**	**0.1843**	0.1851	0.1354	**0.1253**
Weibull(1, 2)	0.1055	**0.0915**	0.05535	**0.04837**	0.00463	**0.00373**	0.001324	**0.001043**

The CDF based approximation is better in more cases but the error measures are similar. Without including the numerical values, we report that in case of $n = 100$ the CDF based approximation resulted in lower $\mathcal{D}_1^{\text{Pdf}}$ and $\mathcal{D}_2^{\text{Pdf}}$ measures for all the three distributions. In contrast, with $n = 10$ the CDF based approximation resulted, surprisingly, in lower $\mathcal{D}_1^{\text{Pdf}}$ and $\mathcal{D}_2^{\text{Pdf}}$ measures for the Uniform$(0, 1)$ and the Weibull$(1, 2)$ distributions, but not for Uniform$(1, 2)$.

Based on these numerical experiences, we can conclude that the PDF and CDF based BPH approximations have rather similar quality. An advantage of the CDF based BPH approximation is that it does not require the application of a normalization constant (denoted by c in (31)) in order to obtain a proper distribution. This constant, in particular cases with low values of n, can also introduce some sort of distortion of the approximation because its effect can be seen as approximating $f(x)/c$ instead of $f(x)$ itself.

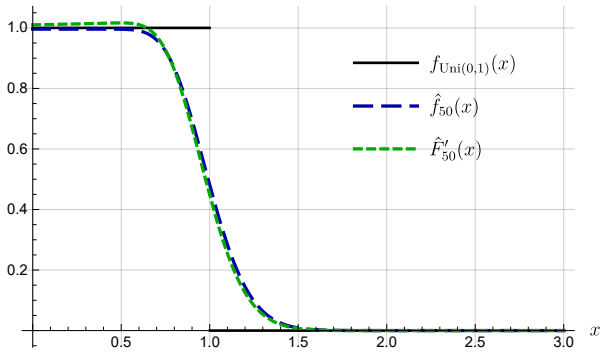

Figure 16: PDF and CDF based BPH approximation of the Uniform$(0, 1)$ distribution with $n = 50$.

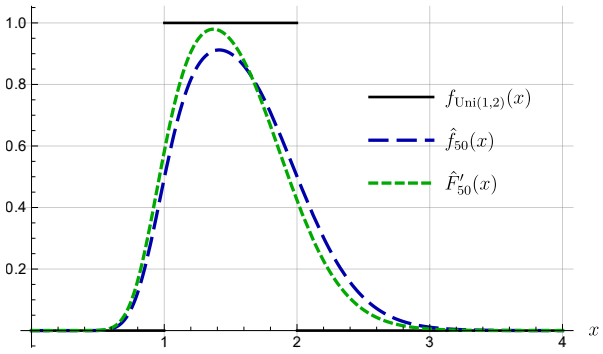

Figure 17: PDF and CDF based BPH approximation of the Uniform$(1, 2)$ distribution with $n = 50$.

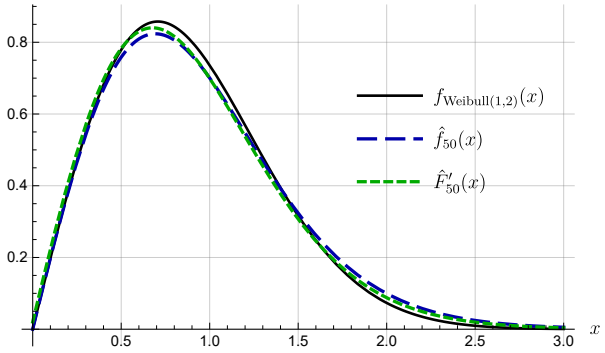

Figure 18: PDF and CDF based BPH approximation of the Weibull$(1, 2)$ distribution with $n = 50$.

7. Scaling Bernstein Phase Type Distributions

To transform Bernstein polynomials to Bernstein exponentials, we applied the change of variable $y = e^{-x}$. Applying a scaled version of this change of variable, namely, $y = e^{-\lambda x}$ (i.e., $x = -\log(y)/\lambda$), we also obtain a mapping from $[0, 1]$ onto $[0, \infty)$ but with an additional free parameter. This parameter, λ, referred to as scaling parameter, can be optimized in order to get better approximations.

The degree n scaled BE approximation of a given CDF $F(x)$ with scaling parameter $\lambda \in (0, \infty)$ is

$$\hat{F}_{n,\lambda}(x) = \sum_{i=0}^{n} F\left(\frac{\log \frac{n}{i}}{\lambda}\right) \cdot \underbrace{\binom{n}{i} e^{-i\lambda x}(1 - e^{-\lambda x})^{n-i}}_{\triangleq T_{n,i,\lambda}(x)}. \tag{32}$$

Similar to the BE approximation, the scaled BE approximation given in (32) also corresponds to a PH distribution as it is shown by the following theorem.

Theorem 5. *The CDF of PH(a, A) with*

$$a = \begin{pmatrix} a_1 & \dots & a_n \end{pmatrix}, \quad a_i = F\left(\frac{\log \frac{n}{i-1}}{\lambda}\right) - F\left(\frac{\log \frac{n}{i}}{\lambda}\right) \tag{33}$$

