Differentiable Fit of Empirical Distributions by Bernstein Phase Types

Laura Carnevali[1], András Horváth[2], Marco Paolieri[3], and Enrico Vicario[1]

[1] Department of Information Engineering, University of Florence, Italy
{laura.carnevali,enrico.vicario}@unifi.it

[2] Department of Computer Science, University of Turin, Italy
horvath@di.unito.it

[3] Electrical and Computer Engineering, University of Southern California, USA
paolieri@usc.edu

Abstract. The empirical cumulative distribution function (ECDF) is an efficient estimator of the distribution of sampled data. Yet, it has discontinuities at each step (where probability mass is concentrated) which result in a probability density function (PDF) with discrete spikes, preventing its use in models that require continuous PDFs. Bernstein approximants bridge this gap, yielding smooth estimators for bounded ECDFs and their PDFs. In this paper, we consider a strategy from the literature to select the order of Bernstein polynomials (BP) over bounded support, and we extend the approach to Bernstein phase types (BPH) in order to approximate distributions with unbounded support. Preliminary experiments confirm the effectiveness of this approach in balancing the bias-variance tradeoff, giving insight into the relation between the sample size and the Bernstein order. Comparative experiments with Gaussian kernel density estimators show that the BPH approximant achieves better or comparable performance on a benchmark of Erlang PDFs with unit mean, while guaranteeing a low memory footprint thanks to its closed form and allowing integration into continuous-time Markov chains. We provide a replication package to support reproducibility of experiments.

Keywords: Empirical Cumulative Distribution Function, Bernstein Polynomials, Bernstein Phase Types, Gaussian Kernel Density Estimation

1 Introduction

In most practical applications, the distributions of stochastic quantities are not known in advance, but they must be estimated from independent observations. Parametric approaches select a family of distributions and fit the distribution parameters from sample data, e.g., to maximize likelihood or to match statistic properties of the data; in contrast, nonparametric approaches construct an estimate of the distribution from the data itself, with tunable assumptions on the smoothness of the real distribution. For example, kernel density estimation (KDE) [2,15,9] is a popular nonparametric approach where the real probability density function (PDF) is estimated as the sum of kernel functions (e.g., Gaussian or Uniform PDFs with zero mean) centered at each data point; the *bandwidth* parameter (e.g., the variance of the kernel functions) is used to control smoothness.

Among parametric approaches, the family of phase-type (PH) distributions [18,13] has received significant attention for its power and flexibility. PH distributions are defined as the time to absorption in Markov chains; they provide a flexible tradeoff between approximation accuracy and complexity, which is obtained by varying the number of transient states, referred to as *phases*, that can be visited before reaching the absorbing state. PH distributions have closed-form densities and Laplace transforms, enjoy closure under finite mixtures and convolutions, and are dense in the set of positive-valued distributions (i.e., they can approximate any positive-valued distribution with arbitrary accuracy, provided enough phases). When used in stochastic models, PH distributions result in Markov chains with regular structures amenable to efficient analysis through matrix analytic methods [19,16]. Fitting methods have been proposed to find PH parameters (transition rates and initial state probabilities) that maximize likelihood [1,4] or that match moments [5,23], tail behavior [8,21], or both [10].

More recently, an approach was proposed to approximate PDFs [14] and cumulative distribution functions (CDFs) [11,12] using *Bernstein exponentials* (BEs), i.e., linear combinations of Bernstein polynomials (BP) [20,22] where the support $x \in [0,1]$ is mapped to $y \in [0,\infty)$ through the change of variable $x = e^{-y}$. This approach results in a subclass of acyclic PH distributions called Bernstein phase type (BPH) distributions, which preserve shape properties of approximated CDFs (including monotonicity, upper and lower bounds, and exact limit values at 0 and $+\infty$), while allowing fast, closed-form fitting of CDFs using only their values at specific points (specifically, one for each phase).

When applied to the approximation of the empirical cumulative distribution function (ECDF) derived from sample data, the selection of the number of phases of the BPH distribution plays an important role in the bias-variance tradeoff [6]: using more phases reduces the approximation error (i.e., the error due to the choice of model) but increases the estimation error (i.e., the error due to the limited size of the data sample). Since a BE of the order N results in a BPH with N phases, selecting the number of phases is equivalent to the selection of the order of a Bernstein exponential. Prior work analyzed the properties of Bernstein polynomials for the approximation of ECDFs; in particular, if the true CDF has a continuous PDF f that is Lipschitz of order 1, the approximation of the ECDF with BPs results in a PDF that converges uniformly to f, provided that the order N of the BP is such that $N = o(M/\log(M))$, where $N \to \infty$ and $M \to \infty$ is the number of samples that define the ECDF [3]. Higher order expansions for the asymptotic mean-squared error are provided in [17].

In this paper, we apply the analysis of BPs of [3] to select the order of BEs, i.e., to select the number of phases of BPHs, providing ECDF smooth estimators within the class of PHs, which has major relevance in stochastic modeling. In our experiments, we characterize the intertwined effects of two distinct mechanisms: *i*) the approximation error inherent to BPH approximants, and *ii*) the estimation error due to the sample size. On the one hand, our experiments use Erlang distributions with unit mean to show that the coefficient of variation of the ECDF impacts the estimation error. On the other hand, experimental results show that, if the sample size M is sufficiently representative of the true CDF, then the accuracy of the estimated PDF increases with N only until $N \approx M/\log(M)$; conversely, if the sample size M is insufficient, increasing the Bernstein order N does not improve estimation accuracy. Based on these considerations, we report key findings that provide guidance on how to apply the proposed approach to derive a BPH estimator of an ECDF. We also compare our parametric approach based on BPHs with nonparametric KDE using Gaussian kernels, showing that BPH estimation achieves better or similar accuracy, while maintaining a low memory footprint thanks to its compact representation, which requires a number of parameters equal to the Bernstein order N (instead of the dataset size M). The paper is accompanied by a replication package [7] with open-source license.

The rest of the paper is organized as follows. First, we illustrate BP and BPH approximants, and we summarize the approach of [3] to select the order of BPs approximating PDFs with bounded support (Section 2). Then, we define our experimental setup using BPHs to approximate PDFs with unbounded support (Section 3) and we present experimental results (Section 4). Finally, we draw our conclusions and discuss next steps (Section 5).

