A First Course in Stochastic Network Calculus

Dipl.-Math. Michael Alexander Beck

January 28, 2016

Contents

Foreword

This *First Course in Stochastic Network Calculus* is a "First" course in two different perspectives. One can see it as a introductory course to Stochastic Network Calculus. On the other side it was my first course I have given about SNC to a few students in June 2012. It builds on the lecture *Performance Modelling in Distributed Systems [?]* at the University of Kaiserslautern. The concepts of stochastic network calclulus parallels those of deterministic network calculus. This is why I reference on the lecture of 2011 at several points to stress these connections. This document however is thought of a stand-alone course and hence a deep study of [?] is not necessary (but recommended).

This course contains a rather large probability primer to ensure the students can really grasp the expressions, which appear in stochastic network calculus. A student familiar with probability theory might skip this first chapter and delve directly into the stochastic network calculus. For each topic exercises are given, which can (and should) be used to strengthen the understanding of the presented definitions and theory.

This document got a major revision concerned with the way of tail-bounding service and is hence in its second version. However, it is also still in process and hopefully will evolve at some day into a fully grown course about Stochastic Network Calculus, providing a good overview over this exciting theory. Please provide feedback to `beck@cs.uni-kl.de`.

- Michael Beck

1 Basics of Probability

In this chapter we give a refresher about the basics of probability theory. We focus on material needed for constructing a stochastic network calculus. This refresher exceeds the probability primer of the lecture *Performance Modelling in Distributed Systems [?]* by far. The following definitions and results serve us as a foundation, to build our calculus upon. The material presented here (and much more) is standard in probability theory and can be found - much more detailed - in many introductions [?, ?, ?, ?].

1.1 Probability Spaces and Random Variables

We start our refresher with the very basic definitions needed to construct a probability space.

Definition 1.1. We call a non-empty set $\Omega \neq \emptyset$ *sample space*. A *σ-algebra* $\mathcal{A}$ on Ω is called *event space*. If $\omega \in A$ for some *event* $A \in \mathcal{A}$, we say that the outcome ω lies in the event A.

A σ-algebra (also called σ-field) of Ω is a collection of subsets of Ω, with the properties:

- $\Omega \in \mathcal{A}$
- $A \in \mathcal{A} \quad \Rightarrow \quad A^c \in \mathcal{A}$
- $A_1, A_2, A_3, \ldots \in \mathcal{A} \quad \Rightarrow \quad \{\bigcup_{i=1}^{\infty} A_i\} \in \mathcal{A}$

Example 1.2. As first example we consider a fair die. The possible outcomes of a die are the following:

$$\begin{aligned}\Omega = \{&\text{The die shows one pip, The die shows two pips,}\\ &\text{The die shows three pips , The die shows four pips,}\\ &\text{The die shows five pips, The die shows six pips}\}\end{aligned}$$

A simple σ-algebra would be just the *power set* of Ω, defined by:

$$\mathcal{A} = \text{Pot}(\Omega) = 2^{\Omega} = \mathcal{P}(\Omega) := \{A : A \subset \Omega\}$$

Applied to our example this would be:

$$\begin{aligned}\mathcal{A} = \{&\emptyset, \{\text{The die shows one pip}\}, \{\text{The die shows two pips}\}, \ldots,\\ &\{\text{The die shows one pip, The die shows two pips}\},\\ &\{\text{The die shows one pip, The die shows three pips}\}, \ldots\}\end{aligned}$$

Example 1.3. Given a subset $A \in \Omega$ we can always construct a σ-algebra, by: $\mathcal{A} := \{\emptyset, A, A^c, \Omega\}$. This is called the σ-algebra over Ω induced by A.

Definition 1.4. The pair $(\Omega, \mathcal{A})$ is called *measurable space*. A function $\mu : \mathcal{A} \to \mathbb{R} \cup \{\infty\}$ with

- $\mu(\emptyset) = 0$
- $\mu(A) \geq 0 \quad \forall\, A \in \mathcal{A}$
- $\mu\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \quad \forall\, A_i \in \mathcal{A} \quad$ with $A_i \cap A_j = \emptyset$ for all $i \neq j$

is a *measure* on Ω. We name a triple $(\Omega, \mathcal{A}, \mu)$ *measure space*.

Definition 1.5. A measure $\mathbb{P}$ on $(\Omega, \mathcal{A})$ is a *probability measure* if $\mathbb{P}(\Omega) = 1$. The triple $(\Omega, \mathcal{A}, \mathbb{P})$ is a *probability space*.

Example 1.6. We continue our previous example. If the die is fair we assume all outcomes (i.e. how many pips the die shows) to be equally likely. Hence we give each outcome the some "chunk" of probability:

$$\mathbb{P}(\{\text{The die shows } i\text{pips}\}) = \frac{1}{6} \quad \forall\, i = 1, 2, \ldots, 6 \tag{1.1}$$

We easily check that $\mathbb{P}(\Omega) = 1$. A sceptic person may ask, how to construct a probability measure from the above equation, since we don't know what the measure of an event like "the die shows one or five pips" is. The answer lies in the physical fact that the die can only show one side at a time. It can not show one pip and five pips at the same time. Or in other words the outcomes in 1.1 are all disjoint. Hence we can construct by the properties of a measure all events by just adding the probabilities of all outcomes, which lie in that event. In this way the probability for the event "the die shows one or five pips" is - as expected - $\frac{2}{6}$.

This intuitive construction is important enough to be generalised in the next example.

Example 1.7. In the discrete case, i.e. if $\Omega = \{\omega_1, \omega_2, \ldots, \}$ is a countable set, we can define probability weights $p_i = \mathbb{P}(\{\omega_i\}) \geq 0$ with $\sum_{i \in \Omega} p_i = 1$. By defining these weights we construct a uniquely defined probability measure on $(\Omega, \mathcal{P}(\Omega))$ by:

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} p_i \delta_A(\omega_i)$$

with

$$\delta_A(\omega_i) = \begin{cases} 1 & \text{if } \omega_i \in A \\ 0 & \text{if } \omega_i \notin A \end{cases}$$

Example 1.8. If $\Omega = \mathbb{R}$ one usually uses the *Borel σ-algebra* denoted by $\mathcal{B}(\mathbb{R})$. This special σ-algebra is the smallest σ-algebra on $\mathbb{R}$, which contains all open intervals (i.e. we intersect all σ-algebras, which contain all open intervals. Since the power set of $\mathbb{R}$ contains all open intervals and is a σ-algebra, we know that this intersection is not empty.). On this measurable space $(\mathbb{R}, \mathcal{B})$ we can define probability measures by the usage of *probability density functions* (pds). A density $f : \mathbb{R} \to \mathbb{R}_0^+$ fulfills

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

and its corresponding probability measure is given by:

$$\mathbb{P}(A) = \int_A f(x)dx$$

There is a good reason why we use the non-intuitive Borel σ-algebra, instead of $\mathcal{P}(\mathbb{R})$. One can show, that the power set is "too large" to define a measure on it. This means, if one tries to define a non-trivial measure (a measure is trivial if $\mu(A) = 0$ for all $A \in \mathcal{A}$) on $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ this leads inevitably to a contradiction (so called Vitali sets).