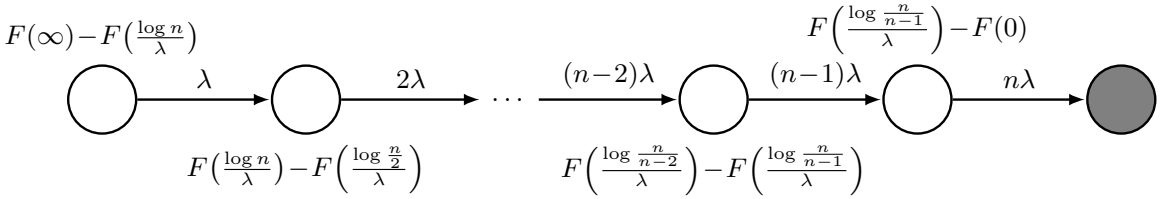

Figure 19: Bernstein PH approximation of a CDF $F(x)$ with scale parameter λ

and

$$A = \begin{pmatrix} -\lambda & \lambda & 0 & & ... & & \\ 0 & -2\lambda & 2\lambda & 0 & & ... & \\ & & & \ddots & & & \\ & ... & & 0 & -(n-2)\lambda & (n-2)\lambda & 0 \\ & & ... & & 0 & -(n-1)\lambda & (n-1)\lambda \\ & & & ... & & 0 & -n\lambda \end{pmatrix} \tag{34}$$

is equal to the approximation in (32)*. That is* $F_{a,A}(x) = \hat{F}_{n,\lambda}(x)$*.*

The graphical representation of PH(a, A) is shown in Figure 19. The proof of Theorem 5 based on (32) follows the same pattern as the proof of Theorem 1 based on (4), and we omit it here, but we spell out some properties of scaled BE approximation of CDFs.

Based on Figure 19, it is easy to verify that $\sum_{i=1}^{n} a_i = 1$ if $F(0) = 0$ and $F(\infty) = 1$ (which we assume in this paper). Moreover, since $F(x)$ is non-decreasing we have $a_i > 0, i = 1, ..., n$, that is, $a = (a_1 \ ... \ a_n)$ is a valid distribution over the phases. A direct consequence of Theorem 5 is the following corollary.

Corollary 2. *The approximation* $\hat{F}_{n,\lambda}(x)$ *of a CDF* $F(x)$ *is a valid CDF, i.e., it is non-decreasing,* $\hat{F}_{n,\lambda}(0) = 0$ *and* $\hat{F}_{n,\lambda}(\infty) = 1$*.*

To indicate the effect of the scaling parameter in BPH approximations of a given CDF $F(x)$, we investigate the behavior of some low order ($n = 5$) approximations of uniform and Weibull distributions with $\hat{F}_{n,\lambda}(x)$ according to (32).

Figure 20 presents the results for the Uniform$(0, 1)$ distribution. It indicates that different λ parameters result in approximating CDFs with different qualitative properties. The increase of the CDF in $(0, 1)$, i.e., $\hat{F}_{n,\lambda}(1) - \hat{F}_{n,\lambda}(0) = \hat{F}_{n,\lambda}(1)$ increases with λ. That is, for large λ values the probability that the approximating BPH random variable is larger than 1 is vanishing. The right side of Figure 20 indicates the price of increasing $\hat{F}_{n,\lambda}(1)$. When λ is large, the uniform density is significantly over-estimated by $\hat{F}'_{n,\lambda}(x)$ in the $(0, 0.5)$ interval.

Figure 21 presents the approximation of the CDF of the Uniform$(1, 2)$ distribution. In this case, high λ values also lead to a sharp increase of the CDF in the $(0, 1)$ interval, but they result in inaccurate approximations. The right side of Figure 21 shows the associated density functions. To better understand the behavior of the fitting curves, Table 3 reports the points where (32) samples the CDF to fit. That is, all sampling points with $\lambda = 4$ and $\lambda = 2$ are in the $(0, 1)$ interval where the CDF is zero. Consequently, $\hat{F}_{n,\lambda}(x) = T_{n,0,\lambda}(x) = (1 - e^{-\lambda x})^n$ in these cases. For a more appropriate fit, it is worth to set λ so that some sample points are present in the $(1, 2)$ interval as well, which is the case with $\lambda = 0.5$ and $\lambda = 1$. This simple example already indicates that the proper choice of λ depends on the distribution to fit.

Figures 22 and 23 present the approximations for the CDF of the Weibull$(1, 0.5)$ and the Weibull$(1, 2)$ distributions where the sample points are also according to Table 3. Similarly to the uniform distributions, sample points close to zero (i.e., when λ is large) result in an over estimation of the PDF and the CDF in the $(0.5, 1)$ interval. Lower λ values shift the sample points as well as the distribution to the right.

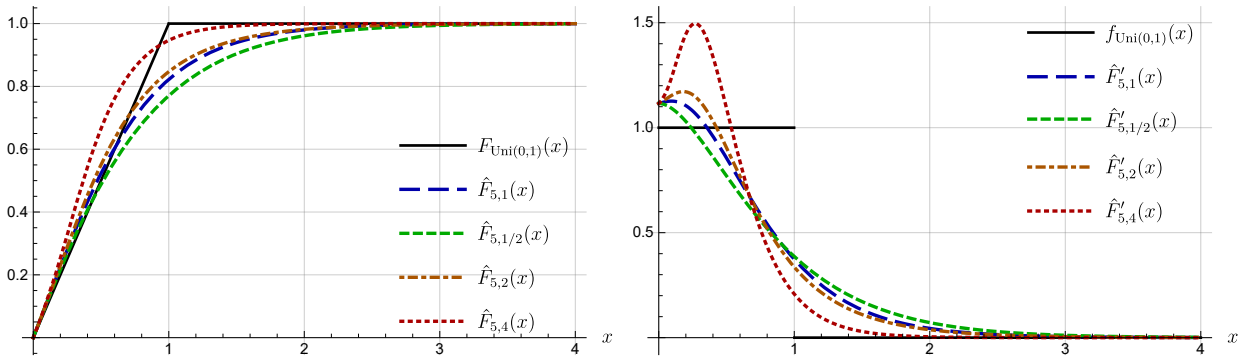

Figure 20: Scaled BPH approximation of the CDF of the Uniform$(0, 1)$ distribution with different λ parameters: approximate CDFs on the left and the corresponding approximate PDFs on the right.