2 Bernstein Approximants

In this section, we report background on BPs [3,20] (Section 2.1) and BPHs [11] (Section 2.2); then, we recall the results of [3] for the selection of the order of BPs and discuss how to apply it to the selection of the number of phases of BPHs (Section 2.3).

2.1 Bernstein Polynomials

Given a CDF $F\colon [0,1] \to [0,1]$ and the ECDF

$$F_M(x) = \frac{1}{M} \sum_{i=1}^{M} I\{X_i \leq x\} \tag{1}$$

obtained from random samples $X_1, X_2, \ldots, X_M$ distributed according to F, the BP approximant of F_M with order N is defined as

$$B_N(F_M; x) := \sum_{n=0}^{N} F_M\left(\frac{n}{N}\right) \binom{N}{n} x^n (1-x)^{N-n} . \tag{2}$$

Note that $B_N(F_M; x)$ preserves the values of F_M at 0 and 1, the lower and upper bounds (i.e., $l \leq F_M(x) \leq u \; \forall x \in [0,1] \implies l \leq B_N(F_M; x) \leq u \; \forall x \in [0,1]$), and monotonicity (i.e., $B_N(F_M; x)$ is monotonically nondecreasing given that $F_M(x)$ is monotonically nondecreasing). According to this, $B_N(F_M; x)$ is a valid

CDF for all $N > 0$. Moreover, the approximation error $|F_M(x) - B_N(F_M; x)|$ converges uniformly to 0 over the support $[0, 1]$ as $N \to \infty$ (i.e., $\forall \epsilon > 0, \exists \bar{N} \in \mathbb{N}$ such that $N > \bar{N} \implies |F_M(x) - B_N(F_M; x)| < \epsilon \; \forall x \in [0, 1]$). To approximate the PDF f of F, the derivative of $B_N(F_M; x)$ for $x \in [0, 1]$ can be calculated as

$$B'_N(F_M; x) = \sum_{n=1}^{N} \left(F_M\left(\frac{n}{N}\right) - F_M\left(\frac{n-1}{N}\right) \right) n \binom{N}{n} x^{n-1}(1-x)^{N-n} \,. \tag{3}$$

Proof. By linearity of differentiation, $B'_N(F_M; x) = \sum_{n=0}^{N} F_M\left(\frac{n}{N}\right) \cdot T'_{N,n}(x)$ where $T_{N,n}(x) = \binom{N}{n} x^n (1-x)^{N-n}$ and, for $n = 0, 1, \ldots, N$ and $x > 0$,

$$\begin{aligned} T'_{N,n}(x) &= n \binom{N}{n} x^{n-1}(1-x)^{N-n} - (N-n) \binom{N}{n} x^n (1-x)^{N-(n+1)} \\ &= \Big(n\, T_{N,n}(x) - (n+1)\, T_{N,n+1}(x) \Big) / x \,. \end{aligned}$$

Note that $T_{N,N+1}(x) = 0$ and that we used $(N-n)\binom{N}{n} = (n+1)\binom{N}{n+1}$. Then,

$$\begin{aligned} B'_N(F_M; x) &= \sum_{n=0}^{N} F_M\left(\frac{n}{N}\right) \cdot \Big(n\, T_{N,n}(x) - (n+1)\, T_{N,n+1}(x) \Big) / x \\ &= \left(\sum_{n=1}^{N} F_M\left(\frac{n}{N}\right) n\, T_{N,n}(x)/x \right) - \left(\sum_{n=1}^{N} F_M\left(\frac{n-1}{N}\right) n\, T_{N,n}(x)/x \right) \\ &= \sum_{n=1}^{N} \left(F_M\left(\frac{n}{N}\right) - F_M\left(\frac{n-1}{N}\right) \right) n\, T_{N,n}(x)/x \,. \end{aligned}$$

The result is also valid at $x = 0$, where $B'_N(F_M; 0) = N\,(F_M(1/N) - F_M(0))$, and, by continuity, for all $x \in [0, 1]$. □

If F has a compact support $[a, b]$, i.e., $F \colon [a, b] \to [0, 1]$, its BP approximant of order N is obtained from Eq. (2) by the change of variables $x = (y-a)/(b-a)$:

$$\begin{aligned} B_N^{[a,b]}(F_M; y) &= B_N(F_M; (y-a)/(b-a)) \\ &= \sum_{n=0}^{N} F_M\left(a + \frac{n}{N}(b-a)\right) \binom{N}{n} \frac{(y-a)^n (b-y)^{N-n}}{(b-a)^N} , \end{aligned} \tag{4}$$

where F_M is the ECDF obtained from a random sample with CDF F and size M. Since the change of variables is continuous and strictly monotonic, $B_N^{[a,b]}(F_M; y)$ inherits the shape preservation properties of $B_N(F_M; x)$. Moreover, to approximate the PDF f of F, the derivative of $B_N^{[a,b]}(F_M; y)$ for $y \in [a, b]$ can be calculated as

$$\frac{d}{dy} B_N^{[a,b]}(F_M; y) = \frac{d}{dy} B_N(F_M; (y-a)/(b-a)) = \frac{1}{b-a} B'_N(F_M; (y-a)/(b-a)) \,.$$

2.2 Bernstein Phase Types

If F has unbounded support $[0, \infty)$, i.e., $F\colon [0,\infty) \to [0,1]$, the BPH approximant of order N of its ECDF is obtained from Eq. (2) through the change of variable $x = e^{-y}$:

$$BPH_N(F_M; y) = \sum_{n=0}^{N} F_M\left(\log\left(\frac{N}{n}\right)\right)\binom{N}{n} e^{-ny}(1-e^{-y})^{N-n}, \tag{5}$$

where F_M is the ECDF obtained from a random sample with CDF F and size M. The division by zero for $n = 0$ is resolved by considering the limiting value of $F_M(x)$ as x tends to infinity, i.e., $F_M\left(\log \frac{N}{0}\right) := \lim_{x\to\infty} F_M(x) = 1$; for $n = N$, we have $F_M(\log(N/N)) = F_M(0)$ which is 0 if the ECDF has no samples equal to 0. Since the change of variables is continuous and strictly monotonic, $BPH_N(F_M; y)$ inherits the shape preservation properties of $B_N(F_M; x)$. Moreover, the derivative of $BPH_N(F_M; y)$, which we use to approximate the PDF f of F, is equal to

$$\begin{aligned} BPH'_N(F_M; y) &= \frac{d}{dy}\, B_N(F_M; e^{-y}) \\ &= -e^{-y}\, B'_N(F_M; e^{-y}) \\ &= \sum_{n=1}^{N}\left(F_M\left(\log\left(\frac{N}{n-1}\right)\right) - F_M\left(\log\left(\frac{N}{n}\right)\right)\right) \\ &\qquad\qquad \cdot n\binom{N}{n} e^{-ny}(1-e^{-y})^{N-n}\,. \end{aligned} \tag{6}$$

