

論 文

確率密度関数の形状に合わせた コックス型分布による確率分布の近似方法について

佐々木 幸恵*, 今井 博英*, 角山 正博**, 石井 郁夫*

Approximation Method for Probability Distribution Functions by Coxian Distribution
considering Feature of its Density Function

Yukie SASAKI, Hiroei IMAI, Masahiro TSUNOYAMA, Ikuo ISHII

要旨：この論文では、マルチメディアシステムなどのシステム解析に用いることを目的とした、コックス型分布による確率分布の近似方法を示す。著者らは文献[1]において、近似の対象とする確率分布を指数分布とアーラン分布の並列結合分布を用いて近似した後、得られた並列結合分布をコックス型分布に変換する、間接的な近似方法を提案した。本論文では、指数分布とアーラン分布の並列結合分布の求め方を改良した、凸型などを含んだ複雑な形状を示す確率密度関数をもつ分布の近似を可能とする方法を提案する。また、実際のインターネットトラフィックのパケット到着間隔を近似した例を示すとともに、既存の方法と比較することで、提案した近似方法の有用性を示す。

キーワード：システム解析、コックス型分布、アーラン分布、指数分布

Abstract : This paper shows the extension of the approximation method proposed in [1] using the Coxian distribution. The formerly proposed method can approximate distribution functions with the monotonic decreasing property. The method proposed in this paper uses the combination of hyper-exponential distribution and Erlang distributions, and enables to approximate non-monotonous distributions such as unimodal-type. We also present examples of the approximation for showing the usefulness of the method.

Keywords: System Evaluation, Coxian Distribution, Erlang Distribution, Exponential Distribution

1 はじめに

フォールトレントシステムやマルチメディアシステムでは、システムの信頼性の向上のために、実装前にあらかじめシステムの性能を解析、評価する必要がある。また、シミュレータやペトリネットなどのモデルを用いて解析、評価を行う場合、システム中のタスクの発生間隔、処理時間、制限時間などを統計モデルで近似する必要がある[5]。従来の大部分のシステム解析では、タスクの発生間隔、処理時間、制限時間などは指数分布に従うものと仮定されてきた。しかし、インターネットやLANにおけるタスクの発生間隔などが指数分布では近似できないことが報告されており、指

数分布に限定されないさまざまな分布を近似できる確率分布をシステム解析に取り入れることが必要となっている[6][7][8][9]。

本研究では、コックス型分布を用いて、タスクの統計的な性質が指数分布で表せないマルチメディアシステムなどの解析を可能にすることを目的としている。このため実測されたタスクの到着間隔や処理時間等のデータに基づいて確率分布の近似を行う。このコックス型分布は任意分布を近似可能であり、さらに指数分布の直列結合であるため、一般化確率ペトリネット等に組込んで解析することが容易であるという利点をもつ分布である。

コックス型分布を用いた確率分布の近似方法と

* 新潟大学大学院自然科学研究科, 〒950-2181 新潟市五十嵐2の町 8050 sasaki@cg.ie.niigata-u.ac.jp

** 新潟工科大学情報電子工学科, 〒945 1195 新潟県柏崎市藤橋 1719

してEMアルゴリズム[11][13]が提案されている。しかし、確率分布全体の尤度だけではなく、確率密度関数の長い裾や値の大きくなる凸部分などの特徴を考慮する必要がある[14][15]。

著者らは、文献[1]において(1)対象となる分布を指数分布の結合分布で近似し、(2)得られた結合分布をコックス型分布で表現する、という間接的な近似方法を提案した。これは、コックス型分布よりもパラメータと分布の形状の関係を把握しやすい指数分布とアーラン分布を組み合わせて用いることで、近似の対象となる分布の密度関数の形状に合わせた近似を行うことができるためである。また、指数分布の直列結合、並列結合さらにそれらの組合せなどの結合分布と等しい確率分布を表すことができるというコックス型分布の特長[16]を用いて、並列結合分布をコックス型分布に統一することで、解析モデルへの適用が容易になる。

しかし、文献[1]で提案した方法では、近似の対象分布が単調減少な確率密度関数をもつ分布に限定されているという問題点があった。そこで本論文では、上記の方法の(1)の手順を改良して、確率密度関数が単調減少ではなく凸型なども含んでいる複雑な分布にも適用できる近似方法を提案する。

本論文では、まず**2.**においてコックス型分布の定義を示す。つぎに**3.**に考案した近似方法の手順を示し、**4.**で実際のインターネットトラフィックのパケット到着間隔を近似した例を示し、既存のEMアルゴリズム[11]を用いて得られる近似結果と比較することで、本研究の近似方法の有用性を示す。最後に**5.**で全体をまとめ、今後の課題について述べる。

2 コックス型分布

コックス型分布は相型分布の一つであり、指数分布に従うノードを用いて表現すると、図1のような滞在時間で表されることが文献[16]において定義されている。

ここで L をノード数、 $a_0 = 1$ は、コックス型分布への到着確率、 a_l ($0 \leq a_l \leq 1, l = 1, 2, \dots, L-1$) をノード l からノード $l+1$ への到着確率、 b_l ($a_l + b_l = 1, l = 1, 2, \dots, L$) をノード l からの

退去確率、 $1/\lambda_l$ ($l = 1, 2, \dots, L$) をノード l の平均滞在時間とする。

客はノード1に到着すると、そのノードに平均 $1/\lambda_1$ の指数分布に従う時間滞在した後、確率 a_1 で次のノード2に進むか、確率 b_1 で網から退去する。最後にノード L での滞在を終えると網から退去する。この網全体の滞在時間の確率分布がコックス型分布で表され、確率密度関数のラプラス変換式 $\mathcal{F}_c(s)$ は以下のように表される。ここで $A_l = a_0 \cdot a_1 \cdots a_{l-1}$ とする。

$$\mathcal{F}_c(s) = \sum_{l=1}^L A_l b_l \prod_{i=1}^l \frac{\lambda_i}{\lambda_i + s} \quad (1)$$

このコックス型分布は、図1に示したように指数分布に従うノードの直列結合となっているため、容易に解析モデルに用いることができる。

また、コックス型分布は任意の累積分布関数を精度よく近似できることが証明されている[16]。しかし文献[16]の近似方法はノード数を無限大としているため、実際の解析モデルに用いることは困難である。このため、解析に用いることを前提とした近似方法を提案する。

3 近似方法

3.1 概要

提案する近似方法は、近似対象とする分布を指数分布やアーラン分布の並列結合で近似した後、その結合分布をコックス型分布に変換する方法[1]の拡張である。

多くの確率密度関数は、単調に減少する区間と一旦増加した後減少する凸型の区間に分けられる。これらをそれぞれ、確率密度関数が単調に減少する指数分布と凸型の確率密度関数を示すアーラン分布を用いて近似することで、近似対象分布の形状の特徴を考慮して精度よく近似することができる。