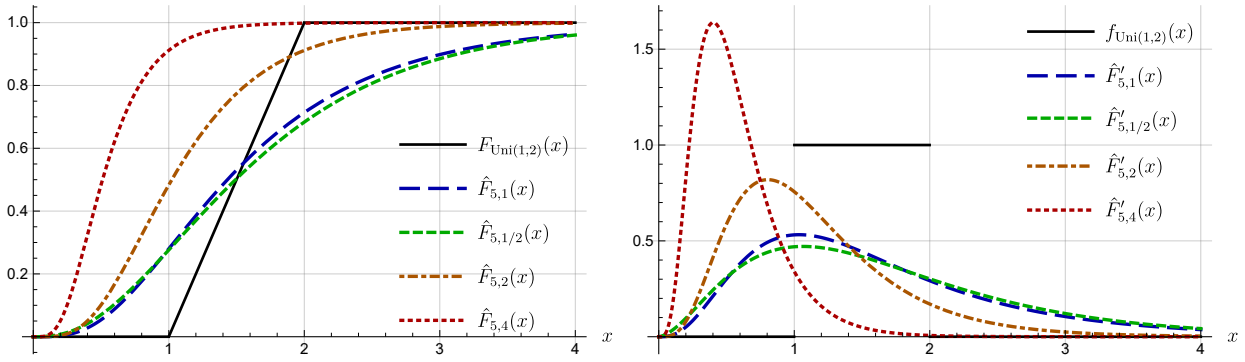

Figure 21: Scaled BPH approximation of the CDF of the Uniform$(1, 2)$ distribution with different λ parameters: approximate CDFs on the left and the corresponding approximate PDFs on the right.

λ	p_1	p_2	p_3	p_4
0.5	0.4462	1.0216	1.832	3.218
1	0.2231	0.5108	0.9162	1.609
2	0.1115	0.2554	0.4581	0.8047
4	0.05578	0.1277	0.2290	0.402

Table 3: The sampling points, $p_{n-i} = \frac{\log \frac{n}{i}}{\lambda}$, of the order $n = 5$ BPH approximation with different λ parameters.

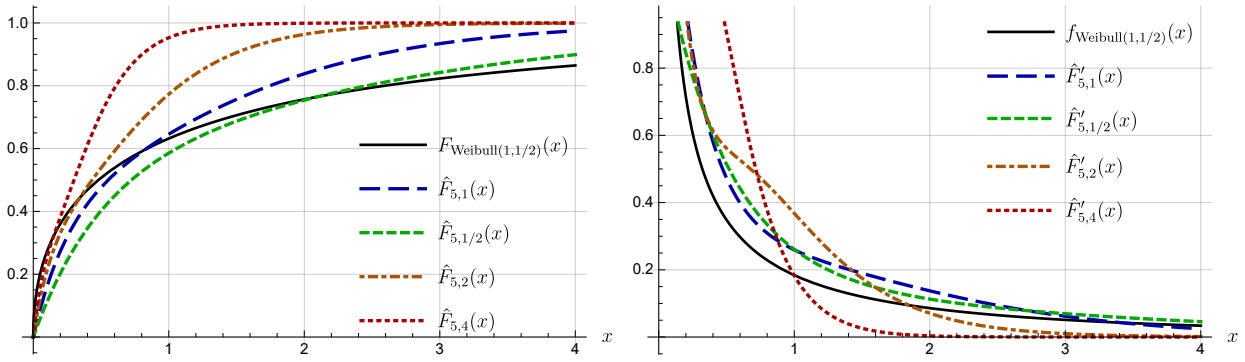

Figure 22: Scaled BPH approximation of the CDF of the Weibull$(1, 0.5)$ distribution with different λ parameters: approximate CDFs on the left and the corresponding approximate PDFs on the right.

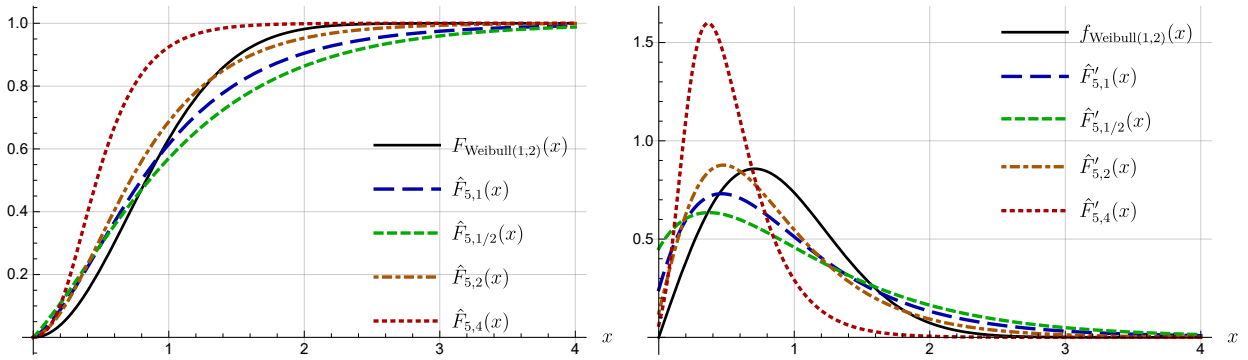

Figure 23: Scaled BPH approximation of the CDF of the Weibull$(1, 2)$ distribution with different λ parameters: approximate CDFs on the left and the corresponding approximate PDFs on the right.

7.1. Effect of the Scaling Parameter on the Accuracy of BPH Approximation

As it was illustrated before, different scaling parameters result in different BPH approximations, which raises the problem of finding the optimal scaling parameter for the scaled BPH approximation of a given distribution function.