2.3 Fitting Approach

Related work [3] studies the asymptotic properties observed when the true CDF $F(x)$ and PDF $f(x)$ with support $[0,1]$ are approximated using BPs $B_N(F_M; x)$, where N is the order of the BP and $F_M(x)$ is an instance of the ECDF of F using M samples. Specifically, if F is continuous over $[0,1]$, then $B_N(F_M; x)$ converges w.p.1 uniformly to F as $N, M \to \infty$. More interestingly, if F is continuous and differentiable, and f is Lipschitz of order 1, then $B_N(F_M; x)$ converges w.p.1 uniformly to F as $N \to \infty$, provided that $M^{2/3} \leq N \leq (M/\log(M))^2$. A similar result is shown for $B'_N(F_M; x)$, proving that, if the CDF F has a continuous PDF f that is Lipschitz of order 1, then $B'_N(F_M; x)$ converges w.p.1 uniformly to f as $N, M \to \infty$, provided that $N = o(M/\log(M))$.

We extend the order selection approach proposed in [3] to the BPH estimators of a PDF f with unbounded support $[0, \infty)$. In particular, we experimentally show that, if the sample size M is sufficiently representative of the CDF F, then the accuracy of the BPH estimator of the PDF f increases as the Bernstein order N increases, until the threshold $M/\log(M)$ is reached, beyond which the approximation accuracy no longer improves and indeed becomes worse. Conversely, if the sample size M is insufficient, then increasing the Bernstein order N does not improve the accuracy of the BPH estimator of f, which actually remains poor.

3 Experimental Setup

We consider Erlang CDFs $\mathcal{E}_K(x)$ having both shape and rate parameters equal to $K \in \{2, 4, 8, 16\}$, thus resulting in expected value equal to 1 and variance $1/K$. As described in Section 2.2, we approximate the PDF of $\mathcal{E}_K(x)$ by the derivative of the BPH approximant of its ECDF, i.e., $bph_N(\mathcal{E}_{K,M}; x) := BPH'_N(\mathcal{E}_{K,M}; x)$, where $\mathcal{E}_{K,M}(x)$ is the ECDF of $\mathcal{E}_K(x)$ for a random sample of size M.

For each value of K, we compute $bph_N(\mathcal{E}_{K,M}; x)$ for each Bernstein order $N \in \{8, 16, 32, 64, 128, 256\}$ and for each number of samples $M \in \{27, 68, 163, 381, 866, 1938\}$. The values of N correspond to the bound $M/\log(M)$ defined in [3], i.e., $8 \approx 27/\log(27)$, $16 \approx 68/\log(68)$, ..., and $256 \approx 1938/\log(1938)$.

We evaluate the accuracy of $bph_N(\mathcal{E}_{K,M}; x)$ with respect to the PDF of $\mathcal{E}_K(x)$, denoted by $\mathcal{E}'_K(x)$, through a discrete approximation of the Kullback-Leibler Divergence (KLD):

$$D_{KL}\left(\mathcal{E}'_K(x) \,||\, bph_N(\mathcal{E}_{K,M}; x)\right) = \sum_{x \in \mathcal{X}} \mathcal{E}'_K(x) \log\left(\frac{\mathcal{E}'_K(x)}{bph_N(\mathcal{E}_{K,M}; x)}\right)\delta \qquad (7)$$

where $\mathcal{X}$ is the set of 500 equidistant time points covering the interval $[10^{-9}, l]$ such that $1 - \mathcal{E}_K(l) = 10^{-3}$ (i.e., the probability that the Erlang exceeds l is equal to 10^{-3}) and δ is their distance. The evaluation of the KLD of $bph_N(\mathcal{E}_{K,M}; x)$ is repeated for 10^5 random samples of size M (each resulting in a different $\mathcal{E}_{K,M}$); we report mean and standard deviation of the KLD across these samples.

For each value of K and M, we compare our BPH approximant with Gaussian Kernel Density Estimation (KDE) using the samples x_m that define $\mathcal{E}_{K,M}$:

$$kde_h(\mathcal{E}_{K,M}; x) = \frac{1}{M}\sum_{m=1}^{M} \frac{1}{h\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x - x_m}{h}\right)^2\right) \qquad (8)$$

where h is a smoothing parameter called *bandwidth*, selected according to Scott's rule $h = \sigma M^{-1/5}$ (σ is the standard deviation of the M values in a sample). Also for KDE, we evaluate the KLD with respect to the ground truth PDF $\mathcal{E}'_K(x)$, i.e., $D_{KL}\left(\mathcal{E}'_K(x) \,||\, kde_h(\mathcal{E}_{K,M}; x)\right)$, over the same 10^5 random samples of size M.

4 Experimental Results

4.1 Bernstein Order Selected from Sample Size

Table 1 shows the accuracy of the BPH and KDE estimators of Erlang PDFs with shape and rate equal to $K \in \{2, 4, 8, 16\}$, i.e., mean value and standard deviation of the KLD computed over 10^5 random samples of size $M \in \{27, 68, 163, 381, 866, 1938\}$. Specifically, Table 1 considers the cases where the Bernstein order N is selected according to $N \approx M/\log(M)$, as suggested in [3] for BPs. A subset of these experimental results is visualized in Figs. 1 to 4: Figs. 1 and 2 plot the BPH and KDE estimators obtained for $K = 2$ and for each value of M;