We have seen in our die example, that the space of outcomes can be something physical. However, we would like to translate this physical descriptions into something more tractable, like real numbers. The next very important definition allows us to do so.

Definition 1.9. A function $X : \Omega \to \Omega'$ between two measurable spaces $(\Omega, \mathcal{A})$ and $(\Omega', \mathcal{A}')$ is *measurable* if

$$\{X \in A'\} \in \mathcal{A} \quad \forall\, A' \in \mathcal{A}'$$

holds ($\{X \in A'\} := \{\omega \in \Omega : X(\omega) \in A'\}$).

A *random variable* is a measurable function from a probability space into a measurable space, i.e. $X : (\Omega, \mathcal{A}, \mathbb{P}) \to (\Omega', \mathcal{A}')$. A random variable X induces a probability measure on $(\Omega', \mathcal{A}')$ by:

$$\mathbb{P}_X(A') := \mathbb{P}(\{X \in A'\}) =: \mathbb{P}(X \in A')$$

The probability measure $\mathbb{P}_X$ is called *distribution* of X.

Often one speaks just of the distribution of X and assumes a corresponding probability space $(\Omega, \mathcal{A}, \mathbb{P})$ to exist. In the very most cases this is unproblematic.

Example 1.10. If we return again to the die example an intuitive real random variable $X : \Omega \to \mathbb{R}$ would look like the following:

$$\begin{aligned}
X(\text{The die shows one pip}) &= 1 \\
X(\text{The die shows two pips}) &= 2 \\
X(\text{The die shows three pips}) &= 3 \\
X(\text{The die shows four pips}) &= 4 \\
X(\text{The die shows five pips}) &= 5 \\
X(\text{The die shows six pips}) &= 6
\end{aligned}$$

Definition 1.11. For a real random variable X we define the *(cumulative) distribution function (cdf)* by:

$$\begin{aligned} F_X :&\mathbb{R} \to [0,1] \\ &x \mapsto F(x) := \mathbb{P}(X \leq x) \end{aligned}$$

By its distribution function F_X the random variable X is uniquely determined.

We can see now, why in most cases we are not interested in the original probability space. If we have a real valued random variable with some distribution function F we can construct the original probability space like follows: Use $\Omega = \mathbb{R}$ and $\mathcal{A} = \mathcal{B}(\mathbb{R})$ and define $\mathbb{P}$ by:

$$\mathbb{P}((-\infty, x]) = F(x)$$

Example 1.12. We say a random variable is exponentially distributed with parameter λ, if it has the following distribution function:

$$F(x, \lambda) = 1 - e^{-\lambda x}$$

The corresponding density $f(x, \lambda)$ is given by differentiating F:

$$F'(x) = \lambda e^{-\lambda x}$$

Example 1.13. A random variable is normally distributed with parameters μ and σ^2, if it has the following distribution function:

$$F(x, \mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{(y-\mu)^2}{2\sigma^2}} dy$$

Example 1.14. A random variable is Pareto distributed with parameters x_m and α, if it has the following distribution function:

$$F(x, \alpha, x_m) = \begin{cases} 1 - \left(\frac{x_m}{x}\right)^{\alpha} & x \geq x_m \\ 0 & x < x_m \end{cases}$$

Its density is given by:

$$f(x, \alpha, x_m) = \frac{\alpha x_m^{\alpha}}{x^{\alpha+1}} \qquad x \geq x_m$$

Exercises

Exercise 1.15. Show that the function $\mathbb{P}$ defined in example 1.7 is a probability measure, i.e. check all the conditions for $\mathbb{P}$ to be a measure, as given in definitions 1.4 and 1.5. Show further that the probability measure is uniquely defined, i.e. assume another function $\mathbb{P}'$ with $\mathbb{P}(\{\omega_i\}) = \mathbb{P}'(\{\omega_i\})$ for all $i \in \mathbb{N}$ and show:

$$\mathbb{P}'(A) = \mathbb{P}(A) \qquad \forall\, A \in \mathcal{A}$$

Exercise 1.16. Let $\mathcal{A}$ and $\mathcal{B}$ be two different σ-algebras on the same space Ω. Show that the intersection of the two σ-algebras $\mathcal{A} \cap \mathcal{B}$ is again a σ-algebra on Ω.

Exercise 1.17. Let X be a real random variable with differentiable densitiy f. Show that the distribution of X is an antiderivative (german: "Stammfunktion") of f, i.e.:

$$\frac{d}{dx}F(x) = f(x)$$

(Hint: Use the Fundamental Theorem of Calculus)

Exercise 1.18. Let $(\Omega, \mathcal{A})$ be a measurable space and $A_1, A_2, A_3, \ldots \in \mathcal{A}$. Show

$$\{\omega : \omega \in A_i \text{ for finite many } i \in \mathbb{N}\} \in \mathcal{A}$$

and

$$\{\omega : \omega \in A_i \text{ for infinite many } i \in \mathbb{N}\} \in \mathcal{A}.$$

Exercise 1.19. Let $(\Omega, \mathcal{A})$ be a measurable space and $A, B \in \mathcal{A}$. Prove in this sequence:

- $\mathbb{P}(A) \geq \mathbb{P}(B)$ if $A \supseteq B$
- $\mathbb{P}(A \setminus B) = \mathbb{P}(A) - \mathbb{P}(A \cap B)$
- $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$
- $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$

Exercise 1.20. Many programming languages offer methods to create random numbers in $[0, 1]$. How can you use this to create realisations of an exponential distributed random variable with parameter λ? Use this method to simulate n different exponential distributed random variables X_i with the same parameter λ and compute the arithmetic mean of these realisations $E = \frac{1}{n}\sum_{i=1}^{n} X_i$. How does $E - \frac{1}{\lambda}$ evolve for large n?

Exercise 1.21. Use the procedure of the previous exercise to generate exponential distributed random numbers $x_1, x_2, x_3, \ldots$ with parameter λ. Now count the occurences of this random numbers in certain intervals. Define these intervals of length $N \in \mathbb{R}$ by:

$$I_n = [n \cdot N, (n+1) \cdot N) \subset \mathbb{R} \qquad n \in \mathbb{N}_0$$

The number of realisations in one interval is then given by:

$$f^*(n) = \sum_{i=1} \delta_{I_n}(x_i)$$

Plot the sequence $(f^*(n))_{n\in\mathbb{N}}$. Compare this sequence with the density of an exponentially distributed random variable X with the same parameter λ.

Exercise 1.22. A geometric distributed random variable X is a discrete random variable with probability weights:

$$\mathbb{P}(X = n) = (1-p)^n p \qquad n \in \mathbb{N}_0$$

for some parameter $p \in (0,1)$. Think about a method to generate geometric distributed random numbers.

(Hint: $\mathbb{P}(X = n) = (1-p)^n p = (1-p) \cdot (1-p)^{n-1} p = (1-p) \cdot \mathbb{P}(X = n-1)$)

Generate some geometrically distributed random variables and plot the result.

Take the sequence $(f^*(n))_{n\in\mathbb{N}}$ of the previous exercise, with parameter $\lambda = -\log(1-p)$ and intervals $I_n = [n, n+1)$ and compare it with the plot of geometrically distributed random variables.