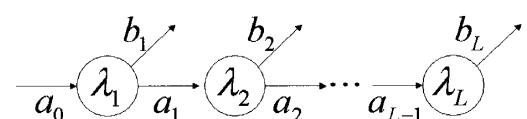


図1: コックス型分布

3.2 諸定義

本論文で提案する近似方法では、まず実測されたデータに基づいて構成された頻度分布に対して、移動平均法を用いてスムージングを行う。ここで用いたデータは標本数が十分多いため、この頻度分布に基づく確率分布を用いて、解析を行う上での母集団の確率分布を近似する。この結果得られた分布を対象分布とよび、累積分布関数を $\hat{F}(x)$ 、確率密度関数を $\hat{f}(x)$ と表記する。これら対象分布 $\hat{F}(x), \hat{f}(x)$ の値は、 K 個のサンプル点 $x_k (k = 0, \dots, K-1, x_0 = 0)$ における値で表される。また、対象分布の確率密度関数 $\hat{f}(x)$ から裾の部分を取り除いた確率密度関数を $f_{rt}(x) (rt : RemoveTail)$ 、単調に減少する区間を取り除いて凸型区間だけとした確率密度関数を $f_{um}(x) (um : UniModal)$ と表記する。これら $f_{rt}(x), f_{um}(x)$ の値は、近似の過程で求める。

これに対して、近似方法を用いて求められる分布を近似分布とよび、累積分布関数を $H(x)$ 、確率密度関数を $h(x)$ と表記する。さらに近似の過程で用いる推定の分布関数 $H(x)$ や推定のパラメータ λ や $p_{i,j}$ は、 $\hat{H}(x)$ 、 $\hat{\lambda}$ 、 $\hat{p}_{i,j}$ のように[^]（ハット）をつけて区別する。

つぎに、補累積分布関数（ccdf:Complementary cummulative distribution function）を次のように定義する。

定義 1 $F^c(x)$ を累積分布関数 $F(x)$ の補累積分布関数とよび、次式で表す。

$$F^c(x) = 1 - F(x) \quad (2)$$

つぎに、区間 $[x_i, x_j]$ の平均誤差を次のように定義する。

定義 2 $RE(x_i, x_j)$ を対象分布の累積分布関数 $\hat{F}(x)$ と近似分布の累積分布関数 $H(x)$ の区間 $[x_i, x_j]$ における平均誤差とよび、次式で表す。

$$RE(x_i, x_j) = \frac{1}{x_j - x_i} \int_{x_i}^{x_j} |\hat{F}(x) - H(x)| dx \quad (3)$$

さらに、指数分布の並列結合を以下のように定義する。

定義 3 指数分布の並列結合の確率密度関数 $h_0(x)$ は、 n_0 個の重みづけられた指数分布 $h_{0,i}(x) (i = 1, \dots, n_0)$ の和である。

$$h_0(x) = \sum_{i=1}^{n_0} h_{0,i}(x) \quad (4)$$

ここで $h_{0,i}(x)$ は、パラメータ $\lambda_{0,i}$ の指数分布と選択確率 $p_{0,i}$ との積であり、密度関数は下式のようになる。

$$h_{0,i}(x) = p_{0,i} \lambda_{0,i} \exp(-\lambda_{0,i} x) \quad (5)$$

最後に、近似手順で用いる重みづけられたアーラン分布を以下のように定義する。

定義 4 重みづけられたアーラン分布 $h_m(x)$ はパラメータ λ_m をもつ n_m 個の指数分布の直列結合と選択確率 p_m の積とする。確率密度関数 $h_m(x)$ は下式で表される。

$$h_m(x) = p_m \frac{\lambda_m^{n_m} x^{n_m-1} \exp(-\lambda_m x)}{(n_m - 1)!} \quad (6)$$

3.3 近似手順

まず、実測されたデータに基づいて構成された頻度分布に対して、移動平均法等を用いてスムージングを行う。次に、この頻度分布から得られた近似対象分布を、確率密度関数 $\hat{f}(x)$ が単調に減少する区間 P_0 と凸型の傾きを示す M 個の区間

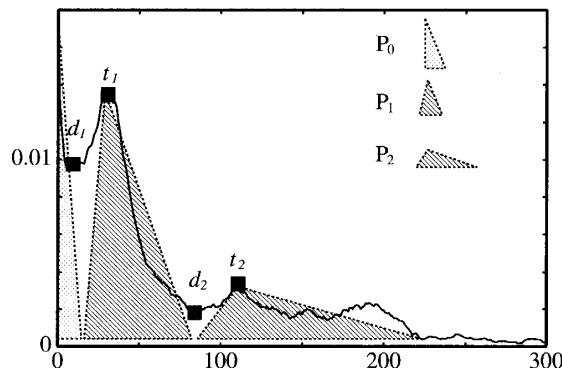


図 2: 対象分布の密度関数の分割例

$P_m (m = 1, 2, \dots, M)$ に分割する。図 2 に単調減少区間 P_0 と二つの凸型区間 P_1, P_2 に分割した例を示す。図 2 の近似対象分布は、団体 OSU の 2000 年に集計された ftp のパケット到着間隔の確率分布である。

つぎに、対象分布の確率密度関数 $\hat{f}(x)$ の単調減少区間 P_0 を近似する指数分布の並列結合 $h_0(x)$ を求める。その後、対象分布 $\hat{f}(x)$ から $h_0(x)$ の値を引いて凸型区間だけの確率密度関数 $f_{\text{sum}}(x)$ を求め、この $f_{\text{sum}}(x)$ の区間 $P_m (m = 1, 2, \dots, M)$ をそれぞれ近似するアーラン分布 $h_m(x)$ を求める。以上の手順で求めた $h_m(x)$ の和 $\sum_{m=0}^M h_m(x)$ が近似分布の確率密度関数 $h(x)$ となる。図 3 に指数分布の並列結合 $h_0(x)$ とアーラン分布 $h_1(x), h_2(x)$ 、そしてこれらの並列結合である $h(x)$ の例を示す。

さらに、 ε を極めて小さい正の実数、 x_ε を対象分布の補累積分布関数 $\hat{F}^c(x)$ が $\hat{F}^c(x_\varepsilon) = \varepsilon$ を満たす x の値と定義して、近似手順の詳細を以下に示す。

3.3.1 近似対象分布の分割

まず、近似対象分布を確率密度関数が単調に減少する区間と、 M 個の凸型の区間に分割する手順の詳細を示す。

近似対象分布の確率密度関数 $\hat{f}(x)$ 上の $x = 0$ から m 番目の極小点の x の値を m 番目の境界点 d_m とする。このとき、近似の対象とする極小点が増加すると、近似分布のノード数が増加し、解析するシステムのモデルに組み込むことができなくなる恐れがある。そこでシステムのモデルに許されるノード数に応じて、直近の極大点との差が大きい極小点から順に境界点としていく。ここ

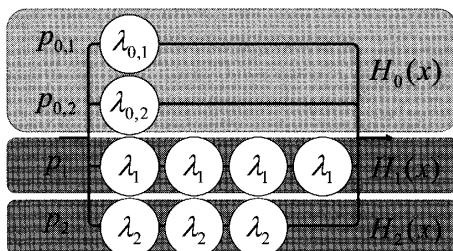


図 3: 並列結合分布の例

で $d_{M+1} = x_{K-1}$ として、 $x = 0$ 付近の $\hat{f}(x)$ の傾きが正である場合は $d_1 = 0$ とする。図 2 に境界点 d_1, d_2 の例を示す。