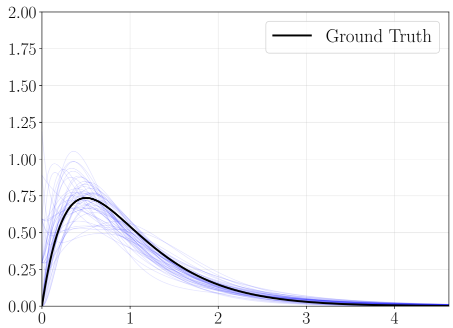
K	M	N	KLD (mean)		KLD (std)	
			bph	*kde*	*bph*	*kde*
2	27	8	0.0381	0.1260	0.0276	0.1391
	68	16	0.0209	0.0607	0.0138	0.0525
	163	32	0.0123	0.0314	0.0070	0.0209
	381	64	0.0076	0.0174	0.0037	0.0084
	866	128	0.0049	0.0104	0.0020	0.0033
	1938	256	0.0032	0.0069	0.0011	0.0013
4	27	8	0.0647	0.1055	0.0303	0.1036
	68	16	0.0285	0.0530	0.0161	0.0362
	163	32	0.0131	0.0291	0.0077	0.0151
	381	64	0.0069	0.0165	0.0036	0.0065
	866	128	0.0042	0.0098	0.0019	0.0030
	1938	256	0.0028	0.0060	0.0011	0.0015
8	27	8	0.1448	0.0880	0.0292	0.0801
	68	16	0.0659	0.0431	0.0190	0.0283
	163	32	0.0264	0.0226	0.0105	0.0123
	381	64	0.0102	0.0120	0.0049	0.0056
	866	128	0.0045	0.0066	0.0022	0.0027
	1938	256	0.0024	0.0037	0.0010	0.0013
16	27	8	0.2777	0.0755	0.0250	0.0695
	68	16	0.1500	0.0362	0.0180	0.0243
	163	32	0.0675	0.0184	0.0124	0.0104
	381	64	0.0259	0.0097	0.0070	0.0049
	866	128	0.0091	0.0052	0.0033	0.0023
	1938	256	0.0033	0.0028	0.0014	0.0012

Table 1: Accuracy of BPHs (with order N) and KDEs in approximating Erlang PDFs with shape and rate equal to K: mean value and standard deviation of the KLD with respect to the ground truth, over 10^5 random samples of size M.

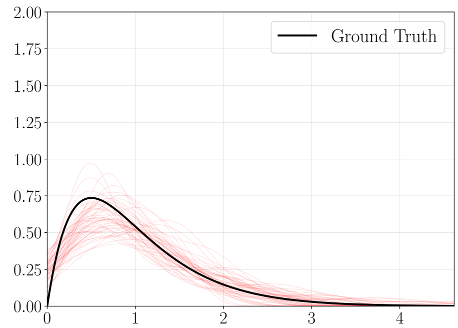
Figs. 3 and 4 plot those obtained for $K = 4$ and $K = 16$, respectively, and for $M \in \{68, 381, 1938\}$. Each plot shows the true Erlang PDF (thick black curve) and the estimators obtained using BPHs (thin blue curves) or KDEs (thin red curves) for 50 random samples of size M (out of the 10^5 samples used in Table 1).

The results in Table 1 show that, for each value of K, the mean value and the standard deviation of the KLD decrease as the sample size M increases, given that increasing the number of samples provides a more accurate empirical estimate of the ground truth PDF. This trend is also evident in Figs. 1 to 4, as both BPHs and KDEs converge to the ground truth Erlang PDF as M grows.

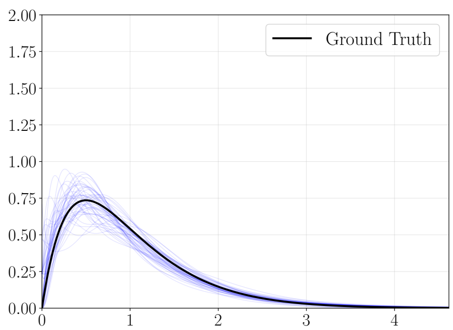
As K increases, the probability mass of the Erlang PDF concentrates over a narrower support, making approximation difficult for BPHs, due to their limitations in fitting data with low coefficient of variation using a small number of phases, and also due to the fact that the parameters of the BPH approximant are estimated in large part using values of the ECDF over the left end of the support. In contrast, KDEs are structurally capable of fitting such data given that the kernel variance can be scaled down arbitrarily. Moreover, as K grows, the Erlang PDFs also become less right-skewed, improving the performance of KDEs due to the symmetric shape of their kernels.



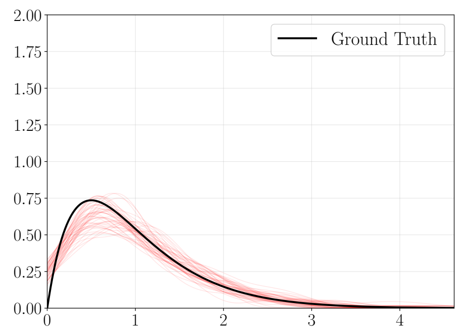
(a) BPH: $M = 27$, $N = 8$



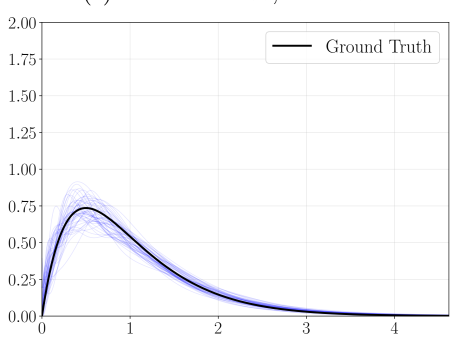
(b) KDE: $M = 27$



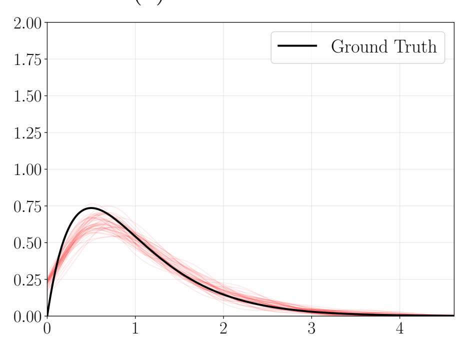
(c) BPH: $M = 68$, $N = 16$



(d) KDE: $M = 68$



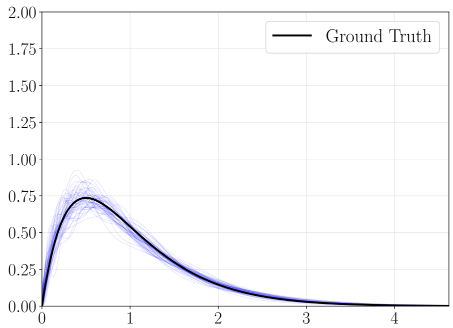
(e) BPH: $M = 163$, $N = 32$



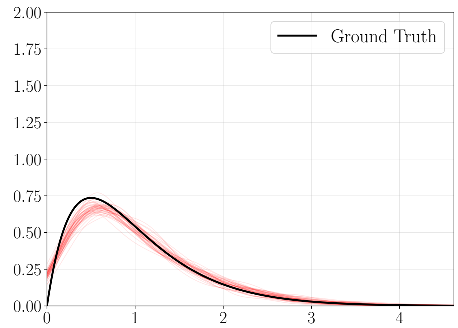
(f) KDE: $M = 163$

Fig. 1: BPH approximants with order N (a,c,e) and KDE approximants (b,d,f) for 50 different samples of size M of an Erlang PDF with shape and rate $\boldsymbol{K = 2}$.