Exercise 1.23. Prove that for X being an exponentially distributed random variable with parameter $\lambda = -\log(1-p)$ holds

$$\mathbb{P}(X \in [n, n+1)) = \mathbb{P}(Y = n)$$

where Y is a geometrically distributed random variable with parameter p.

1.2 Moments and Stochastic Independence

Definition 1.24. Let X, Y be real random variables with densities. The *expectaction* of X is given by:

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} x f(x) dx$$

where f is the density of X. Further is

$$\mathbb{E}(h(X)) = \int_{-\infty}^{\infty} h(x) f(x) dx$$

For the special case of $h(x) = |x|^n$ we talk about the *n-th moment* of X:

$$\mathbb{E}(|X|^n) = \int_{-\infty}^{\infty} |x|^n f(x) dx$$

The *variance* of X is given by:

$$\mathrm{Var}(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$$

The *covariance* of X and Y is given by:

$$\mathrm{Cov}(X, Y) = \mathbb{E}((X - \mathbb{E}(X))(Y - \mathbb{E}(Y)))$$

Example 1.25. The expectation of an exponential random variable with parameter λ can be calculated by:

$$\begin{aligned}\mathbb{E}(X) &= \int_0^\infty x \cdot \lambda e^{-\lambda x} dx = \lim_{a\to\infty}(-ae^{-\lambda a} - (-0 \cdot e^{-\lambda \cdot 0})) - \int_0^\infty -e^{-\lambda x} dx \\ &= 0 - \frac{1}{\lambda}\lim_{a\to\infty}(e^{-\lambda a} - e^{-\lambda \cdot 0}) = \frac{1}{\lambda}\end{aligned}$$

Similarly one has for the variance $\mathrm{Var}(X) = \frac{1}{\lambda^2}$

Example 1.26. The expectation and variance of a normal distributed random variable with parameters μ and σ^2 are $\mathbb{E}(X) = \mu$ and $\mathrm{Var}(X) = \sigma^2$.

Example 1.27. The expectation of the Pareto distribution with parameters x_m and α is given by:

$$\begin{aligned}\mathbb{E}(X) &= \int_{x_m}^\infty x \cdot \frac{\alpha x_m^\alpha}{x^{\alpha+1}} dx = \alpha x_m^\alpha \int_{x_m}^\infty x^{-\alpha} dx \\ &= \alpha x_m^\alpha \lim_{a\to\infty} \frac{1}{1-\alpha} a^{-\alpha+1} - \frac{1}{1-\alpha} x_m^{-\alpha+1} \\ &= -\alpha x_m^\alpha \cdot \frac{1}{1-\alpha} \cdot \frac{1}{x_m^{\alpha-1}} = \frac{\alpha x_m}{\alpha - 1}\end{aligned}$$

if $\alpha > 1$. For the cases that $\alpha \leq 1$ the integral $\int_{x_m}^\infty x^{-\alpha} dx = \infty$ does not converge. In this case we say that the expectation of X is not existent. Analogous one gets for $\alpha \leq 2$ that the variance of X (or otherwise stated the second moment) does not exist.

From the definition of the expectation as an integral, we know immediately that the expectation is linear and monotone:

$$\mathbb{E}(aX + Y) = a\mathbb{E}(X) + \mathbb{E}(Y) \qquad \forall\, a \in \mathbb{R}$$

$$\mathbb{E}(X) \geq \mathbb{E}(Y) \qquad \text{if } X \geq Y \text{ a.s.}$$

(The expression "a.s." is the abbreviation for almost surely and means that some event occurs with probability one, i.e. $\mathbb{P}(X \geq Y) = 1$). With a bit more effort (which we not invest here) one can proof for a sequence of non-negative random variables X_n that even

$$\mathbb{E}\left(\sum_{n=1}^\infty X_n\right) = \sum_{n=1}^\infty \mathbb{E}(X_n) \tag{1.1}$$

holds.

The variance is also called the centered second moment, it can be seen as the expected deviation from the expectation. Also it can be reformulated by:

$$\begin{aligned}\mathrm{Var}(X) &= \mathbb{E}(X^2 - 2X\mathbb{E}(X) + \mathbb{E}(X)^2) \\ &= \mathbb{E}(X^2) - 2\mathbb{E}(X)\mathbb{E}(X) + \mathbb{E}(\mathbb{E}(X)^2) \\ &= \mathbb{E}(X^2) - \mathbb{E}(X)^2\end{aligned}$$

Further we know that the variance fulfills:

$$\mathrm{Var}(aX + b) = a^2 \mathrm{Var}(X)$$

To calculate the covariance we need the two-dimensional density of the random vector (X, Y):

$$f(x, y) \geq 0 \qquad \forall\, (x, y) \in \mathbb{R}^2$$
$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx\, dy = 1$$

We can define the marginal densities f_X and f_Y of the single random variables X and Y. They can be computed by:

$$f_X(x) = \int_{-\infty}^{\infty} f(x, y) dy$$
$$f_Y(y) = \int_{-\infty}^{\infty} f(x, y) dx$$

The covariance can then be expressed by:

$$\mathrm{Cov}(X, Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y)(x - \mathbb{E}(X))(y - \mathbb{E}(Y)) dx\, dy$$

Further we have:

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}(Y, X) \tag{1.2}$$
$$\mathrm{Cov}(X, X) = \mathrm{Var}(X) \tag{1.3}$$
$$\mathrm{Cov}(aX + b, cY + d) = ac\mathrm{Cov}(X, Y) \tag{1.4}$$
$$\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) \tag{1.5}$$
$$\mathrm{Var}(X + Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X, Y) \tag{1.6}$$

If $\mathrm{Cov}(X, Y) = 0$ we say the two random variables X and Y are uncorrelated.

Definition 1.28. Two events $A, B \in \mathcal{A}$ are *stochastically independent* if

$$\mathbb{P}(A \cap B) = \mathbb{P}(A) \cdot \mathbb{P}(B)$$

Let $\mathcal{A}_1, \mathcal{A}_2, \ldots, \mathcal{A}_n \subset \mathcal{A}$. The sets $\mathcal{A}_i$ are *stochastically independent* if for each selection $\mathcal{A}_{i_1}, \mathcal{A}_{i_2}, \ldots, \mathcal{A}_{i_k}$ of the $\mathcal{A}_i$ and all subsets $A_{i_j} \in \mathcal{A}_{i_j}$ holds:

$$\mathbb{P}(A_{i_1} \cap A_{i_2} \cap \ldots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \ldots \cdot \mathbb{P}(A_{i_k})$$

The second definition is stronger than the *pairwise stochastic independence*, which is given by $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$ for all $i \neq j$ and a family of sets $A_i \in \mathcal{A}$ and also stronger than the demand $\mathbb{P}(\bigcap_{i=1}^n A_i) = \prod_{i=1}^n \mathbb{P}(A_i)$. The following simple example makes this clear.