つぎに、区間 $[0, d_1]$ を単調減少区間 P_0 、区間 $[d_m, d_{m+1}]$ を m 番目の凸型区間 P_m として、近似対象分布を分割する。また P_m 内の極大点の x の値を頂点 t_m とする。図 2 に頂点 t_1, t_2 の例を示す。

最後に、各区間 $P_m (m = 0, 1, \dots, M)$ を近似する分布 $h_m(x)$ の最大ノード数 N_m と、目標とする平均誤差 TE_m を与える。ノード数に上限を設ける理由は、この近似方法が実際のシステム解析に用いることを前提としており、近似分布のノード数が多い場合、解析モデルの状態数が増加して解析が困難になるためである。

3.3.2 単調減少区間の近似

つぎに、単調減少区間 P_0 を近似する n_0 個の指数分布の並列結合 $h_0(x)$ を求める。

指数分布の並列結合は、ひとつの指数分布だけでは近似できない裾の長い分布を近似することができる。

このため、近似対象分布の裾が長い場合は $h_0(x)$ をノード数 $n_0 > 1$ の並列結合として、 $x = 0$ 付近を $h_{0,1}(x)$ 、裾の部分を $h_{0,i}(x) (i = 2, \dots, n_0)$ で近似する。他方で、近似対象分布の裾が短い場合は $n_0 = 1$ 、つまり $h_0(x)$ を 1 つの指数分布とする。

$h_0(x)$ のパラメータは、 n_0 個の指数分布 $h_{0,i}(x)$ に分けて求める。求め方は大きく以下の 3 つの手順に分けられる。

1. ノード数 n_0 の決定

対象分布の補累積分布関数 $\hat{F}^c(x)$ の裾の長さを判定して、指数分布の並列結合 $h_0(x)$ を 1 つの指数分布とするべきか、 $n_0 > 1$ の並列結合とするべきかを決定する。

2. 裈部分の近似

上記の手順で、対象分布の裾が長く $h_0(x)$ を $n_0 > 1$ の並列結合とした場合、裾部分を近似する指数分布の並列結合 $h_{0,2}(x) + \dots + h_{0,n_0}(x)$ を求める。

3. 指数分布 $h_{0,1}(x)$ の算出

区間 P_0 を近似する指数分布 $h_{0,1}(x)$ を求め る。

以下にこれら 3 つの手順を手順 1 から 3 まで詳しく述べる。

まず、指数分布の並列結合 $h_0(x)$ を 1 つの指数分布、つまり $n_0 = 1$ とするべきか、より裾の長い並列結合、つまり $n_0 > 1$ とするべきかを、対象分布の補累積分布関数 $\hat{F}^c(x)$ の裾の長さから判定する。

これは、 $n_0 = 1$ と仮定した指数分布の補累積分布関数 $H_0^c(x)(= H_{0,1}^c(x))$ や、凸型区間を近似するアーラン分布の補累積分布関数 $H_m^c(x)(m = 1, \dots, M)$ よりも、対象分布の補累積分布関数 $\hat{F}^c(x)$ の裾が長い場合、 $h_0(x)$ をより裾の長いノード数 $n_0 > 1$ の並列結合として裾部分の近似を行うことで、近似精度を向上させるためである。

詳細な手順を以下に示す。

手順 1 ノード数 n_0 の決定

1. $n_0 = 1$ と仮定したときの指数分布の推定の補累積分布関数 $\hat{H}_0^c(x)(= \hat{H}_{0,1}^c(x))$ のパラメータを求める。

推定パラメータ $\hat{\lambda}_{0,1}, \hat{p}_{0,1}$ は、区間 P_0 で最小二乗法用いて算出する。ただし、区間 $P_0 = [0, 0]$ である場合は、以下の値とする。

$$\hat{\lambda}_{0,1} = \hat{f}(0), \quad \hat{p}_{0,1} = 1 \quad (7)$$

2. 凸型区間 $P_m(m = 1, \dots, M)$ を近似するアーラン分布の推定の補累積分布関数 $\hat{H}_m^c(x)$ のパラメータを求める。

ノード数を $\hat{n}_m = N_m$ と仮定して、推定パラメータ $\hat{\lambda}_m, \hat{p}_m$ を、確率密度関数 $\hat{h}_m(x)$ の頂点の値が凸型区間 P_m の頂点 $\hat{f}(t_m)$ に一致するように下式より求める。

$$\begin{aligned} \hat{\lambda}_m &= (\hat{n}_m - 1)/t_m \\ \hat{p}_m &= \frac{\hat{f}(t_m)}{\hat{\lambda}_m^{\hat{n}_m} t_m^{\hat{n}_m-1} \exp(-\hat{\lambda}_m t_m) / (\hat{n}_m - 1)!} \end{aligned} \quad (8)$$

3. 指数分布 $\hat{H}_0^c(x)$ とアーラン分布 $\hat{H}_m^c(x)(m = 1, \dots, M)$ によって近似する範囲の上限を x_α と表し、下式のように求める。

$$x_\alpha = \max(x_{\varepsilon,m}) \quad (m = 0, \dots, M) \quad (9)$$

ここで $x_{\varepsilon,m}$ は以下の条件を満たす x の値である。

$$\hat{H}_m^c(x_{\varepsilon,m}) = \varepsilon \quad (10)$$

4. x_α と x_ε の値を比較して、対象分布の裾が長いか否かを判定し、 $h_0(x)$ のノード数を $n_0 > 1$ とするか否かを決定する。

- $x_\alpha < x_\varepsilon$ のときは、対象分布の補累積分布関数 $\hat{F}^c(x)$ の裾が長く、指数分布 $H_{0,1}^c(x)$ やアーラン分布 $H_m^c(x)$ では近似できないため、 $n_0 > 1$ として手順 2 へと進む。
- $x_\alpha \geq x_\varepsilon$ のときは、対象分布の補累積分布関数 $\hat{F}^c(x)$ の裾が短いため、 $n_0 = 1$ として手順 3 に進む。

上記の手順において対象分布の補累積分布関数 $\hat{F}^c(x)$ の裾を長いと判定し、 $h_0(x)$ のノード数を $n_0 \geq 2$ とした場合は、つぎに裾の部分を近似する 1 つ以上の指数分布の並列結合 $h_{0,2}(x) + \dots + h_{0,n_0}(x)$ を求める。