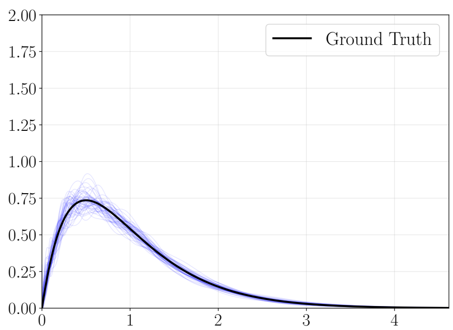
In Table 1, for small values of M, the mean KLD of BPHs nearly doubles as K doubles, e.g., for $M = 27$, it is equal to 0.0381, 0.0647, 0.1448, and 0.2777 for K equal to 2, 4, 8, and 16, respectively. Similarly, the minimum value of N needed to achieve mean KLD close to 10^{-2} tends to increase with K, i.e., it is equal to 32, 32, 64, and 128 for $K = 2$, $K = 4$, $K = 8$, and $K = 16$, respectively.



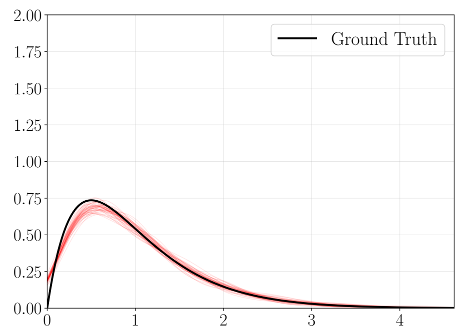
(a) BPH: $M = 381$, $N = 64$



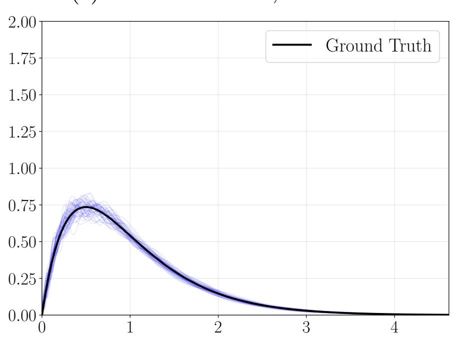
(b) KDE: $M = 381$



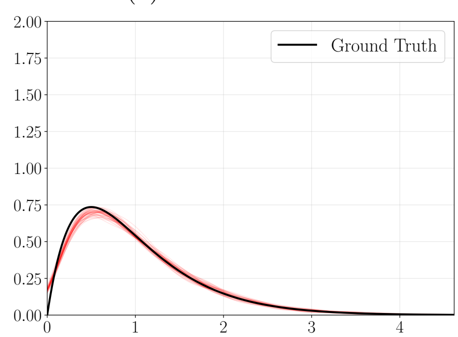
(c) BPH: $M = 866$, $N = 128$



(d) KDE: $M = 866$



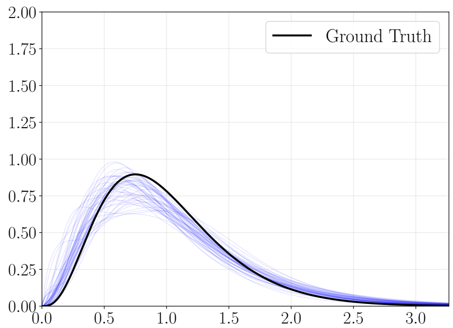
(e) BPH: $M = 1938$, $N = 256$



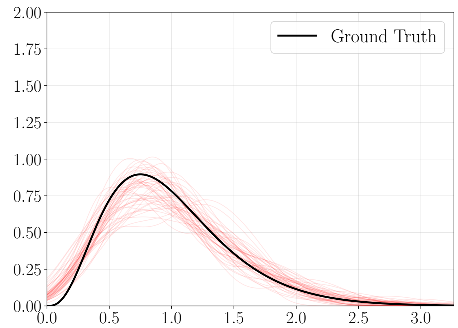
(f) KDE: $M = 1938$

Fig. 2: BPH approximants with order N (a,c,e) and KDE approximants (b,d,f) for 50 different samples of size M of an Erlang PDF with shape and rate $\boldsymbol{K = 2}$.

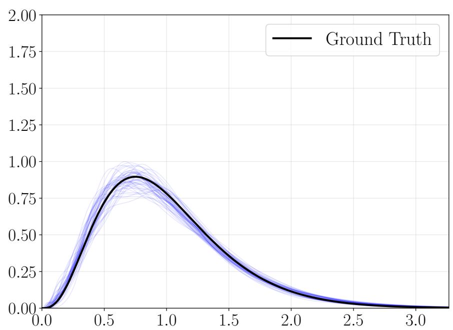
Nevertheless, such accuracy can be achieved for all the values of K considered in this experimental evaluation. Conversely, for KDEs, the approximation complexity does not increase with K: in fact, for a given sample size M, KDEs exhibit a slight reduction in the KLD as K increases, due to the fact that the shape of the Erlang PDF more closely resembles a Gaussian PDF. Notably, compared to KDEs, BPHs achieve lower mean value and standard deviation of the KLD