Assume one throws two dice. We define three events:

$$A = \{\text{The first die shows an even number of pips}\}$$
$$B = \{\text{The second die shows an even number of pips}\}$$
$$C = \{\text{The sum of pips shown by both dice is even}\}$$

We can easily check that $\mathbb{P}(A) = \mathbb{P}(B) = \mathbb{P}(C) = \frac{1}{2}$ and that $\mathbb{P}(A \cap B) = \mathbb{P}(A \cap C) = \mathbb{P}(B \cap C) = \frac{1}{4}$. But this pairwise stochastic independence does not infer stochastic independence of the familiy $\mathcal{D} = \{A, B, C\}$ since:

$$\mathbb{P}(A \cap B \cap C) = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$$

Definition 1.29. Let $X_i : (\Omega, \mathcal{A}) \to (\Omega', \mathcal{A}')$ be a set of random variables $(i = 1, \ldots, n)$. They are *stochastically independent*, if their pre-image-σ-algebras are stochastically independent sets. In expression, if for every selection $\{i_1, \ldots, i_k\} \subset \{1, \ldots, n\}$ and all $A'_{i_j} \in \mathcal{A}'$ $(j = 1, \ldots, k)$ holds:

$$\mathbb{P}\left(\bigcap_{j=1}^{k} \{X_{i_j} \in A'_{i_j}\}\right) = \prod_{j=1}^{k} \mathbb{P}(X_{i_j} \in A'_{i_j})$$

Let X_i be a set of stochastically independent random variables $(i = 1, \ldots, n)$. Denote by $X = (X_1, \ldots, X_n)$ the corresponding random vector, by F_X its cdf and by f_X its density (if existent). Then holds:

- $F_X(x_1, \ldots, x_n) = \prod_{i=1}^{n} F_{X_i}(x_i) \qquad \forall\, x_i \in \mathbb{R}$
- $f_X(x_1, \ldots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i) \qquad \forall\, x_i \in \mathbb{R}$
- $\mathbb{E}(\prod_{i=1}^{n} X_i) = \prod_{i=1}^{n} \mathbb{E}(X_i)$ and $\mathrm{Cov}(X_i, X_j) = 0$ for all $i \neq j$[1]

In the modelling of stochastic behaviour it is more common to *assume* the independence of events, instead of *proving* that something is stochastically independent. Usually, when one assumes that two outcomes have no influence on each other, this is modelled via stochastic independence. One example would be the (simultaneous) throwing of two fair dice. You can model each die as one random variable, which are stochastically independent of each other. Alternatively you could model both dice by just one random variable with values in $\{1, \ldots, 6\} \times \{1, \ldots, 6\}$ and show that all events concerning only one die (e.g. the first die shows one pip) are stochastically independent from the events, which concern only the other die (see for this the upcoming example).

Definition 1.30. Let A, B be two events in the probability space $(\Omega, \mathcal{A}, \mathbb{P})$ with $\mathbb{P}(B) > 0$. The *conditional probability* of A given B is defined by:

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$

[1] Watch out, that from $\mathrm{Cov}(X, Y) = 0$ does not necessarily follow stochastic independence between X and Y! See exercise 1.41

The conditional probability of some event A given B is the probability that A occurs, if we already know, that B will occur (or has occured).

Example 1.31. We consider the before mentioned two dice experiment. We model the throwing of two dice as a random variable X, mapping from a suitable probability space $(\Omega, \mathcal{A}, \mathbb{P})$ into the measurable space $(\Omega' := \{1, \dots, 6\}^2, \mathcal{P}(\Omega'))$. Since the dice are fair, we assume each outcome in Ω' to be equally likely, i.e. each pair $(i, j) \in \Omega'$ has probability $\mathbb{P}_X((i, j)) = \frac{1}{36}$.

We investigate two events. The first event denoted by C, is that the first die shows at least 5 pips. We calculate:

$$\mathbb{P}_X(C) = \mathbb{P}_X(\{5, 6\} \times \{1, \dots, 6\}) = \frac{12}{36} = \frac{1}{3}$$

Next we investigate the event A, which is that the sum of both dice is at least 5 pips, if we already know, that the first die shows an even number of pips. For this, we denote by B the event, that the first die shows an even number of pips. The probability of A is given now, by the conditional probability:

$$\mathbb{P}_X(A|B) = \frac{\mathbb{P}_X(A \cap B)}{\mathbb{P}_X(B)} = \frac{\mathbb{P}_X(\{(2, j)|j \in \{3, \dots, 6\}\} \cup \{4, 6\} \times \{1, \dots, 6\})}{\mathbb{P}_X(\{2, 4, 6\} \times \{1, \dots, 6\})} = \frac{\frac{16}{36}}{\frac{1}{2}} = \frac{8}{9}$$

Example 1.32. Another typical example of dependence is sampling without replacement. Imagine an urn with colored balls in it. Someone draws randomly the balls from the urn, one by one and we might ask, what is the probability that the n-th ball has a certain color. For our example imagine the urn is filled with two blue and one green ball. We ask for the probability that the second ball drawn, will be the green one, this event will be denoted by G_2. To model this we need to find a suitable event space Ω first. If we number the blue balls by b_1 and b_2 as well as the green ball by g, we see that the following event space describes all possibilites to draw the balls from the urn:

$$\Omega = \{(b_1, b_2, g), (b_2, b_1, g), (b_1, g, b_2), (b_2, g, b_1), (g, b_1, b_2), (g, b_2, b_1)\}$$

Since the balls are drawn completely random, we can assume each of the outcomes has the same probability $\frac{1}{6}$, hence we have $\mathbb{P}(G_2) = \frac{1}{3}$. If we now condition on the event, that the first ball is a blue one and want to know again the probability that the second ball is the green one, we have (denoting by B_1 the event of first drawing a blue ball):

$$\mathbb{P}(G_2|B_1) = \frac{\mathbb{P}(G_2 \cap B_1)}{\mathbb{P}(B_1)} = \frac{\frac{1}{3}}{\frac{2}{3}} = \frac{1}{2}$$

The equality $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$ for two stochastically independent random variables X and Y carries over to many functions of X and Y, i.e.

$$\mathbb{E}(f(X) \cdot g(Y)) = \mathbb{E}(f(X))\mathbb{E}(g(Y))$$

For this to hold f and g need to be only measurable functions mapping from $\mathbb{R}$ to $\mathbb{R}$. We omit the prove of this statement, as it is rather intuitive.

Definition 1.33. The *moment generating function (MGF)* of a real random variable X is given by:

$$\begin{aligned}\phi_X :&\mathbb{R} \to \mathbb{R}\\ \phi_X :&\theta \mapsto \phi_X(\theta) := \mathbb{E}(e^{\theta X})\end{aligned}$$

If $\mathbb{E}(e^{\theta X}) = \infty$ we say that the MGF of X is not existent at θ.