ここでまず補累積分布関数 $\hat{F}_i^c(x)$ を定義する。

定義 5 補累積分布関数 $\hat{F}_i^c(x)(i = 2, \dots, n_0)$ を対象分布の補累積分布関数 $\hat{F}^c(x)$ から指数分布の補累積分布関数 $H_{0,j}^c(x)(j = 2, \dots, i-1)$ の和を引いた値と定義して、下式のように求める。

$$\hat{F}_i^c(x) = \hat{F}^c(x) - \sum_{j=2}^{i-1} H_{0,j}^c(x) \quad (11)$$

ここで $\hat{F}_2^c(x) = \hat{F}^c(x)$ であり、 $\hat{F}_{i+1}^c(x)$ は

$$\hat{F}_{i+1}^c = \hat{F}_i^c(x) - H_{0,i}^c(x) \quad (12)$$

となり、 $\hat{F}_i^c(x)$ よりも裾の短い分布となる。

指数分布 $h_{0,i}(x)$ のパラメータを、上記で定義した補累積分布関数 $\hat{F}_i^c(x)$ の裾上の 2 点において、 $\hat{F}_i^c(x)$ の値と補累積分布関数 $H_{0,i}^c(x)$ の値が等しくなるように算出する。その後、式(12)から補累積分布関数 $\hat{F}_{i+1}^c(x)$ を求め、この $\hat{F}_{i+1}^c(x)$ に対して同様の手順で指数分布 $h_{i+1}(x)$ を求める。上記の手順を繰り返すことで、対象分布 $\hat{F}^c(x)$ の裾全体を近似する指数分布の並列結合 $h_{0,2}(x) + \dots + h_{0,n_0}(x)$ を求める。

また、近似精度は指数分布 $h_{0,i}(x)$ のパラメータを算出するときに基準とする 2 点（基準点とよぶ）の値に大きく依存する。このため、基準点の候補の点 x_k （候補点とよぶ）を複数用意し、各候補点から候補となるパラメータ $\lambda_{0,i,k}, p_{0,i,k}$ と、この候補パラメータの補累積分布関数 $H_{0,i,k}^c(x)$ と $\hat{F}_i^c(x)$ との平均誤差を求める。誤差が最小となった候補パラメータ $\lambda_{0,i,k}, p_{0,i,k}$ を真のパラメータ $\lambda_{0,i}, p_{0,i}$ とする。

詳細な手順を以下に示す。

手順 2 裾部分の近似

1. $h_0(x)$ のノード数を $n_0 = 2$ とする。
2. 2 つの基準点のうち、 $x = 0$ 側の基準点の候補点の範囲 C_{0,n_0} を以下のように決定する。

$$C_{0,n_0} = [x_\alpha, x_\alpha + \frac{N_0 - n_0}{N_0 - n_0 + 1} (x_\varepsilon - x_\alpha)] \quad (13)$$

3. 候補範囲 C_{0,n_0} 内のすべての候補点 x_k について、候補パラメータ $\lambda_{0,n_0,k}, p_{0,n_0,k}$ を基準点 x_ε と x_k で補累積分布関数 $H_{0,n_0,k}^c(x)$ の値が $\hat{F}_{n_0}^c(x)$ と等しくなるように下式より求め、

$$\begin{aligned} \lambda_{0,n_0,k} &= \frac{\ln(\hat{F}_{n_0}^c(x_k)/\hat{F}_{n_0}^c(x_\varepsilon))}{x_\varepsilon - x_k} \quad (14) \\ p_{0,n_0,k} &= \hat{F}_{n_0}^c(x_k)/\exp(-\lambda_{0,n_0,k} x_k) \end{aligned}$$

得られた候補分布 $H_{0,n_0,k}^c(x)$ と $\hat{F}_{n_0}^c(x)$ の区間 $[x_\alpha, x_{K-1}]$ の平均誤差 $RE_k(x_\alpha, x_{K-1})$ を求める。

4. ステップ 3 で求めた平均誤差 $RE_k(x_\alpha, x_{K-1})$ の値が最小となった候補パラメータ $\lambda_{0,n_0,k}, p_{0,n_0,k}$ を真のパラメータ $\lambda_{0,n_0}, p_{0,n_0}$ に決定する。
5. 式(12)からより裾の短い補累積分布関数 $\hat{F}_{n_0+1}^c(x)$ の値を求め、さらに $\hat{F}_{n_0+1}^c(x)$ に対する x_ε を求める。
6. 平均誤差が目標とする平均誤差を満たしている、つまり $\min_k(RE_k(x_\alpha, x_{max})) \leq TE_0$ 、もしくは最大ノード数に達している、つまり $n_0 = N_0$ の場合は、つぎの手順 3 へと進む。それ以外の場合はノード数 n_0 を 1 つ増やして、ステップ 2 へと戻る。

上記の手順で求めた指数分布 $h_{0,i}(x)(i = 2, \dots, n_0)$ の和を、裾部分を近似する指数分布の並列結合とする。さらに、対象分布の確率密度関数 $\hat{f}(x)$ から求めた並列結合 $\sum_{i=2}^{n_0} h_{0,i}(x)$ の値を引いた結果を、対象分布から裾部分を取り除いた確率密度関数 $f_{rt}(x)$ とする。

つぎに、対象分布の密度関数から裾部分を除いた $f_{rt}(x)$ の単調減少区間 P_0 を近似する指数分布 $h_{0,1}(x)$ のパラメータを算出する。まず二種類のパラメータ算出方法、(方法 a) (方法 b) のどちらが適切であるかを区間 P_0 の傾きから判定し、選択された方法を用いてパラメータを算出する。

手順 3 指数分布 $h_{0,1}(x)$ の算出

1. $h_{0,1}(x)$ の推定パラメータ $\hat{\lambda}_{0,1}, \hat{p}_{0,1}$ を、区間 P_0 内で最小二乗法を用いて求める。
2. $x_{\varepsilon,0}$ と x_ε を比較して、指数分布 $h_{0,1}(x)$ の適切な算出方法を選択し、パラメータを算出する。ここで、 $x_{\varepsilon,0}$ は $\hat{H}_{0,1}^c(x_{\varepsilon,0}) = \varepsilon$ を満たす x の値である。
 - $x_{\varepsilon,0} < x_\varepsilon$
指数分布 $\hat{h}_{0,1}(x)$ の裾が短いため、(方法 a) を用いてパラメータを算出する。

- $x_{\varepsilon,0} \geq x_{\varepsilon}$
- $\hat{h}_{0,1}(x)$ 部分の裾が長いため、(方法 b) を用いてパラメータを算出する。

上記のステップ3で用いる(方法a)(方法b)の詳細な手順を示す。

(方法a)は、指数分布 $h_{0,1}(x)$ のパラメータを単調減少区間 P_0 で最小二乗法を用いて算出する方法である。

区間 $P_0 = [0, d_1]$ 内の確率密度関数 $f_{\text{rt}}(x)$ に対して最小二乗法を適用した場合、この後の手順で凸型区間 $P_m (m = 1, \dots, M)$ を近似するアーラン分布 $h_m(x)$ の値が上乗せされて、近似精度が悪化してしまう。このため、あらかじめアーラン分布の推定の確率密度関数 $\hat{h}_m(x)$ の値を $f_{\text{rt}}(x)$ の値から差し引いた値 $\hat{f}_{\text{rt}}(x)$ に最小二乗法を適用する。図4に(方法a)を用いて求めた指数分布 $h_0(x)$ の例を示す。