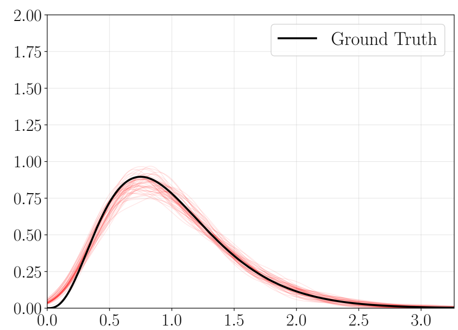
(a) BPH: $M = 68$, $N = 16$



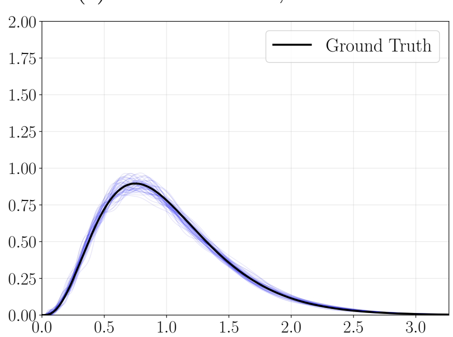
(b) KDE: $M = 68$



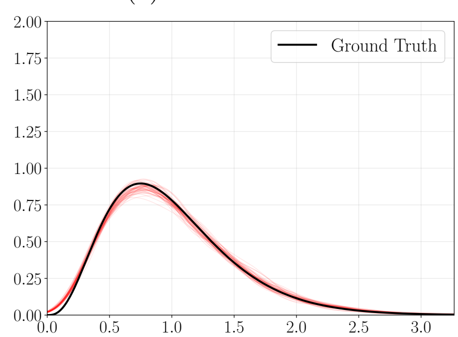
(c) BPH: $M = 381$, $N = 64$



(d) KDE: $M = 381$



(e) BPH: $M = 1938$, $N = 256$

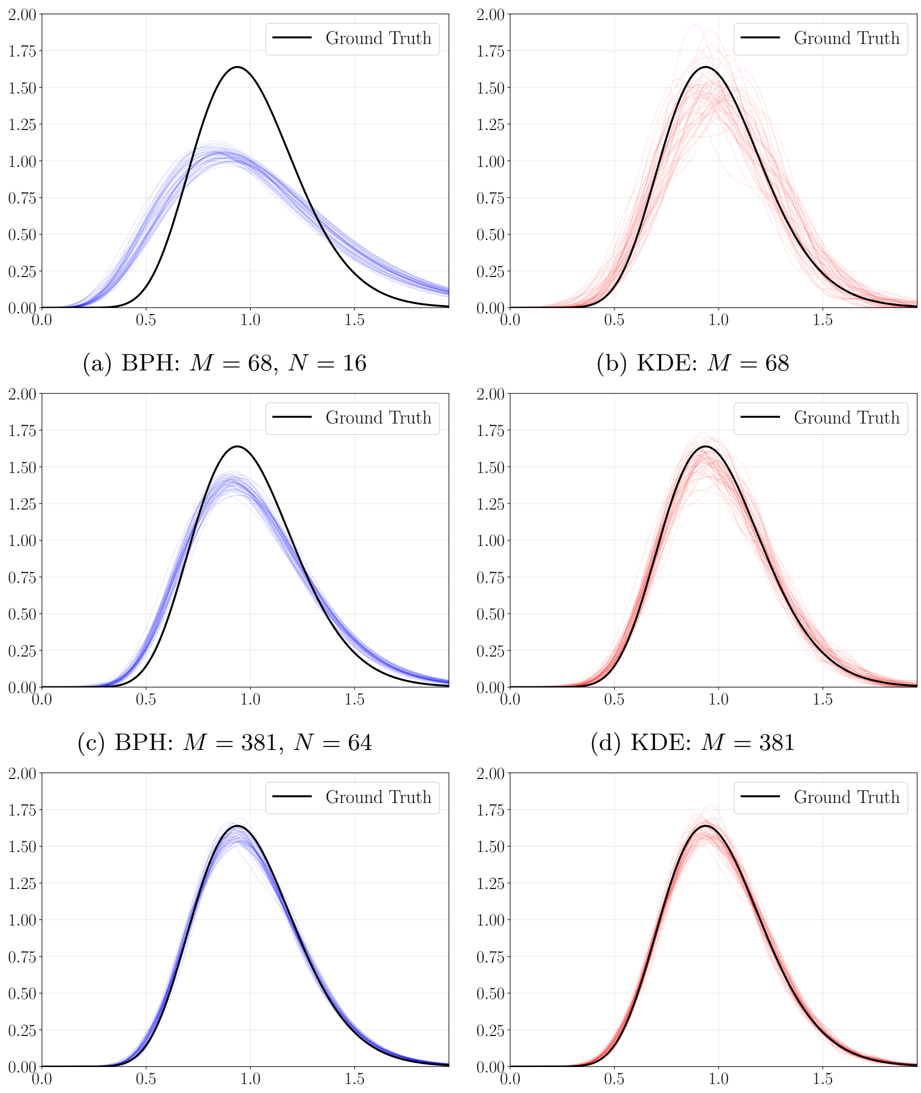


(f) KDE: $M = 1938$

Fig. 3: BPH approximants with order N (a,c,e) and KDE approximants (b,d,f) for 50 different samples of size M of an Erlang PDF with shape and rate $\boldsymbol{K = 4}$.

for $K \in \{2, 4\}$ and any value of M. For $K = 8$, BPHs achieve better accuracy than KDEs for sample size $M \geq 381$. For $K = 16$, BPHs achieve slightly larger KLD than KDEs for sample size $M = 1938$; for other sample sizes, the KLD is significantly larger for BPHs although still lower than $7 \cdot 10^{-2}$ for sample size $M \in \{163, 381, 866\}$.