Corollary 1.34. *Let X_i be a family of stochastically independent, identically distributed (i.i.d.) variables ($i = 1, \ldots, n$) and define $S_n = \sum_{i=1}^n X_i$. Assume $\phi_{X_i}(\theta)$ exists, then holds:*

$$\phi_{S_n}(\theta) = (\phi_{X_i}(\theta))^n$$

Proof. We have:

$$\begin{aligned}\mathbb{E}(e^{\theta S_n}) = \mathbb{E}(e^{\theta \sum_{i=1}^n X_i}) &= \mathbb{E}\left(\prod_{i=1}^n e^{\theta X_i}\right)\\ &= \prod_{i=1}^n \mathbb{E}(e^{\theta X_i}) = (\phi_{X_i}(\theta))^n\end{aligned}$$

□

Example 1.35. Let X_i be i.i.d. exponentially distributed random variables with parameter λ. Define the random variable $S_n = \sum_{i=1}^n X_i$. We say that such a random variable is Erlang distributed with parameters n and λ. Its MGF is given by:

$$\mathbb{E}(e^{\theta S_n}) = (\mathbb{E}(e^{\theta X_i}))^n = \left(\lambda \int_0^\infty e^{(\theta-\lambda)x} dx\right)^n$$

Let now $\theta < \lambda$:

$$\begin{aligned}\mathbb{E}(e^{\theta S_n}) &= \left(\lambda \lim_{a\to\infty} \frac{1}{\theta-\lambda}(e^{(\theta-\lambda)a} - e^{(\theta-\lambda)\cdot 0})\right)^n\\ &= \left(\frac{\lambda}{\lambda-\theta}\right)^n\end{aligned}$$

We see, that for $\theta \geq \lambda$ the MGF is not existent.

Example 1.36. Let X be a Pareto distributed random variable with parameters x_m and α. You are asked in the exercises to show that the MGF of X does not exist for any $\theta > 0$, no matter how the parameters x_m and α are chosen.

We state here some usefull properties of the moment generating function, without explicitly proving them:

- If two random variables have the same existent MGF they also have the same distribution.

- For non-negative random variables and varying θ the MGF exists in an interval $(-\infty, a)$ with $a \in \mathbb{R}_0^+$.
- The MGF always exists for $\theta = 0$ and if it exists in the interval $(-a, a)$ it is infinitely often differentiable on that interval.
- The MGF is convex.

Exercises

Exercise 1.37. Prove equations (1.1)-(1.4).

Exercise 1.38. What is the n-th moment of a Pareto distributed random variable and when does it exist?

Exercise 1.39. Prove that the variance of an exponentially distributed random variable is equal to $\frac{1}{\lambda^2}$.

Exercise 1.40. Construct an example, in which for three events $A, B, C \subset \mathcal{A}$ holds $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$ but at least one of the following equalities *does not* hold:

$$\begin{aligned}\mathbb{P}(A \cap B) &= \mathbb{P}(A)\mathbb{P}(B)\\ \mathbb{P}(A \cap C) &= \mathbb{P}(A)\mathbb{P}(C)\\ \mathbb{P}(B \cap C) &= \mathbb{P}(B)\mathbb{P}(C)\end{aligned}$$

(Hint: Draw three intersecting circles and give each area a certain probability weight, such that the weights sum up to 1.)

Exercise 1.41. Let X and Y be independent and Bernoulli distributed with the same parameter p (this means $\mathbb{P}(X = 1) = p$ and $\mathbb{P}(X = 0) = 1 - p$). Show that the two new random variables $X + Y$ and $X - Y$ are uncorrelated, yet not independent.

Exercise 1.42. Let X be some real random variable, such that the MGF of X exists in an interval around 0. Calculate the first derivative of the MGF at zero $\frac{d}{d\theta}\phi_X(\theta)\big|_{\theta=0}$. Calculate the n-th derivative of the MGF at zero $\left(\frac{d}{d\theta}\right)^n \phi_X(\theta)\big|_{\theta=0}$. (Hint: Use that $e^x = \sum_{i=0}^{\infty} \frac{x^i}{i!}$ and use (1.1).)

Exercise 1.43. In this exercise prove first the law of total probability: If $A \in \Omega$ and $B \in \Omega$ are some events in a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ it holds:

$$\mathbb{P}(A) = \mathbb{P}(A|B)\mathbb{P}(B) + \mathbb{P}(A|B^c)\mathbb{P}(B^c)$$

Convince yourself that for some partition $(B_n)_{n\in\mathbb{N}}$ of Ω we similarly have:

$$\mathbb{P}(A) = \sum_{n=1}^{\infty} \mathbb{P}(A|B_n)\mathbb{P}(B_n)$$

(A partition fulfills: $\bigcup_{n=1}^{\infty} B_n = \Omega$ and $B_n \cap B_m = \emptyset$ for all $n \neq m$)

Now if we have two real random variables X and Y with densities f_X and f_Y, we can reformulate the law of total probability:

$$\mathbb{P}(X \in A) = \int_{-\infty}^{\infty} \mathbb{P}(X \in A | Y = y) f_Y(y) dy$$

Assume now that X is an exponentially distributed random variable with parameter λ and Y is stochastically independent of X. Show that the probability that Y is smaller than X is equal to the MGF of Y at the point λ, in expression:

$$\mathbb{P}(X > Y) = \mathbb{E}(e^{-\lambda Y})$$

(Hint: Use that $\mathbb{E}(h(Y)) = \int_{-\infty}^{\infty} h(y) f_Y(y) dy$)

Exercise 1.44. Show that for the Pareto distribution and every $\theta > 0$ the corresponding MGF does not exist.

Exercise 1.45. The log-normal distribution is defined by the following density:

$$f(x, \mu, \sigma^2) = \frac{1}{x\sigma\sqrt{2\pi}} e^{-\left(\frac{(\ln x - \mu)^2}{2\sigma^2}\right)} \qquad \forall\, x > 0$$

Its moments are given by $\mathbb{E}(X^n) = e^{n\mu + \frac{1}{2}\sigma^2 n^2}$. Show that for all $\theta > 0$ the corresponding MGF does not exist.

Exercise 1.46. This exercise continues exercises 1.20 and 1.21 in which we have seen, that the arithmetic mean converges to the expectation. Simulate again n exponentially distributed numbers and compute $V = \frac{1}{n} \sum_{i=1}^{n} (x - \frac{1}{\lambda})^2$. How does $V - \frac{1}{\lambda^2}$ evolve for large n? Plot $V - \frac{1}{\lambda^2}$ against n for different values of λ. When converges $V - \frac{1}{\lambda^2}$ fast and when slow?

2 Stochastic Arrivals and Service

In this chapter we carefully motivate the need for stochastic arrivals and service and see two approaches how such stochastic processes can be described and bounded. Each of them leads to its own stochastic extension of the deterministic network calculus (see [?, ?, ?, ?]). So we cannot talk about *the* stochastic network calculus, but rather have to distinguish between several approaches, each with its own characteristics and strengths.

Overall we use a discrete time-scale, since it keeps formulas somewhat easier.