(方法a) 区間 P_0 で最小二乗法を用いる

- 凸型区間 $P_m (m = 1, \dots, M)$ を近似するアーラン分布の推定の確率密度関数 $\hat{h}_m(x)$ を求める。

ノード数を $\hat{n}_m = N_m$ と仮定して、推定パラメータ $\hat{\lambda}_m, \hat{p}_m$ を確率密度関数 $\hat{h}_m(x)$ の頂点の値が凸型区間 P_m の頂点 $f_{\text{rt}}(t_m)$ に一致

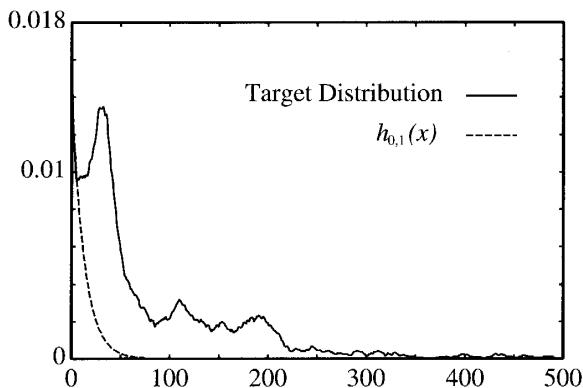


図4: 補の短い $h_0(x)$ の例

するように下式より求める。

$$\begin{aligned}\hat{\lambda}_m &= (\hat{n}_m - 1)/t_m \\ \hat{p}_m &= \frac{f_{\text{rt}}(t_m)}{\hat{\lambda}_m^{\hat{n}_m} t_m^{\hat{n}_m - 1} \exp(-\hat{\lambda}_m t_m) / (\hat{n}_m - 1)!}\end{aligned}\quad (15)$$

- 確率密度関数 $f_{\text{rt}}(x)$ の値からステップ1で求めたアーラン分布の推定分布 $\hat{h}_m(x)$ の値を引いた値を $\hat{f}_{\text{rt}}(x)$ とする。

$$\hat{f}_{\text{rt}}(x) = f_{\text{rt}}(x) - \sum_{m=1}^M \hat{h}_m(x) \quad (16)$$

- 単調減少区間 $P_0 = [0, d_1]$ 内の $\hat{f}_{\text{rt}}(x)$ に対して最小二乗法を適用し、指数分布 $h_{0,1}(x)$ のパラメータ $\lambda_{0,1}, p_{0,1}$ を算出する。

つぎに、(方法b)について述べる。(方法b)は、単調減少区間 P_0 を近似する $h_{0,1}(x)$ の裾が長い場合に用いる方法で、 $x = 0$ と裾上の1点、合わせて2点を基準点としてパラメータを算出する。

裾上の基準点の決定方法は、手順2と同様である。つまり、基準点の候補点 x_k を複数用意し、各候補点から候補となるパラメータ $\lambda_{0,1,k}, p_{0,1,k}$ と、この候補パラメータの累積分布関数 $H_{0,1,k}(x)$ と $F_{\text{rt}}(x)$ との平均誤差を求め、誤差が最小となつた候補パラメータ $\lambda_{0,1,k}, p_{0,1,k}$ を真のパラメータ $\lambda_{0,1}, p_{0,1}$ とする。

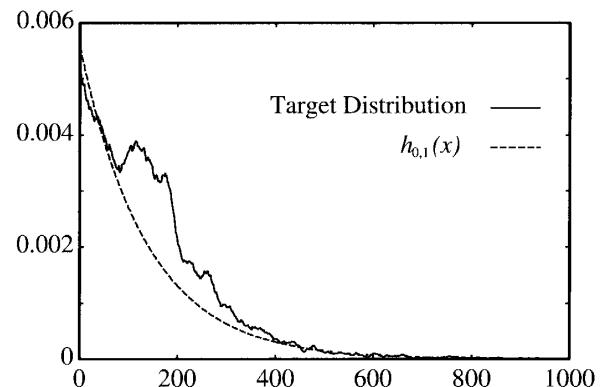


図5: 補の長い $h_0(x)$ の例

このとき、候補パラメータ $\lambda_{0,1,k}, p_{0,1,k}$ の累積分布関数 $H_{0,1,k}(x)$ と $F_{\text{rt}}(x)$ の誤差だけで判定した場合、後に凸型区間 $P_m(m = 1, \dots, M)$ を近似する $H_m(x)$ の値が上乗せされてしまい精度が悪化する。このため、評価基準に用いる誤差は、候補の累積分布関数 $H_{0,1,k}(x)$ とアーラン分布の推定の累積分布関数 $\hat{H}_{m,k}(x)$ との和

$$H_{0,1,k}(x) + \sum_{m=1}^M \hat{H}_m(x)$$

と、対象分布 $F_{\text{rt}}(x)$ との誤差とする。図5に（方法 b）を用いて求めた指数分布 $h_0(x)$ の例を示す。

（方法 b） $x = 0$ および裾上の点を基準点とする

1. 候補となるパラメータ $\lambda_{0,1,k}, p_{0,1,k}$ を $x = 0$ と候補点 $x_k(\in C_1 = [x_{\varepsilon,M}, x_{\varepsilon}])$ において、確率密度関数の値が $f_{\text{rt}}(x)$ と等しくなるように下式より求める。

$$\lambda_{0,1,k} = \frac{\ln(f_{\text{rt}}(0)/f_{\text{rt}}(x_k))}{x_k} \quad (17)$$

$$p_{0,1,k} = f_{\text{rt}}(0)/\lambda_{0,1,k} \quad (18)$$

ここで $x_{\varepsilon,M}$ は最も裾側の凸型区間 P_M を近似するアーラン分布の推定の累積分布関数 $\hat{H}_M^c(x)$ の値が ε となる x である。

2. 候補の確率密度関数 $h_{0,1,k}(x)$ に対する、アーラン分布の推定確率密度関数 $h_m(x)(m = 1, \dots, M)$ のパラメータを求める。

$\hat{n}_{m,k} = N_m$ と仮定して、推定パラメータ $\hat{\lambda}_{m,k}, \hat{p}_{m,k}$ を、下式より算出する。

$$\hat{\lambda}_{m,k} = (\hat{n}_m - 1)/t_m \quad (19)$$

$$\hat{p}_{m,k} = \frac{f_{\text{rt}}(t_m) - h_{0,1}(t_m)}{\hat{\lambda}_{m,k}^{\hat{n}_m} t_m^{\hat{n}_m - 1} \exp(-\hat{\lambda}_{m,k} t_m) / (\hat{n}_m - 1)!}$$

3. ステップ1で求めた候補の指数分布 $H_{0,1,k}(x)$ とステップ2で求めたアーラン分布の $H_{m,k}(x)$ の和と、対象分布の裾部分を取り除いた累積分布関数 $F_{\text{rt}}(x)$ との平均誤差 $\text{RE}_k(0, d_1)$ が最小となった候補 $\lambda_{0,1,k}, p_{0,1,k}$ を真のパラメータ $\lambda_{0,1}, p_{0,1}$ とする。