The scaling parameter has two effects on the approximation (see (32)):

- it determines the points where the original function is sampled, $\left(\frac{\log \frac{n}{i}}{\lambda}\right)$ with $i = 1, ..., n-1$, and
- it affects the exponential functions applied in the approximation (i.e., the terms become $\binom{n}{i} e^{-i\lambda x}(1 - e^{-\lambda x})^{n-i}$).

The complex interdependencies of these two effects inhibit an analytical treatment of the optimal scaling parameter, and lead us to resort to numerical investigations.

We investigate the effect of the scaling parameter on some CDF based error measures of the scaled BPH approximation. Namely, we analyze the error measures

$$\mathcal{D}_1^{\text{Cdf}} = \int_{x=0}^{\infty} \left| \hat{F}_{n,\lambda}(x) - F(x) \right| dx, \quad \text{and} \quad \mathcal{D}_2^{\text{Cdf}} = \int_{x=0}^{\infty} \left(\hat{F}_{n,\lambda}(x) - F(x) \right)^2 dx$$

for some non-negative distributions with CDF given by $F(x)$.

The two error measures are plotted as function of the scaling parameters in Figures 24 to 27 in case of approximating the Uniform(0,1), the Uniform(1,2), the Weibull$(1, 0.5)$ and the Weibull$(1, 2)$ distributions.

The figures suggest the following conclusions:

- increasing the order decreases the error in all cases;
- the error has a large scale U-shape as a function of λ;
- for some distributions (e.g., the uniform distributions) this U-shape behavior is altered by fluctuation, which decreases when increasing the order of the approximation;
- the optimal λ value, denoted as $\hat{\lambda}$, depends both on the distribution to approximate and the error measure;
- for some distributions (e.g., the uniform distributions and the Weibull$(1, 2)$ distribution) the optimal λ value is rather insensitive to the order (for those orders where the fluctuation is not present), while for other distributions (e.g., the Weibull$(1, 0.5)$ distribution) the optimal λ value depends on the order.

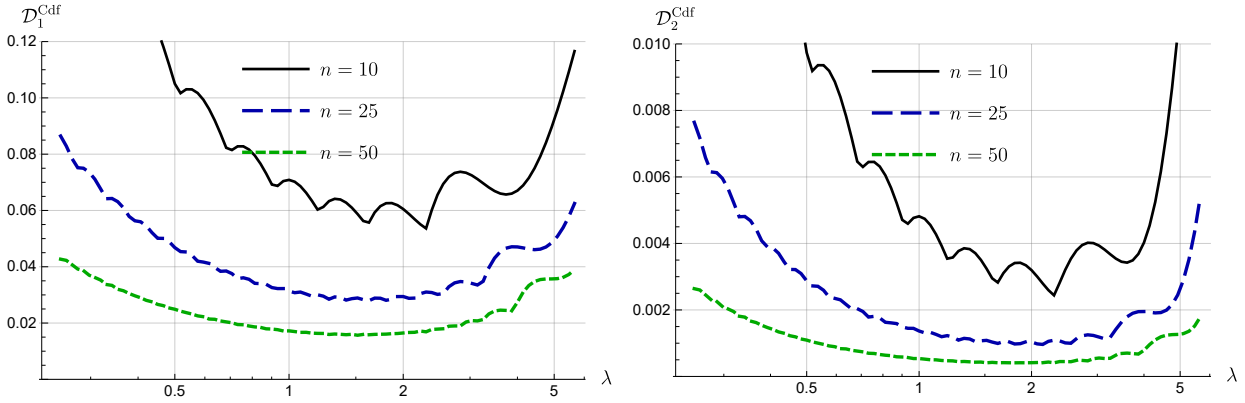

Figure 24: $\mathcal{D}_1^{\mathrm{Cdf}}$ (left) and $\mathcal{D}_2^{\mathrm{Cdf}}$ (right) error measures of order 10, 25 and 50 scaled BPH approximation of the CDF of the Uniform$(0, 1)$ distribution as a function of the scaling parameter. On the left, the minimum is attained at $\hat{\lambda}_{10} = 2.297$, $\hat{\lambda}_{25} = 1.625$, $\hat{\lambda}_{50} = 1.516$. On the right, at $\hat{\lambda}_{10} = 2.297$, $\hat{\lambda}_{25} = 2.144$, $\hat{\lambda}_{50} = 2.297$.

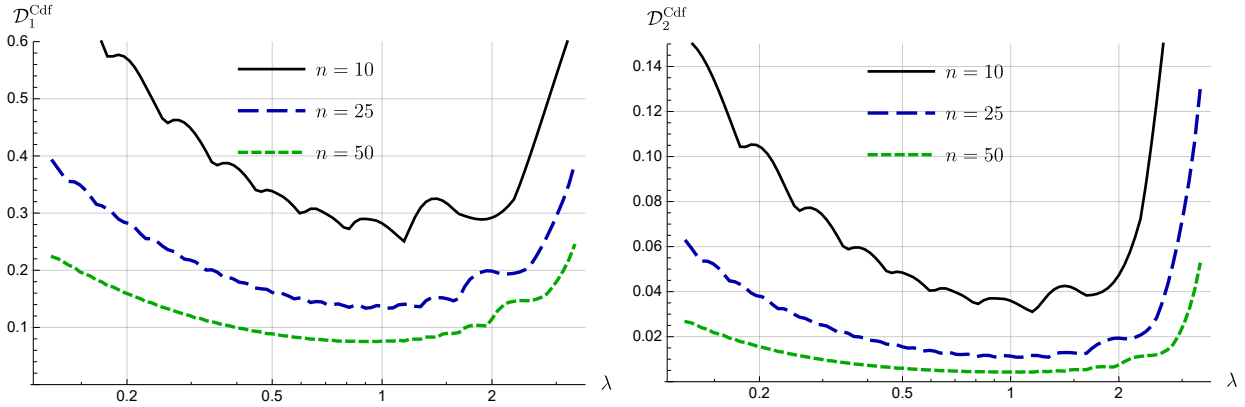

Figure 25: $\mathcal{D}_1^{\mathrm{Cdf}}$ (left) and $\mathcal{D}_2^{\mathrm{Cdf}}$ (right) error measures of order 10, 25 and 50 scaled BPH approximation of the CDF of the Uniform$(1, 2)$ distribution as a function of the scaling parameter. On the left, the minimum is attained at $\hat{\lambda}_{10} = 1.149$, $\hat{\lambda}_{25} = 0.9013$, $\hat{\lambda}_{50} = 0.9659$. On the right, at $\hat{\lambda}_{10} = 1.149$, $\hat{\lambda}_{25} = 1.072$, $\hat{\lambda}_{50} = 1.149$.