2.1 Stochastic Arrivals

As in deterministic network calculus we abstract data arriving at some service element as flows. The service element is abstracted as a node, which takes the arriving flow as input and relays it to a departing flow under a certain rate and scheduling policy. A node can have many different arriving flows and produces the same number of departing flows, each of them corresponding to one arrival flow. To describe such a flow we use cumulative functions satisfying:

$$\begin{aligned} A(n) &\geq 0 \qquad \forall\, n \in \mathbb{N}_0 \\ A(n) - A(m) &\geq 0 \qquad \forall\, n \geq m \in \mathbb{N}_0 \end{aligned} \tag{2.1}$$

Here the number of packets/data/bits/"whatever is flowing" up to time t is given by $A(t)$. In deterministic network calculus such a flow is given and no random decisions are involved. Now we want to generalise this idea, in the sense that chance plays a role. To do so we need the definition of stochastic processes:

Definition 2.1. Let $(\Omega, \mathcal{A}, \mathbb{P})$ be a probability space, $(\Omega', \mathcal{A}')$ a measurable space and I an index set. A familiy of random variables $X = (X_i)_{i \in I}$ with $X_i : \Omega \to \Omega'$ is called *stochastic process*, with *state space* Ω' and *time space* I.

Remark 2.2. If I is totally ordered (e.g. when $I \subset \mathbb{R}$) we can fix one $\omega \in \Omega$ and interpret X_i as a function in the "time"-variable i:

$$\begin{aligned} X(\omega) :& I \to \Omega' \\ X(\omega) :& i \mapsto X_i(\omega) \end{aligned}$$

This mapping is called *trajectory* or *(sample) path* of X under ω.[1]

[1] This means, that ω contains "enough" information, to deduce complete trajectories from it. In fact, we often use more than one stochastic process simultaneously, i.e. have several families of random vari-

We give a few examples to illustrate what kind of structures the above definition can cover. However, we are not going into much detail here.

Example 2.3. Markov Chains are stochastic processes with a countable state space and discrete or continuous time space, i.e. $I \in \{\mathbb{N}_0, \mathbb{R}_0^+\}$. In the special case of $\Omega' = \mathbb{N}_0$ and the condition that $X_n \in [X_{n-1} - 1, X_{n-1} + 1]$ our Markov Chain becomes a Birth-Death-Process.

Example 2.4. The most famous continuous time stochastic process is the Brownian Motion or the Wiener Process (see for example §40 in [**?**]). The one-dimensional Wiener Process W_t has state space $\mathbb{R}$, time space $\mathbb{R}_0^+$ and satisfies:

- $W_0 = 0$
- The trajectories of W_t are continuous, a.s.
- $W_t - W_s$ has normal distribution with parameters $\mu = 0$ and $\sigma^2 = t - s$. $(t \geq s \geq 0)$
- If $t \geq s > u \geq v$, then the increments $W_t - W_s$ and $W_u - W_v$ are stochastically independent. (A similar condition holds for n different increments. See Definition 1.28)

Showing the existence of such a process lies out of the scope of this course and is hence omitted.

Example 2.5. The time space is very often something "natural" like $\mathbb{N}_0$ or $\mathbb{R}_0^+$, but one should be aware, that the above definition, covers much more. Other appearing index sets are $[0, 1] \subset \mathbb{R}$ for the Brownian Bridge or $I = T$ for tree-indexed random processes, T being the set of nodes in a finite or infinite tree.

In our context of arrivals, we use a stochastic process with time space $\mathbb{N}_0$, while the state space is kept continuous ($\Omega' = \mathbb{R}_0^+$). Further, we want to preserve the cumulative nature of (2.1).

Definition 2.6. Let $(a(i))_{i \in \mathbb{N}}$ be a sequence of non-negative real random variables. We call the stochastic process A with time space $\mathbb{N}_0$ and state space $\mathbb{R}_0^+$ defined by

$$A(n) := \sum_{i=1}^{n} a(i)$$

a *flow*[2]. The $a(i)$ are called *increments* of the flow.

ables $X_i, Y_j, Z_k, \ldots$ each defining its own stochastic process. How this is possible, is best understood if one assumes $I = \mathbb{N}$. In this case, we might represent each event by $\omega = (\omega_{x_1}, \omega_{y_1}, \omega_{z_1}, \omega_{x_2}, \ldots)$ such that each X_i only depends on ω_{x_i}. Of course, the corresponding event space Ω is then quite large. In fact if one sets I equal to an uncountable set, the construction of the original probability space $(\Omega, \mathcal{A}, \mathbb{P})$ is a non-trivial task. In this course, however, we do not delve into this topic and are just happy that someone has done the work of constructing $(\Omega, \mathcal{A}, \mathbb{P})$ for us.

[2] As usual we define the empty sum as zero: $\sum_{i=1}^{0} a(i) = 0$

We see in the following two examples that it is quite hard to bound flows with *deterministic* arrival curves. (A flow has a deterministic arrival curve $\alpha : \mathbb{N}_0 \to \mathbb{R}_0^+$ if $A(n) - A(m) \leq \alpha(n-m)$ for all $n \geq m \geq 0$.)

Example 2.7. First consider the $a(i)$ to be i.i.d. Bernoulli distributed. This means for each $a(i)$ holds:

$$\begin{aligned}\mathbb{P}(a(i) = 0) &= p \\ \mathbb{P}(a(i) = 1) &= 1 - p\end{aligned}$$

for some parameter $p \in (0,1)$. Is it possible that $A(n) - A(m) = n - m$? The answer is yes and we show this, by proving that the corresponding probability is larger than zero:

$$\begin{aligned}\mathbb{P}(A(n) - A(m) = n - m) &= \mathbb{P}\left(\sum_{k=m+1}^{n} a(k) = n - m\right) \\ &= \mathbb{P}\left(\bigcap_{k=m+1}^{n} \{a(k) = 1\}\right) = \prod_{k=m+1}^{n} \mathbb{P}(a(k) = 1) \\ &= (1-p)^{n-m} > 0\end{aligned}$$

Hence it is possible to encounter exactly $n - m$ arrivals in an interval of length $n - m$ (although the corresponding probability might be tiny). Hence the best deterministic arrival curve, we can give for such an arrival is $\alpha(n) = n$. With our deterministic "worst-case-glasses" we see A sending 1 in each time step. The possibility that the flow could send nothing in one time step is completely ignored! This buries any hopes to effectively bound a stochastic process by *deterministic* arrival curves. The next example shows, that the situation can become even worse.

Example 2.8. Next assume that the increments $a(i)$ are i.i.d. and exponentially distributed with parameter λ. In this case the best possible deterministic "arrival curve" is given by $\alpha(n) = \infty$ for all $n \in \mathbb{N}$, since we have for any $K \geq 0$:

$$\mathbb{P}(a(i) > K) = e^{-\lambda K} > 0$$

Hence no matter how large we choose K, there is always a small probability, that the arrivals in one time step exceeds K.

The previous two examples are bad news for someone trying to bound stochastic arrivals by deterministic arrival curves. The following corollary summarizes the possibilities one has to deterministically bound i.i.d. arrivals. The proof is generalised easily from example 2.7 and hence omitted.