以上の手順から、単調減少区間を近似する指数分布の並列結合 $h_0(x) = \sum_{i=1}^{n_0} h_{0,i}(x)$ が求まる。近似対象分布の確率密度関数 $\hat{f}(x)$ から $h_0(x)$ の値を引くと、凸型区間のみの確率密度関数 $f_{\text{um}}(x)$ が得られる。

$$f_{\text{um}} = \hat{f}(x) - h_0(x) \quad (20)$$

この $f_{\text{um}}(x)$ を近似するアーラン分布 $h_m(x)(m = 1, \dots, M)$ の求め方を次節で述べる。

3.3.3 凸型区間の近似

対象分布の凸型区間のみの確率密度関数 $f_{\text{um}}(x)$ の M 個の凸型区間 $P_m(m = 1, 2, \dots, M)$ をアーラン分布 $h_m(x)$ で近似する。ここで近似対象分布の補累積分布関数 $\hat{F}^c(x)$ から $H_0^c(x)$ の値を引いた補累積分布関数を $\hat{F}_{\text{um}}^c(x) (= \hat{F}^c(x) - H_0^c(x))$ とする。図6に確率密度関数 $f_{\text{um}}(x) (= \hat{f}(x) - h_0(x))$ の例を示す。

裾側の区間 P_M から P_1 まで降順に、凸型区間を近似するアーラン分布 $h_m(x)$ を求める。パラメータ λ_m, p_m の求め方は、手順2と同様に、複数の候補点に対するパラメータと平均誤差を求め、誤差が最小となる候補パラメータを真のパラメータとする。

また、 $h_m(x)$ のノード数 n_m は、平均誤差が目標とする平均誤差 TE_m 以下となるか、ノード数

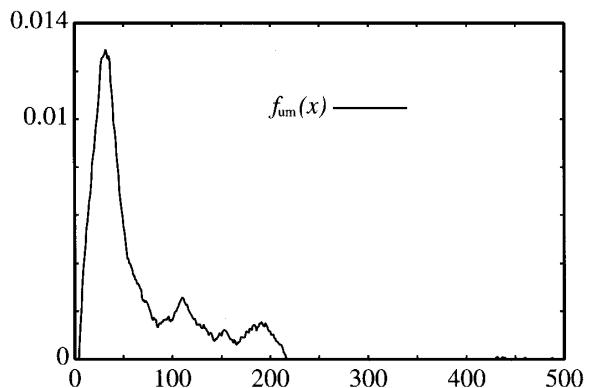


図 6: $f_{\text{um}}(x) = \hat{f}(x) - h_0(x)$ の例

n_m が最大ノード数 N_m に達するまで、一つずつ増やしていく。

手順 4 凸型部分の近似

- 裾側からアーラン分布 $h_m(x)$ を求める。

$$m = M$$

- $h_m(x)$ の初期ノード数を $n_m = 2$ とする。
- 候補範囲 $C_m = [d_m, d_{m+1}]$ 内のすべての候補点 x_k に対して、 x_k が頂点となるように下式から候補パラメータ $\lambda_{m,k} p_{m,k}$ を算出し、得られたパラメータに対する区間 $[d_m, x_{\max}]$ の平均誤差 $RE_k(d_m, x_{\max})$ を求める。

$$\lambda_{m,k} = (n_m - 1)/x_k \quad (21)$$

$$p_{m,k} = \begin{cases} \hat{F}_{\text{um}}^c(d_m) - \sum_{i=m+1}^M H_i^c(d_m) & (m > 1) \\ 1 - \sum_{i=0, i \neq 1}^M p_i & (m = 1) \end{cases} \quad (22)$$

- 平均誤差 $RE_k(d_m, x_{\max})$ が最小となった候補パラメータ $\lambda_{m,k}, p_{m,k}$ を真のパラメータ λ_m, p_m とする。
- 平均誤差が目標誤差以下となる、つまり $\min_k (RE_k(d_m, x_{\max})) \leq TE_m$ 、もしくはノード数が最大ノード数に達した、つまり $n_m = N_m$ の場合は、ステップ 6 へと進む。
それ以外の場合は、ノード数 n_m を一つ増やしてステップ 3 へ戻る。
- $m = 1$ の場合は、近似を終了する。

それ以外は $m \leftarrow m - 1$ として、ステップ 2 へもどる。

以上の手順により求めた指数分布の並列結合 $h_0(x)$ と M 個のアーラン分布 $h_m(x)$ ($m = 1, \dots, M$) の並列結合が、近似対象分布を近似する近似分布 $h(x)$ となる。

$$h(x) = \sum_{m=0}^M h_m(x) \quad (23)$$

3.4 コックス型分布への変換

近似分布 $h(x)$ と等しい確率分布を示すコックス型分布のパラメータを求める。

手順 5 コックス型分布のパラメータの求め方

- $h(x)$ の $L = \sum_{m=0}^M n_m$ 個のパラメータ、 $\lambda_{0,1}, \lambda_{0,2}, \dots, \lambda_{0,n_0}$ と λ_m を n_m 個ずつ ($m = 1, 2, \dots, M$) を降順に並べ、コックス型分布のノードのパラメータ λ_l ($l = 1, 2, \dots, L$) とする。
- ノード k までのラプラス変換式 $Y_k(s) = \prod_{i=1}^k \lambda_i/(s + \lambda_i)$ を、部分分数を用いて展開する。

$$Y_k(s) = \sum_{i=1}^k \frac{\beta_{k,i}}{(s + \lambda_i)^{n_i}} \quad (24)$$

ここで、 n_i はノード $1, 2, \dots, i$ で λ_i とパラメータ λ の値が等しいノードの数、 $\beta_{k,i}$ は係数である。

- 係数 $\beta_{k,i}$ を用いて、コックス型分布のパラメータ $A_l = (a_0 \cdot a_1 \cdots a_{l-1})$ を算出する。

$$A_l = \frac{p_l \lambda_l^{n_l} - \sum_{i=l+1}^L A_i (\beta_{i,l} - \beta_{i-1,l})}{\beta_{l,l}} \quad (25)$$

ここで、 p_l はパラメータ λ_l が n_l 個直列に結合した分布への選択確率である。

- A_l より、ノード l から $l+1$ への到着確率 a_l を求める。

$$a_l = A_{l+1}/A_l \quad (l = 1, 2, \dots, L-1) \quad (26)$$

4 近似例

本論文で示した近似方法を用いた近似例を示す。近似対象として、NLANR(National Laboratory for Applied Network Research) で公開されている実測データ [18] を集計した統計データを用いた。