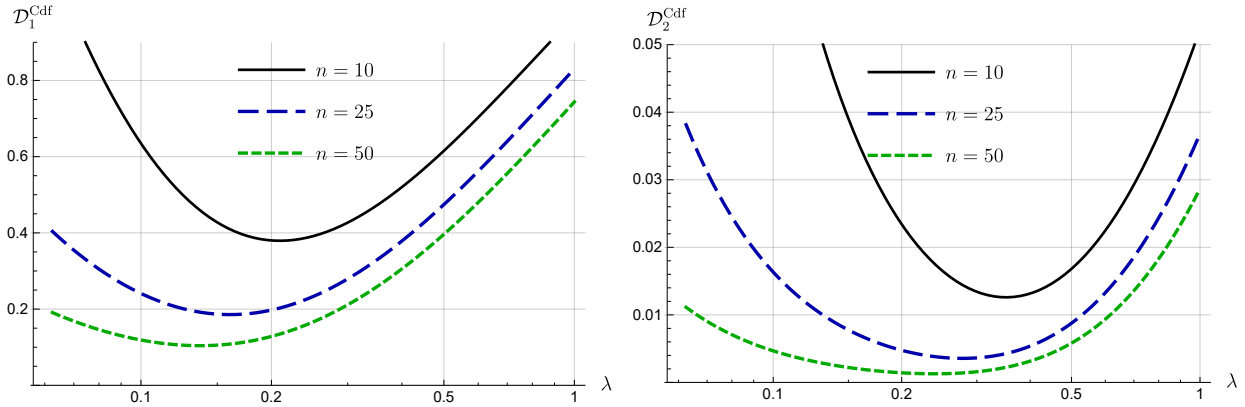

Figure 26: $\mathcal{D}_1^{\mathrm{Cdf}}$ (left) and $\mathcal{D}_2^{\mathrm{Cdf}}$ (right) error measures of order 10, 25 and 50 scaled BPH approximation of the CDF of the Weibull$(1, 0.5)$ distribution as a function of the scaling parameter. On the left, the minimum is attained at $\hat{\lambda}_{10} = 0.2102$, $\hat{\lambda}_{25} = 0.1593$, $\hat{\lambda}_{50} = 0.1387$. On the right, at $\hat{\lambda}_{10} = 0.3536$, $\hat{\lambda}_{25} = 0.2774$, $\hat{\lambda}_{50} = 0.2415$.

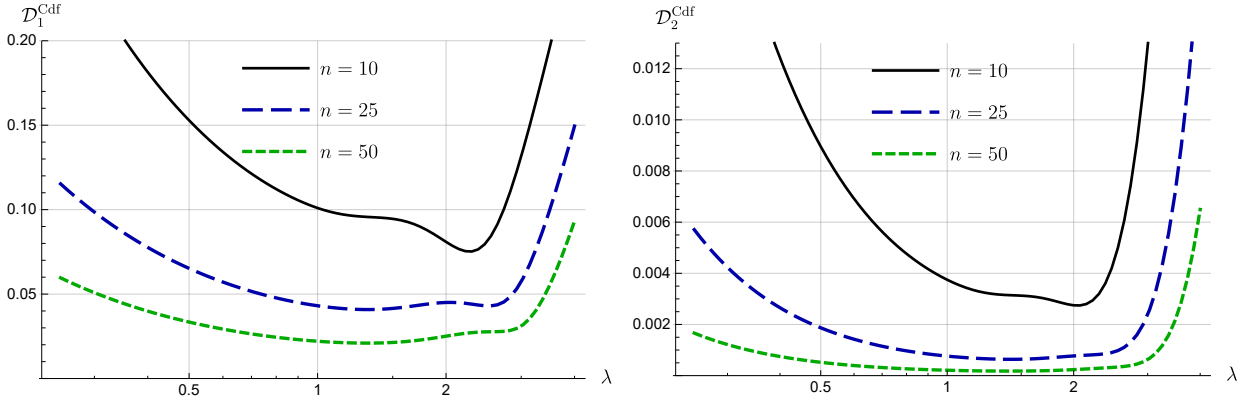

Figure 27: $\mathcal{D}_1^{\text{Cdf}}$ (left) and $\mathcal{D}_2^{\text{Cdf}}$ (right) error measures of order 10, 25 and 50 scaled BPH approximation of the CDF of the Weibull$(1, 2)$ distribution as a function of the scaling parameter. On the left, the minimum is attained at $\hat{\lambda}_{10} = 2.297$, $\hat{\lambda}_{25} = 1.319$, $\hat{\lambda}_{50} = 1.319$. On the right, at $\hat{\lambda}_{10} = 2.071$, $\hat{\lambda}_{25} = 1.414$, $\hat{\lambda}_{50} = 1.414$.

8. Conclusions

We applied Bernstein exponentials to the approximation of CDFs and showed that the resulting CDFs are valid and describe random variables that belong to a subclass of acyclic PH distributions, allowing efficient approximations of non-Markovian models. We also provided an approach to obtain stochastically ordered approximations, which opens the way to the application in problems where a safe approximation of performance metrics is required.