Corollary 2.9. *Let the increments $a(i)$ of some flow be i.i.d. and define $x_+ := \inf\{x : F(x) = 1\} \in [0, \infty]$, where F is the cdf of $a(i)$. Then the best possible arrival curve for A is given by: $\alpha(n) = n \cdot x_+$.*

We see it is a somewhat pointless enterprise to describe stochastic arrivals by deterministic arrival curves. The solution to this problem is to exclude events, which are unlikely to happen. So instead of asking for some curve α such that $A(n) - A(m) \leq \alpha(n-m)$ we are rather interested in bounds, which hold with a high probability. Stated in a simple form, this looks like

$$\mathbb{P}(A(n) - A(m) \leq \alpha(n-m, \varepsilon)) \geq 1 - \varepsilon \qquad \forall\, n > m \in \mathbb{N}$$

which is equivalent to:

$$\mathbb{P}(A(n) - A(m) > \alpha(n-m, \varepsilon)) < \varepsilon \qquad \forall\, n > m \in \mathbb{N} \tag{2.2}$$

Here $\alpha(., \varepsilon)$ is an arbitrary function, with some parameter $\varepsilon > 0$. [3] Starting from (2.2) two different kinds of bounds can be formulated. The first one is the so-called *tail-bound*, for which we need two more definitions in place:

Definition 2.10. An *envelope function,* or just *envelope,* is a function α with:

$$\alpha : \mathbb{N}_0 \times \mathbb{R}^+ \to \mathbb{R}_0^+$$

Definition 2.11. An *error function*, or just *error,* is a function with:

$$\eta : \mathbb{N}_0 \times \mathbb{R}^+ \to [0, 1]$$

Often we will see errors fulfilling

$$\eta(k, \varepsilon) \leq \eta(l, \varepsilon) \qquad \forall k \geq l,\, \varepsilon > 0$$

and we can interpret this by saying, that in the "long run" the properties at which the error applies will hold with ever increasing probability - i.e. its error decreases, the larger the time index k grows.

Definition 2.12. A flow A is *tail-bounded* by envelope α with error η, if for all $\varepsilon > 0$ and $n \geq m \in \mathbb{N}_0$ it holds that:

$$\mathbb{P}(A(n) - A(m) > \alpha(n-m, \varepsilon)) \leq \eta(n-m, \varepsilon).$$

Example 2.13. If (2.2) holds for all $\varepsilon > 0$, we have indeed a tail-bound. In practice one often formulates the tail-bound in such a way, that α is of a simple form (e.g. linear). Take, for example, the tail-bound

$$\mathbb{P}\left(A(n) - A(m) > r \cdot (n-m) - \frac{1}{d} \log\left(\frac{\varepsilon}{M}\right)\right) \leq \varepsilon = \eta(k, \varepsilon)$$

with $\eta(k, \varepsilon) = 1$ for all $\varepsilon > 1$. This can be reformulated, as for any $\epsilon > 0$ we can choose $\varepsilon := Me^{-d\epsilon}$ in aboves inequality and get:

$$\mathbb{P}\left(A(n) - A(m) > r \cdot (n-m) + \epsilon\right) \leq Me^{-d\epsilon} \tag{2.3}$$

[3] A common, in literature often found, choice for α is $\alpha(n-m) = r \cdot (n-m) - \frac{1}{d}\log(\frac{\varepsilon}{M})$ for some positive d and M.

Since ϵ was choosen arbitrary, the above holds for all $\epsilon > 0$. We can define now $\alpha'(n-m,\epsilon) = r \cdot (n-m) + \epsilon$. This expression describes the probability, that flow A exceeds a maximal rate of r (in the interval $[m,n]$), by at least ϵ. That probability in turn decreases exponentially with decay-rate d. A flow fulfilling (2.3) is said to be bounded by an EBB (exponentially bounded burstiness) envelope.

Another kind of bound involving (2.2) is obtained by using Chernoff's inequality.

Theorem 2.14. *Let X be a non-negative random variable and $x > 0$, then holds* Markov's inequality*:*

$$\mathbb{P}(X > x) \leq \frac{\mathbb{E}(X)}{x}$$

Let X be a real random variable and $x \in \mathbb{R}$, then holds for all $\theta > 0$ Chernoff's inequality*:*

$$\mathbb{P}(X > x) \leq e^{-\theta x}\phi_X(\theta)$$

Proof. Let f be the density[4] of X, then it holds for all $x \geq 0$

$$\mathbb{P}(X > x) = \int_x^\infty f(y)dy \leq \int_x^\infty f(y)\frac{y}{x}dy = \frac{1}{x}\int_x^\infty yf(y)dy \leq \frac{\mathbb{E}(X)}{x}$$

which proves Markov's inequality. Now let X be a real random variable, then $e^{\theta X}$ is a non-negative real random variable. Hence, by the monotonicity of the function $h(x) = e^{\theta x}$

$$\mathbb{P}(X > x) = \mathbb{P}(e^{\theta X} > e^{\theta x}) \leq \frac{\mathbb{E}(e^{\theta X})}{e^{\theta x}} = e^{-\theta x}\phi_X(\theta)$$

where we have used Markov's inequality. □

If we apply Chernoff's inequality in (2.2) we get:

$$\mathbb{P}(A(n) - A(m) > \alpha(n-m,\varepsilon)) \leq e^{-\theta\alpha(n-m,\varepsilon)}\phi_{A(n)-A(m)}(\theta)$$

Returning to example (2.13) this reads:

$$\mathbb{P}(A(n) - A(m) > r \cdot (n-m) + \epsilon) \leq e^{-\theta r\cdot(n-m)+\epsilon}\phi_{A(n)-A(m)}(\theta)$$

For further computations the expression $\phi_{A(n)-A(m)}(\theta)$ needs to be taken care of, which triggers the following definitions:

Definition 2.15. A flow A is $(\sigma(\theta),\rho(\theta))$-*bounded* for some $\theta > 0$, if for all $n \geq m \geq 0$ the MGF $\phi_{A(n)-A(m)}(\theta)$ exists and

$$\phi_{A(n)-A(m)}(\theta) \leq e^{\theta\rho(\theta)(n-m)+\theta\sigma(\theta)}$$

holds.

A flow A is $f(\theta,\cdot)$-*bounded* for some $\theta > 0$ if for all $n \geq m \geq 0$ the MGF $\phi_{A(n)-A(m)}(\theta)$ exists and

$$\phi_{A(n)-A(m)}(\theta) \leq f(\theta, n-m)$$

holds.

[4]We give here the proof for distributions with densities. The general proof is almost the same, but needs knowledge about Lebesgue integration.

Obviously $f(\theta,\cdot)$-boundedness is a generalisation of $(\sigma(\theta),\rho(\theta))$-boundedness[5].

Before we study a few examples on tail- and MGF-bounds, we proof a theorem, which shows, that each of them can be converted to the other one, if mild requirements are met. This theorem connects the two branches of stochastic network calculus and we will always keep it in mind, when we derive results in one of the two branches.

Theorem 2.16. *Assume a flow A. If A is tail-bounded by envelope α with error $\eta(k,\varepsilon) = \varepsilon$, it is also $f(\theta,\cdot)$-bounded with:*

$$f(\theta,\cdot) := \int_0^1 e^{\theta\alpha(n-m,\varepsilon)}d\varepsilon$$

Conversely if A is $f(\theta,\cdot)$-bounded it is also tail-bounded with $\alpha(n-m,\varepsilon) = \varepsilon$ and $\eta(n-m,\varepsilon) = f(\theta,n-m)e^{-\theta\varepsilon}$.