Overall, these results yield key findings concerned with the intertwined effects of the sample size M and model complexity. As M grows, both BPHs and



(a) BPH: $M = 68$, $N = 16$

(b) KDE: $M = 68$

(c) BPH: $M = 381$, $N = 64$

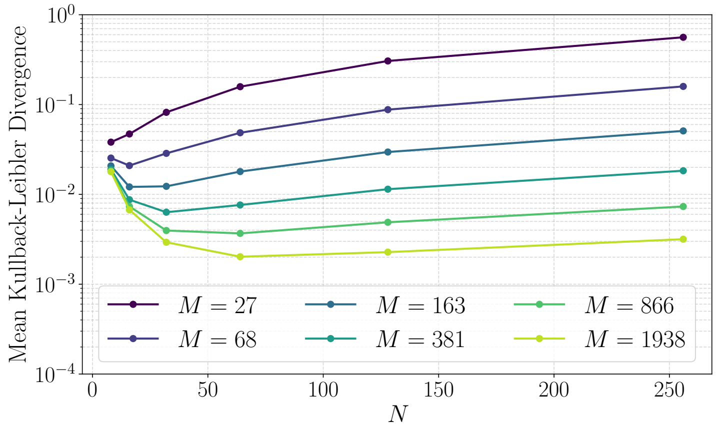
(d) KDE: $M = 381$

(e) BPH: $M = 1938$, $N = 256$

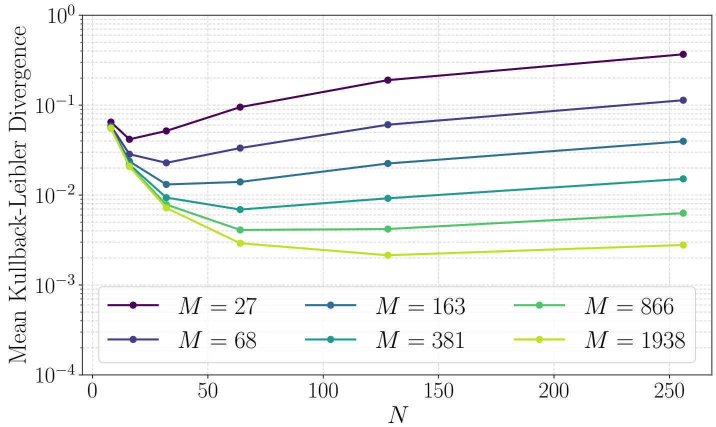
(f) KDE: $M = 1938$

Fig. 4: BPH approximants with order N (a,c,e) and KDE approximants (b,d,f) for 50 different samples of size M of an Erlang PDF with shape and rate $\boldsymbol{K = 16}$.

KDEs benefit from reduced sampling variability. While BPHs excel at capturing the inherent asymmetry of more right-skewed distributions ($K \leq 4$), the symmetric kernels of KDEs become more effective as the PDF becomes more concentrated ($K > 4$). Notably, as the coefficient of variation decreases, BPHs require a higher order and a significantly larger sample size, yet maintaining a better memory footprint (only $N \approx M/\log(M)$ BPH parameters are stored, instead of M sampled values) and allowing integration in Markov models.



(a) $K = 2$

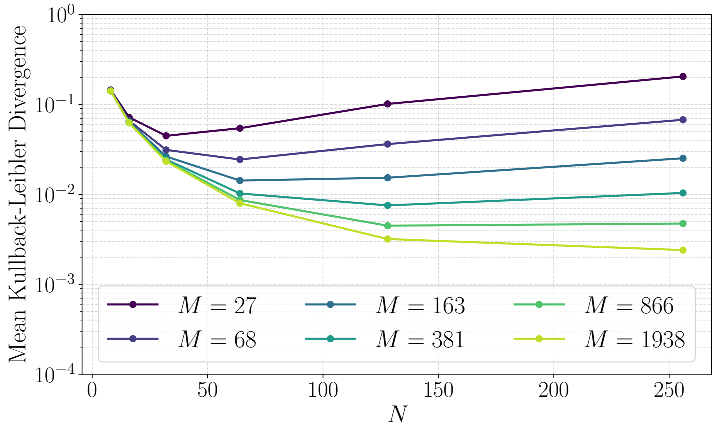


(b) $K = 4$

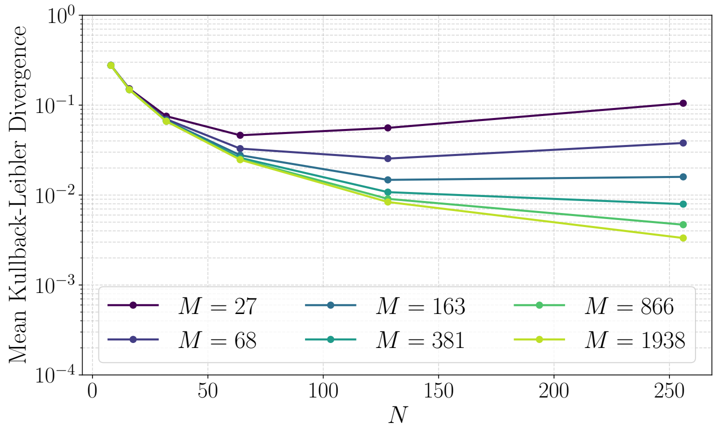
Fig. 5: Accuracy of BPHs (with order N) in approximating Erlang PDFs with shape and rate equal to K: mean value of the KLD with respect to the ground truth, over 10^5 random samples of size M.

4.2 All Combinations of Bernstein Order and Sample Size

Figs. 5 and 6 illustrate the behavior of the BPH approximant for $K \in \{2, 4, 8, 16\}$ and for all combinations of sample size $M \in \{27, 68, 163, 381, 866, 1938\}$ and Bernstein order $N \in \{8, 16, 32, 64, 128, 256\}$. Specifically, in the case of the Erlang PDF with shape and rate $K = 2$ shown in Fig. 5a, the mean value of the KLD increases when N exceeds the threshold $M/\log(M)$, pointing out the effectiveness, for BPH approximants, of the strategy proposed in [3] to select the order of BP approximants. For instance, for $M = 27$, corresponding to $N = 27/\log(27) \approx 8$, the mean KLD increases with N for any $N \geq 8$; for



(a) $K = 8$



(b) $K = 16$

Fig. 6: Accuracy of BPHs (with order N) in approximating Erlang PDFs with shape and rate equal to K: mean value of the KLD with respect to the ground truth, over 10^5 random samples of size M.

$M = 68$, corresponding to $N = 68/\log(68) \approx 16$, the mean KLD slightly decreases as N increases from 8 to 16, and then increases with N for any $N \geq 16$; and so on. Also note that, when the sample size M is insufficient to provide an accurate empirical estimate of the ground truth PDF, then increasing N does not yield a good accuracy (i.e., the mean KLD is much greater than 10^{-2}) and indeed it worsens accuracy for $N > M/\log(M)$, as already remarked. For instance, with a small sample size $M \in \{27, 68, 163\}$, the lowest value of the mean KLD, achieved for $N \approx M/\log(M)$, remains greater than 10^{-2}.