近似例 1 は、団体 OSU の 2000 年に集計された ftp のパケット到着間隔の確率分布、近似例 2 は、

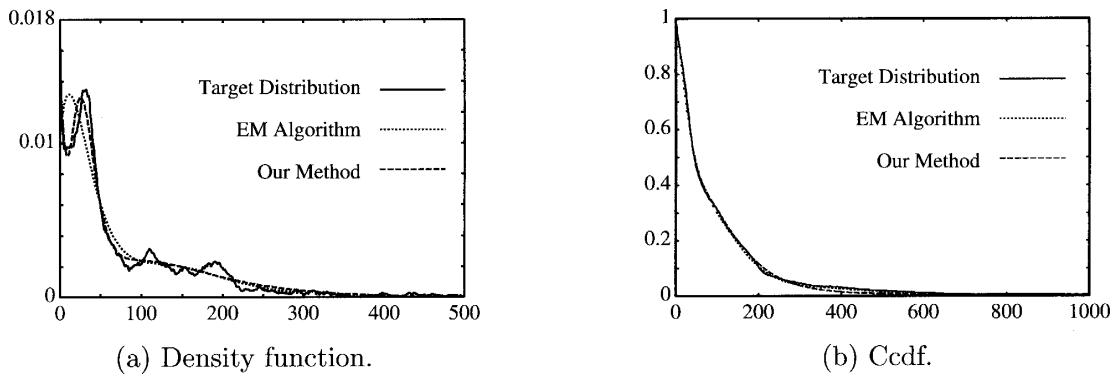


図 7: 近似例 1 の近似結果。

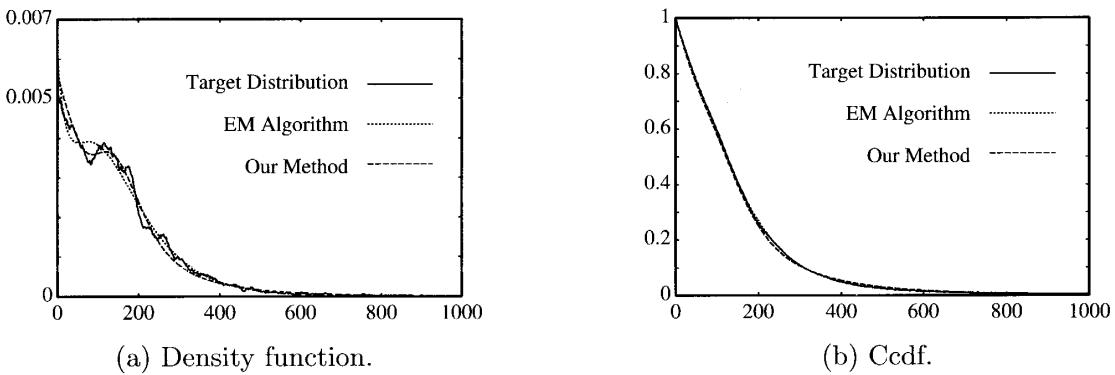


図 8: 近似例 2 の近似結果。

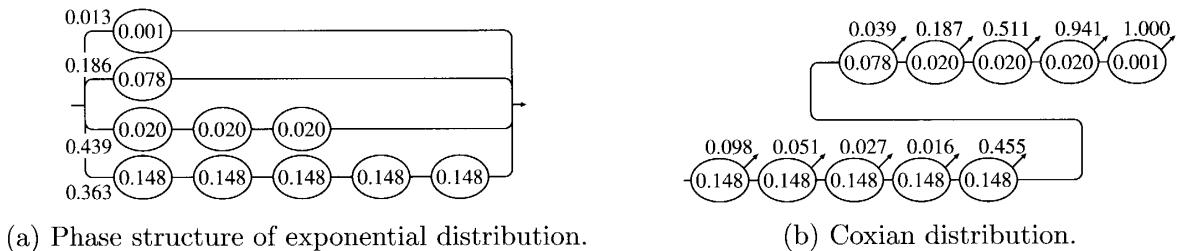


図 9: 近似例 1 の近似分布。

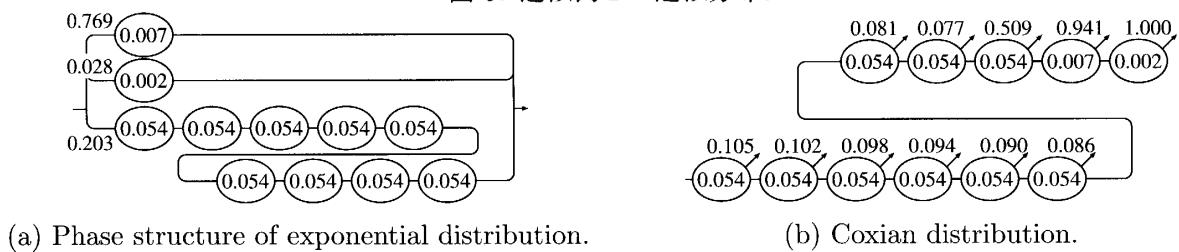


図 10: 近似例 2 の近似分布。

団体 APN の 1999 年に集計された telnet のパケットの到着間隔の確率分布の近似結果である。

近似対象分布と本論文で提案した方法で得た近似結果、さらに既存の EM アルゴリズムを用いて求めた近似結果を図 7,8 に示す。EM アルゴリズムでは、スマージング後の $\hat{F}(x)$ を対象分布として、提案法で得た近似結果と同じ位相、つまり同じ数の推定パラメータをもつコックス型分布により近似を行なった。なお、この計算には、[11] で使用されているプログラム EMphit [12] を使用した。さらに、近似の結果得られた指数分布の結合分布と、結合分布を変換したコックス型分布を図 9,10 に示し、表 1 に提案した方法と EM アルゴリズムの Kullback-Leibler 情報量、確率密度関数と ccdf の平均誤差を示す。

近似例 1 は、 $x = 30$ 周辺の確率密度関数の凸部分は高い精度で近似できたが、 $x = 200$ 周辺の複数の凸型部分を一つの凸型区間とみなして近似しているため、この部分の近似精度が悪くなってしまっている。EM アルゴリズムを用いて得られた近似結果と比較すると、補確率分布の平均誤差においてはわずかに提案した方法の方が誤差が小さい。さらに、Kullback-Leibler 情報量、確率密度関数の平均誤差では提案した方法の方が値が小さくなっている、図 7 から分かるように確率密度関数の凸型部分が精度よく近似できている。

また近似例 2 は、確率密度関数、補累積分布関数ともに近似精度の高い結果が得られている。EM アルゴリズムを用いて得た近似結果と比較すると、補累積分布関数に関しては EM アルゴリズムの方が誤差が小さく、Kullback-Leibler 情報量に関しても EM アルゴリズムの方が値が小さい。しかし確率密度関数に関しては提案した方法の方が誤差が小さいといえる。

以上の結果より、既存の EM アルゴリズムと比較して、同程度の Kullback-Leibler 情報量で、確率密度関数をより精度よく近似できているといえる。

5 まとめ

システム解析に用いることを目的としたコックス型分布による確率分布の近似方法を示した。近似対象とする分布の確率密度関数が単調減少を示

表 1: 近似結果

(a) 近似例 1

	EM algorithm	Our method
KL divergence	9.8185E-02	4.5938e-02
Error (Density)	5.8992E-05	4.0947e-05
Error (Ccdf)	2.3323E-03	1.5139e-03

(b) 近似例 2

	EM algorithm	Our method
KL divergence	6.3643e-03	7.7581e-03
Error (Density)	2.0365e-05	1.8685e-05
Error (Ccdf)	0.5703e-03	1.4344e-03