Additionally, we compared the CDF based BPH approximation, studied in this paper, and the previously studied PDF based BPH approximation and have found that the quality of the two approximations are rather similar, while CDF based BPH approximation has some additional advantages (does not require additional normalization). Finally, we studied the application of a scaling parameter in the BPH approximation and shown that the application of an appropriate scaling parameter is important for the accuracy of the approximation.

A. Proof of Theorem 3

We start with proving necessary lemmas assuming the conditions of Theorem 3.

Lemma 1. *There exists an $y_m > 0$ and n_m such that for any $n > n_m$,*

$$G(y) \leq \tilde{B}_n^+(G_{+\epsilon}; y) \qquad \forall y \in [0, y_m]\,. \tag{A.1}$$

Proof. Let

$$y_m := \min\left\{\hat{y}, \frac{\epsilon}{2C}\right\}, \text{ and } \quad n_m := \min\left\{n : (1-(1-y_m)^n) \geq \frac{1}{2}\right\}; \tag{A.2}$$

such an n_m exists since $\lim_{n\to\infty}(1-(1-y_m)^n) = 1$. The choice of y_m and n_m guarantees $\epsilon(1-(1-y_m)^{n_m}) \geq Cy_m$. The function $\epsilon(1-(1-y)^{n_m})$ is concave for $y \in [0, 1]$, so

$$\epsilon(1-(1-y)^{n_m})|_{y=0} \geq Cy|_{y=0} \quad \text{and} \quad \epsilon(1-(1-y)^{n_m})|_{y=y_m} \geq Cy|_{y=y_m}\,,$$

imply

$$\epsilon(1-(1-y)^{n_m}) \geq Cy \quad \text{for any} \quad y \in [0, y_m]\,.$$

For any fixed $y \in [0, 1]$, $\epsilon(1-(1-y)^n)$ is increasing in n, so for any $n > n_m$,

$$\epsilon(1-(1-y)^n) \geq \epsilon(1-(1-y)^{n_m}) \geq Cy \qquad \forall y \in [0, y_m]\,. \tag{A.3}$$

Finally, for any $n > n_m$ and $y \in [0, y_m]$, we have

$$\begin{aligned}\tilde{B}_n^+(G_{+\epsilon}; y) - G(y) &= B_n(G_{+\epsilon}; y) - \epsilon(1-y)^n - G(y) \\ &\overset{y\leq\hat{y}}{=} B_n(G_{+\epsilon}; y) - \epsilon(1-y)^n + \epsilon - G_{+\epsilon}(y) \\ &= \underbrace{\epsilon(1-(1-y)^n)}_{\geq Cy} + \underbrace{B_n(G_{+\epsilon}; y)}_{\geq \epsilon} - \underbrace{G_{+\epsilon}(y)}_{\leq \epsilon + Cy} \geq Cy + \epsilon - (\epsilon + Cy) = 0,\end{aligned}$$

where the first inequality is by (A.3), the second by the bound preserving property of the Bernstein operator and the third by the condition of Theorem 3. □

Lemma 2. *For any positive integer n_0, any $y_0 \in (0,1)$ and any $c > 0$ there exist $y_1, y_2 \in (0,1)$ such that for any $n \geq \lceil \frac{n_0+1}{y_0} \rceil$,*

$$\sum_{i=\lfloor ny_0 \rfloor}^{n} b_{n,i}(y) \leq cy^{n_0} \qquad \forall y \in [0, y_1] \tag{A.4}$$

and

$$\sum_{i=\lceil ny_0 \rceil}^{n} b_{n,i}(y) \geq 1 - c(1-y)^{n_0} \qquad \forall y \in [y_2, 1]\,, \tag{A.5}$$

where

$$b_{n,i}(y) = \binom{n}{i} y^i (1-y)^{n-i}\,. \tag{A.6}$$

Remark 3. *A possible interpretation of Lemma 2 is examining how smooth the Bernstein approximation is around $y = 0$ for the (discontinuous) function $h(y) = U(y - y_0)$, where U is the Heaviside function. While the lemma does not give an exact answer to that question, it will be sufficient for our purposes.*

Remark 4. *The main difficulty in Lemma 2 is finding a y_1 independent of n. The assumption $\lfloor ny_0 \rfloor \geq n_0$ guarantees that the leading term in the sum in* (A.4) *is of order smaller than y^{n_0} around 0, so for any fixed n, there clearly exists a proper y_1; however, the coefficients in the sum increase rapidly with n, so more work is necessary to obtain y_1 uniformly in n.*

Proof. (Lemma 2) We start by rearranging (A.4) as

$$\sum_{i=\lfloor ny_0 \rfloor}^{n} \binom{n}{i} y^{i-n_0}(1-y)^{n-i} \leq c\,. \tag{A.7}$$

We aim to prove that for a fixed n, the terms in (A.7) decay rapidly (at least exponentially) in i for $y < \frac{y_0}{2-y_0}$. This will allow us to bound the entire sum with the dominant term. Comparing consecutive terms in the sum gives

$$\frac{\binom{n}{i+1} y^{i+1-n_0}(1-y)^{n-(i+1)}}{\binom{n}{i} y^{i-n_0}(1-y)^{n-i}} = \frac{n-i}{i+1} \cdot \frac{y}{1-y} \leq \frac{n-ny_0}{ny_0} \cdot \frac{y}{1-y} = \frac{1-y_0}{y_0} \cdot \frac{y}{1-y},$$

where $\frac{1-y_0}{y_0}$ is constant, so the right-hand side can be made smaller than $1/2$ by choosing $y < \frac{y_0}{2-y_0}$, i.e.,

$$y < \frac{y_0}{2-y_0} \qquad \Longrightarrow \qquad \frac{1-y_0}{y_0} \cdot \frac{y}{1-y} \leq 1/2\,.$$