This behavior is evident also for $K \in \{4, 8, 16\}$, as illustrated in Fig. 5b, Fig. 6a, and Fig. 6b, respectively. Contrary to expectations, even with a small sample size $M \in \{27, 68\}$, for growing values of K we observe that accuracy improves as the Bernstein order N increases up to $N \in \{32, 64\}$, exceeding the $M/\log(M)$ threshold. This behavior may be attributed to the inherent complexity of the ground truth PDF, which increases with K, thus allowing BPHs to benefit from a higher order. Nevertheless, note that, with small sample sizes, the best achieved mean KLD remains significantly above 10^{-2}. Overall, these results show how BPH approximants effectively support the selection of an ideal Bernstein order N based on the available sample size M, as well as a tradeoff between N and the approximation accuracy.

5 Conclusions

We extended the consolidated method of [3] for the selection of the order of BP approximants of CDFs with bounded support, exploiting it to select the order of BPH approximants of PDFs with unbounded support, thus providing smooth estimators within the PH class while leveraging a rigorous framework to control the bias-variance tradeoff. By considering Erlang distributions with unit mean, our analysis characterized the interaction between the intrinsic approximation error due to using BPHs and the estimation error due to the sample size. We compared the accuracy of the approach with respect to Gaussian KDEs, again considering a set of Erlang PDFs with unit mean. The results showed the effectiveness of the BPH approximant in balancing the bias-variance tradeoff, achieving better or comparable performance with respect to KDEs while maintaining a low memory footprint. In fact, the BPH closed-form representation only requires the storage of a number of parameters equal to the Bernstein order.

Future work includes extending the evaluation to other PDFs and optimizing the placement of the points used to construct the approximant BPH to cover the portion of the support where most of the probability mass is located.

Acknowledgment

The authors would like to kindly thank Francesco Bettazzi for his help in performing the experimental evaluation and in preparing the replication package.

References

1. Asmussen, S., Nerman, O.: Fitting phase-type distributions via the EM algorithm. In: Proceedings of Symposium i Advent Statistik. pp. 335–346. Copenhagen (1991)
2. Azzalini, A.: A note on the estimation of a distribution function and quantiles by a kernel method. Biometrika **68**(1), 326–328 (1981). doi:10.2307/2335836
3. Babu, G.J., Canty, A.J., Chaubey, Y.P.: Application of Bernstein polynomials for smooth estimation of a distribution and density function. J. Stat. Plan. Inference **105**(2), 377–392 (2002). doi:10.1016/S0378-3758(01)00265-8
4. Bobbio, A., Cumani, A.: ML estimation of the parameters of a PH distribution in triangular canonical form. In: Computer Performance Evaluation. pp. 33–46. Elsevier (1992)

5. Bobbio, A., Horváth, A., Telek, M.: Matching three moments with minimal acyclic phase type distributions. Stochastic Models **21**, 303–326 (2005). doi:10.1081/STM-200056210
6. Brown, G., Ali, R.: Bias/variance is not the same as approximation/estimation. Trans. Mach. Learn. Res. **2024** (2024), `https://openreview.net/forum?id=4TnFbv16hK`
7. Carnevali, L., Horváth, A., Paolieri, M., Vicario, E.: Differentiable fit of empirical distributions by Bernstein phase types: Replication package (Apr 2026). doi:10.5281/zenodo.19751636
8. Feldman, A., Whitt, W.: Fitting mixtures of exponentials to long-tail distributions to analyze network performance models. Perform. Evaluation **31**, 245–279 (1998)
9. Härdle, W.: Smoothing Techniques. Springer Series in Statistics, Springer, New York (Dec 1991). doi:10.1007/978-1-4612-4432-5
10. Horváth, A., Telek, M.: Approximating heavy tailed behavior with phase-type distributions. In: Proc. of 3rd International Conference on Matrix-Analytic Methods in Stochastic models. pp. 191–214. Leuven, Belgium (June 2000)
11. Horváth, A., Horváth, I., Paolieri, M., Telek, M., Vicario, E.: Approximation of cumulative distribution functions by Bernstein phase-type distributions. In: Proceedings of QEST+FORMATS. Lecture Notes in Computer Science, vol. 14996, pp. 90–106. Springer (2024). doi:10.1007/978-3-031-68416-6_6
12. Horváth, A., Horváth, I., Paolieri, M., Telek, M., Vicario, E.: Approximation of cumulative distribution functions by Bernstein phase-type distributions. Perform. Evaluation **168**, 102480:1–102480:24 (2025). doi:10.1016/J.PEVA.2025.102480
13. Horváth, A., Telek, M.: Phase Type Distributions: Theory and Application. John Wiley & Sons (2024). doi:10.1002/9781119419808
14. Horváth, A., Vicario, E.: Construction of phase type distributions by Bernstein exponentials. In: Proceedings of EPEW. Lecture Notes in Computer Science, vol. 14231, pp. 201–215. Springer (2023). doi:10.1007/978-3-031-43185-2_14
15. Jones, M.: The performance of kernel density functions in kernel distribution function estimation. Statistics & Probability Letters **9**(2), 129–132 (1990). doi:10.1016/0167-7152(92)90006-Q
16. Latouche, G., Ramaswami, V.: Introduction to Matrix Analytic Methods in Stochastic Modeling. SIAM (1999)
17. Leblanc, A.: On estimating distribution functions using Bernstein polynomials. Ann. Inst. Stat. Math. **64**(5), 919–943 (2012). doi:10.1007/s10463-011-0339-4
18. Neuts, M.: Probability distributions of phase type. In: Liber Amicorum Prof. Emeritus H. Florin. pp. 173–206. University of Louvain (1975)
19. Neuts, M.: Matrix Geometric Solutions in Stochastic Models. Johns Hopkins University Press, Baltimore (1981)
20. Phillips, G.M.: Interpolation and Approximation by Polynomials. Springer (2003)
21. Riska, A., Diev, V., Smirni, E.: An EM-based technique for approximating long-tailed data sets with PH distributions. Performance Evaluation **55**(1-2), 147–164 (2004). doi:10.1016/S0166-5316(03)00101-9
22. Rivlin, T.J.: An Introduction to the Approximation of Functions. Courier Corporation (1981)
23. Telek, M., Horváth, G.: A minimal representation of Markov arrival processes and a moments matching method. Perform. Evaluation **64**(9-12), 1153–1168 (Aug 2007). doi:10.1016/j.peva.2007.06.001