さない場合にも適用できるように、既に提案したコックス型分布を用いた確率分布の近似方法を改良した近似方法の詳細な手順を示した。

また近似例を示すことで、確率密度関数に凸型部分を含んだ分布も高い精度で近似できることを示した。

今後の課題として、本手法を本論文中に示した例以外の種々のシステムやトラフィックに適用して本手法を評価すると共に、選択確率がマイナスを示す結合分布も表現できるというコックス型分布の特性をいかして、より近似精度の高い方法を考案することが挙げられる。また、EM アルゴリズムなどの近似方法に対して、本方法におけるメモリ使用量や計算時間等を比較し定量的な評価を行うと共に、近似方法の実システムへの適用方法を検討する予定である。更に本手法を用いて得られたコックス型分布の各パラメータの値と、評価対象であるシステムの関係についても今後研究を進めて行く予定である。

参考文献

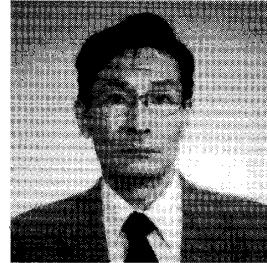
- [1] 佐々木幸恵、今井博英、角山正博、石井郁夫，“マルチメディアシステム解析のためのコックス型分布による確率分布の近似方法について,” 信学論 (D-I), vol.J85-D-I, no.9, pp.887-895, Sept. 2002.
- [2] Y. Sasaki, H. Imai, M. Tsunoyama and I. Ishii, “Approximation method for probability distribution functions using Cox distribution to evaluate multimedia sys-

- tems," Proc.2001 Pacific Rim International Symp. on Dependable Computing, pp.333-340, Seoul, Korea, Dec. 2001.
- [3] 佐々木幸恵, 今井博英, 角山正博, 石井郁夫, “リアルタイムシステム解析のためのコックス型分布の確率分布の近似方法について,” 日本信頼性学会第9回研究発表会, REAJ誌, vol.23, no.4, pp.401-402, May. 2001.
- [4] Y. Sasaki, H. Imai, M. Tsunoyama and I. Ishii, “Approximation method for probability distribution functions by Coxian distribution,” AIWARM 2004, Hiroshima, Japan, Aug. 2004.
- [5] A. Thummler, P. Buchholz and M. Telek, “A Novel Approach for Fitting Probability Distributions to Real Trace Data with the EM Algorithm,” Proc. 2005 Int. Conf. on Dependable Systems and Networks, pp.712-721, IEEE Computer Society Press, 2005.
- [6] W. Willinger, M. S. Taqqu, R. Sherman and D. V. Wilson, “Self similarity through high variability:statistical analysis on Ethernet LAN traffic at the source level,” IEEE/ACM Trans. Networking, vol.5, no.1, pp.71-86, Feb. 1997.
- [7] W. T. Marshall and S. P. Morgan, “Statistics of mixed data traffic on a local area network,” Computer Networks ISDN Systems, vol.10, no.3-4, pp.185-195, Oct./Nov. 1985.
- [8] R. Caceres, P. B. Danzig, S. Jamin and D. J. Mitzel, “Characteristics of wide-area TCP/IP conversations”, Computer Communication Review, vol.21, no.4, pp.101-112, Sept. 1991.
- [9] J. Beran, R. Sherman, M. S. Taqqu and W. Willinger, “Long-range dependence in variable-bit-rate video traffic,” IEEE Trans. Communications, vol.43, no.2-4, pp.1566-1579, Feb./Mar./Apr. 1995.
- [10] Haddad. S, Moreaux. P, Chiola. G, “Efficient Handling of Phase-type Distributions in Generalized Stochastic Petri Nets,” Proc. 18th Int. Conf. on Application and Theory of Petri Nets, Springer Verlag, LNCS 1248, pp.75-194, March 1997.
- [11] S. Asumussen, O. Nerman and M. Olsson, “Fitting phase-type distributions via the EM algorithm,” Scandinavian J. Statist., vol.23, pp.419-441, 1996.
- [12] EMph, <<http://home.imf.au.dk/asmus/pspapers.html>>
- [13] M. C. Heijden, “On the three moment approximation of a general distribution by a Coxian distribution,” Probability in the Engineering and Information Science, vol.2, pp.257-261, 1988.
- [14] K. J. Grinnemo and A. Brunstrom, “A Simulation-Based Performance Analysis of a TCP Extension for Best-Effort Multimedia Applications,” Proceedings of the 35th Annual Simulation Symposium, pp.327-336, San Diego, California, 2002.
- [15] M. S. Borella, “Source Models of Network Game Traffic,” Computer Communications, vol. 23, no. 4, pp.403 - 410, Feb. 15, 2000.
- [16] E. Gelénbe and I. Mitrani, Analysis and Synthesis of Computer Systems, Academic Press, London, 1988.
- [17] A. Feldmann and W. Whitt, “Fitting mixtures of exponentials to long-tail distributions to analyze network performance models,” Perform. Eval., vol.31, no.3-4, pp.245-279, Jan. 1998.
- [18] 阿多信吾, 村田正幸, 宮原秀夫, “高速レイヤ3スイッチングルータ設計のためのネットワークトラヒックの分析,” 信学技報(CQ98-35), pp.29-36, Sept. 1998.
 (ささき ゆきえ/いまい ひろえい/つのやま
 まさひろ/いしい いくお)



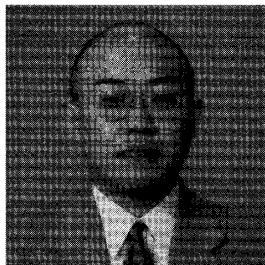
佐々木幸恵

平11新潟大・工・情報卒。現在、同大大学院博士後期課程在学中。日本信頼性学会会員



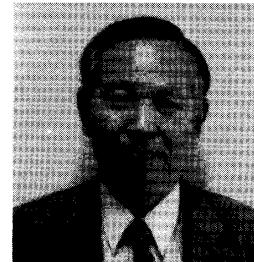
角山 正博

昭44新潟大・工・電子卒。同年東京芝浦電気(株)入社。昭46横河ヒューレットパッカード(株)入社。主に測定器の研究開発に従事。長岡技術科学大学教務職員、助手、長岡工業高等専門学校教授を経て、現在新潟工科大学教授。デペンドブルシステム、コンピュータネットワークの性能評価と高信頼化、リアルタイムシステム等の研究に従事。平2工博。日本信頼性学会、情報処理学会、IEEE等各会員。



今井 博英

平5新潟大・工・情報卒。平7同大大学院修士課程了、平10同大大学院自然科学研究科博士後期課程修了。平10同大大学院自然科学研究科助手、現在同大工学部情報工学科助手。ネットワークシステムの性能評価、共有仮想環境の研究に従事。平10工博。電子情報通信学会会員。



石井 郁夫

1963年新潟大・工・電気卒。同大・助手、助教授を経て、現在同大・大学院自然科学研究科教授。人工現実感、ロボットビジョン、画像処理、マルチメディア通信などの研究に従事。工博。情報処理学会、日本VR学会、日本ME学会各会員。

投稿受付：2005年3月22日

改 訂：2005年10月25日

再改訂：2006年1月30日

再々改訂：2006年2月2日