Then

$$\sum_{i=\lfloor ny_0\rfloor}^{n}\binom{n}{i}y^{i-n_0}(1-y)^{n-i}\le 2\binom{n}{\lfloor ny_0\rfloor}y^{\lfloor ny_0\rfloor-n_0}(1-y)^{n-\lfloor ny_0\rfloor}=2\binom{n}{\lfloor ny_0\rfloor}y^{\lfloor ny_0\rfloor-n_0-1}(1-y)^{n-\lfloor ny_0\rfloor}\cdot y,$$

where a factor of y is separated in the last expression for having a bounded term and a y dependent term. The term $2\binom{n}{\lfloor ny_0\rfloor}y^{\lfloor ny_0\rfloor-n_0-1}(1-y)^{n-\lfloor ny_0\rfloor}$ has its maximum at $y=(\lfloor ny_0\rfloor-n_0-1)/(n-n_0-1)$. We define

$$y^*=\min_n\left\{\frac{\lfloor ny_0\rfloor-n_0-1}{n-n_0-1}\right\},$$

where the minimum exists because $\lim_{n\to\infty}(\lfloor ny_0\rfloor-n_0-1)/(n-n_0-1)=y_0$.

Then for any $y<\min(y^*,\frac{y_0}{2-y_0})$, we have

$$2\binom{n}{\lfloor ny_0\rfloor}y^{\lfloor ny_0\rfloor-n_0-1}(1-y)^{n-\lfloor ny_0\rfloor}\cdot y\le 2\binom{n}{\lfloor ny_0\rfloor}(y^*)^{\lfloor ny_0\rfloor-n_0-1}(1-y^*)^{n-\lfloor ny_0\rfloor}\cdot y.$$

Next we use the fact that $\binom{n}{i}y^i(1-y)^{n-i}\le 1$ for any choice of n,i and y to get

$$\begin{aligned}&2\binom{n}{\lfloor ny_0\rfloor}(y^*)^{\lfloor ny_0\rfloor-n_0-1}(1-y^*)^{n-\lfloor ny_0\rfloor}\cdot y\\&\quad=\frac{2\binom{n}{\lfloor ny_0\rfloor}}{\binom{n-n_0-1}{\lfloor ny_0\rfloor-n_0-1}}\underbrace{\binom{n-n_0-1}{\lfloor ny_0\rfloor-n_0-1}(y^*)^{\lfloor ny_0\rfloor-n_0-1}(1-y^*)^{n-\lfloor ny_0\rfloor}}_{\le 1}\cdot y.\end{aligned}$$

Finally,

$$\frac{2\binom{n}{\lfloor ny_0\rfloor}}{\binom{n-n_0-1}{\lfloor ny_0\rfloor-n_0-1}}=2\frac{n(n-1)\dots(n-n_0)}{\lfloor ny_0\rfloor(\lfloor ny_0\rfloor-1)\dots(\lfloor ny_0\rfloor-n_0)}\to\frac{2}{y_0^{n_0+1}}\qquad\text{as } n\to\infty,$$

which is a finite constant. Setting

$$C^*=\max_{n>\lceil\frac{n_0+1}{y_0}\rceil}\left(\frac{2\binom{n}{\lfloor ny_0\rfloor}}{\binom{n-n_0-1}{\lfloor ny_0\rfloor-n_0-1}}\right)\quad\text{and}\quad y_1=\min\left(\frac{c}{C^*},y^*,\frac{y_0}{2-y_0}\right)$$

ensures (A.7) and thus (A.4).

To prove (A.5), we use the operator $J(G;y)=1-G(1-y)$. We obtain (A.5) by applying J to (A.4), proving Lemma 2. □

Proof. (Theorem 3.) The $\tilde{B}_n^+(G_{+\epsilon};y)\le G(y)+2\epsilon$ part of (23) is fairly straightforward. Applying B_n to the continuous function $G_{+\epsilon}$, uniform convergence guarantees that there exists an n_ϵ such that for any $n>n_\epsilon$,

$$B_n(G_{+\epsilon};y)\le G_{+\epsilon}(y)+\epsilon\,.$$

Then for any $n>n_\epsilon$, we have

$$\tilde{B}_n^+(G_{+\epsilon};y)=B_n(G_{+\epsilon};y)-\epsilon(1-y)^n\le B_n(G_{+\epsilon};y)\le G_{+\epsilon}(y)+\epsilon\le G(y)+2\epsilon\,.$$

The main challenge of (23) is the $G(y)\le\tilde{B}_n^+(G_{+\epsilon};y)$ part. Its proof relies on analyzing the behavior over the intervals $[0,y_m]$, $[y_m,y_M]$ and $[y_M,1]$ separately for $0<y_m<y_M<1$, where Lemma 1 proves the statement for $[0,y_m]$ and we continue with the $[y_M,1]$ interval.

Using (20), $\tilde{B}_n^+(G_{+\epsilon}; y)$ can be rewritten and estimated as

$$\tilde{B}_n^+(G_{+\epsilon}; y) = \sum_{i=1}^{\lceil \hat{y}n \rceil - 1} \left(G\left(\frac{i}{n}\right) + \epsilon \right) b_{n,i}(y) + \sum_{i=\lceil \hat{y}n \rceil}^{n} b_{n,i}(y) \geq \sum_{i=\lceil \hat{y}n \rceil}^{n} b_{n,i}(y)\,. \tag{A.8}$$

We apply (A.5) with $n_0 = n_d$ (where n_d is from Theorem 3), $y_0 = \hat{y}$ and

$$c = \frac{1}{2} \left| \frac{\mathrm{d}^{n_d}}{\mathrm{d}y^{n_d}} G(y) \bigg|_{y=1} \right| \tag{A.9}$$

(where the absolute value is necessary since the sign depends on the parity of n_d) to obtain a y_2 such that for any $n > \lceil \frac{n_d+1}{y_0} \rceil$,

$$\tilde{B}_n^+(G_{+\epsilon}; y) \geq 1 - c(1-y)^{n_d} \qquad \forall y \in [y_2, 1]\,.$$

Choosing c according to (A.9) guarantees

$$1 - c(1-y)^{n_d} \geq G(y)$$

on some interval $[y_3, 1]$, since the first $n_d - 1$ derivatives of both sides are zero and the n_dth derivative of the left-hand side is double of the derivative of the right-hand side at $y = 1$. Consequently, we can choose

$$y_M = \max\{y_2, y_3\}$$

to obtain that for any $n \geq \lceil \frac{n_d+1}{y_0} \rceil$,

$$\tilde{B}_n^+(G_{+\epsilon}; y) \geq G(y) \qquad \forall y \in [y_M, 1]\,. \tag{A.10}$$

At this point, we have $\tilde{B}_n^+(G_{+\epsilon}; y) \geq G(y)$ on $[0, y_m]$ and also on $[y_M, 1]$; what remains is the interval $[y_m, y_M]$. Let

$$\delta = \min(\epsilon, 1 - G(y_M)) > 0\,.$$

For this δ we have

$$G_{+\epsilon}(y) - G(y) \geq \delta > 0 \qquad \forall y \in [y_m, y_M], \tag{A.11}$$

because if $\hat{y} \geq y_M$ then $G_{+\epsilon}(y) - G(y) = \epsilon$ for $\forall y \in [y_m, y_M]$, and if $\hat{y} \leq y_M$ then $G_{+\epsilon}(y) - G(y) \geq 1 - G(y_M)$ for $\forall y \in [y_m, y_M]$.

Due to uniform convergence, there exists an n_c such that for any $n > n_c$,

$$|B_n(G_{+\epsilon}; y) - G_{+\epsilon}(y)| \leq \delta \qquad \forall y \in [y_m, y_M]; \tag{A.12}$$

from (A.11) and (A.12) it follows that

$$B_n(G_{+\epsilon}; y) \geq G(y) \qquad \forall y \in [y_m, y_M]\,. \tag{A.13}$$

Putting together (A.1), (A.10) and (A.13), we obtain that

$$\hat{n} = \max\{n_\epsilon, n_m, \left\lceil \frac{n_d + 1}{y_0} \right\rceil, n_c\}$$

is a suitable choice so that, for any $n > \hat{n}$,

$$B_n(G_{+\epsilon}; y) \geq G(y) \qquad \forall y \in [0, 1], \tag{A.14}$$

proving (23) and Theorem 3. □

References

[1] M. Neuts, Probability distributions of phase type, in: Liber Amicorum Prof. Emeritus H. Florin, University of Louvain, 1975, pp. 173–206.

[2] S. Asmussen, O. Nerman, Fitting phase-type distributions via the EM algorithm, in: Proceedings of Symposium i Advent Statistik, Copenhagen, 1991, pp. 335–346.

[3] A. Bobbio, A. Cumani, ML estimation of the parameters of a PH distribution in triangular canonical form, in: Computer Performance Evaluation, Elsevier, 1992, pp. 33–46.

[4] A. Bobbio, A. Horváth, M. Telek, Matching three moments with minimal acyclic phase type distributions, Stochastic Models 21 (2005) 303–326.

[5] M. Telek, G. Horváth, A minimal representation of Markov arrival processes and a moments matching method, Performance Evaluation 64 (9-12) (2007) 1153–1168.

[6] A. Feldman, W. Whitt, Fitting mixtures of exponentials to long-tail distributions to analyze network performance models, Performance Evaluation 31 (1998) 245–279.

[7] A. Riska, V. Diev, E. Smirni, An EM-based technique for approximating long-tailed data sets with PH distributions, Performance Evaluation 55 (1-2) (2004) 147–164.

[8] A. Horváth, M. Telek, Approximating heavy tailed behavior with phase-type distributions, in: Proc. of 3rd International Conference on Matrix-Analytic Methods in Stochastic models, Leuven, Belgium, 2000, pp. 191–214.

[9] M. Neuts, Matrix Geometric Solutions in Stochastic Models, Johns Hopkins University Press, Baltimore, 1981.

[10] G. Latouche, V. Ramaswami, Introduction to Matrix Analytic Methods in Stochastic Modeling, SIAM, 1999.

[11] A. Horváth, E. Vicario, Construction of phase type distributions by Bernstein exponentials, in: Proceedings of ASMTA 2023, Vol. 14231 of Lecture Notes in Computer Science, Springer, 2023, pp. 201–215.

[12] G. M. Phillips, Interpolation and Approximation by Polynomials, Springer New York, 2003.

[13] T. J. Rivlin, An Introduction to the Approximation of Functions, Courier Corporation, 1981.

[14] J.-M. Fourneau, N. Pekergin, A numerical analysis of dynamic fault trees based on stochastic bounds, in: International Conference on Quantitative Evaluation of Systems, Springer, 2015, pp. 176–191.

[15] F. Baccelli, A. M. Makowski, Multidimensional stochastic ordering and associated random variables, Operations Research 37 (3) (1989) 478–487.

[16] A. Horváth, I. Horváth, M. Paolieri, M. Telek, E. Vicario, Approximation of cumulative distribution functions by Bernstein phase-type distributions, in: Proceedings of QEST+FORMATS 2024, Vol. 14996 of Lecture Notes in Computer Science, Springer, 2024, pp. 90–106.

[17] D. Aldous, L. Shepp, The least variable phase type distribution is Erlang, Stochastic Models 3 (3) (1987).