Proof. We proof the first part under the assumption that for $A(n)-A(m)$ we have a density function $f_{m,n}$ (remember example 1.8). The more general version needs the notion of Lebesgue's integral and - since it works out in the very same fashion - is left out. Denote by $F_{m,n}(x) = \int_0^x f_{m,n}(t)dt$ the cumulative distribution function of $A(n)-A(m)$ and by G the inverse function of $1-F_{m,n}$. We have then for every $\varepsilon \in (0,1]$:

$$\mathbb{P}(A(n)-A(m) > G(\varepsilon)) = 1-\mathbb{P}(A(n)-A(m) \leq G(\varepsilon)) = 1-F(G(\varepsilon)) = \varepsilon$$

On the other hand, from the definition of the tail-bound we have:

$$\mathbb{P}(A(n)-A(m) > \alpha(n-m,\varepsilon)) < \varepsilon = \mathbb{P}(A(n)-A(m) > G(\varepsilon))$$

We can read this as follows: For some value ε the probability of $A(n)-A(m)$ being larger than $G(\varepsilon)$ is larger than the probability of $A(n)-A(m)$ being larger than $\alpha(n-m,\varepsilon)$. This implies:

$$G(\varepsilon) < \alpha(n-m,\varepsilon)$$

Writing the MGF of $A(n)-A(m)$ with the help of $f_{m,n}$ reads:

$$\phi_{A(n)-A(m)}(\theta) = \int_0^\infty e^{\theta x} f_{m,n}(x)dx$$

We substitute the variable x by $G(\varepsilon)$ and multiply by the formal expression $\frac{d\varepsilon}{d\varepsilon}$ (note that $\lim_{a\to 0} 1-F_{m,n}(a) = 1$ and $\lim_{a\to\infty} 1-F_{m,n}(a) = 0$):

$$\begin{aligned}
\phi_{A(n)-A(m)}(\theta) &= \int_1^0 e^{\theta G(\varepsilon)} f_{m,n}(G(\varepsilon))\frac{dG(\varepsilon)}{d\varepsilon}d\varepsilon \\
&= -\int_0^1 e^{\theta G(\varepsilon)} f_{m,n}(G(\varepsilon))\frac{1}{-f_{m,n}(G(\varepsilon))}d\varepsilon \\
&= \int_0^1 e^{\theta G(\varepsilon)}d\varepsilon \leq \int_0^1 e^{\theta\alpha(n-m,\varepsilon)}d\varepsilon
\end{aligned}$$

[5]The parameter $\rho(\theta)$ corresponds to the *effective bandwidth* of the flow A. This means if A is $(\sigma(\theta),\rho(\theta))$-bounded we have that $\rho(\theta) \geq \mathbb{E}(a(i))$ and if $a(i)$ is bounded by c, we further have $\rho(\theta) \leq c$ (or can improve to a $(\sigma(\theta),c)$-bound). In words: $\rho(\theta)$ lies between the average rate and the peak rate of the arrivals. (see also Lemma 7.2.3 and 7.2.6 in [?].)

We have used the rule of differentation for inverse functions in the transition from first to second line ($(F^{-1})' = \frac{1}{F'(F^{-1})}$).

For the second part of the theorem we use Chernoff's inequality:

$$\mathbb{P}(A(n) - A(m) > \varepsilon) \leq \phi_{A(n)-A(m)}(\theta)e^{-\theta\varepsilon} \leq f(\theta, n-m)e^{-\theta\varepsilon}$$

□

Next we show how MGF-bounds and tail-bounds can be derived for a given flow.

Example 2.17. Let the increments $a(n)$ of some flow A be i.i.d. exponentially distributed to the parameter λ. This means that our random variables $A(n) - A(m)$ are Erlang distributed with parameters $n - m$ and λ and we know from example 1.35

$$\phi_{A(n)-A(m)}(\theta) = \left(\frac{\lambda}{\lambda - \theta}\right)^{n-m}$$

for all $\theta < \lambda$. Hence A is $(\sigma(\theta), \rho(\theta))$-bounded for all $\theta < \lambda$ with $\sigma(\theta) = 0$ and $\rho(\theta) = \frac{1}{\theta}\log\left(\frac{\lambda}{\lambda-\theta}\right)$.

Example 2.18. Let the increments of a flow A be i.i.d. Bernoulli with parameter p. We have then for all $\theta > 0$:

$$\begin{aligned}\phi_{A(n)-A(m)}(\theta) &= (\phi_{a(i)}(\theta))^{n-m} = (\mathbb{E}(e^{\theta a(i)}))^{n-m} \\ &= \left(p \cdot e^{\theta\cdot 0} + (1-p)\cdot e^{\theta\cdot 1}\right)^{n-m} = (p + (1-p)e^{\theta})^{n-m}\end{aligned}$$

Hence A is $f(\theta, n)$ bounded with $f(\theta, n) = (p + (1-p)e^{\theta})^n$ for all $\theta > 0$.

Example 2.19. Let the increments of a flow A be i.i.d. Pareto distributed with parameters x_m and α. We know already that the MGF of the Pareto distribution does not exist for any $\theta > 0$, hence no MGF-bound can be found for this flow.

Example 2.20. Assume the increments of a flow to be i.i.d. with bounded variance $\mathrm{Var}(a(n)) \leq \sigma^2 < \infty$. Note that $\mathbb{E}(A(n) - A(m)) = (n-m)\mathbb{E}(a(1))$ and hence $\mathbb{E}(a(1)) = \frac{\mathbb{E}(A(n)-A(m))}{n-m}$. Using the Chebyshev inequality[6] a tail-bound is constructed by:

$$\mathbb{P}(A(n)-A(m) > (n-m)(\mathbb{E}(a(n))+\varepsilon)) \leq \mathbb{P}\left(\left|\frac{A(n)-A(m)}{n-m} - \mathbb{E}(a(n))\right| > \varepsilon\right) \leq \frac{1}{\varepsilon^2(n-m)}\sigma^2$$

In the following example, we see how a multiplexed flow can be bounded, if we already have bounds for the single flows.

[6] Chebyshev's inequality states that for a random variable X with finite variance it holds that: $\mathbb{P}(X - \mathbb{E}(X) > \varepsilon) \leq \varepsilon^{-2}\mathrm{Var}(X)$.

Lemma 2.21. *Let A and B be two flows, which are $(\sigma_A(\theta), \rho_A(\theta))$- and $(\sigma_B(\theta), \rho_B(\theta))$-bounded, respectively (for the same $\theta > 0$). Further let A and B be stochastically independent. We call $A \oplus B(n) := A(n) + B(n)$ the multiplexed flow[7]. We have then:*

$$\begin{aligned}\phi_{A\oplus B(n)-A\oplus B(m)}(\theta) &= \mathbb{E}(e^{\theta(A(n)-A(m)+B(n)-B(m))}) = \mathbb{E}(e^{\theta(A(n)-A(m))})\mathbb{E}(e^{\theta(B(n)-B(m))}) \\ &\leq e^{\theta\rho_A(\theta)(n-m)+\theta\sigma_A(\theta)}e^{\theta\rho_B(\theta)(n-m)+\theta\sigma_B(\theta)} \\ &= e^{\theta(\rho_A(\theta)+\rho_B(\theta))(n-m)+\theta(\sigma_A(\theta)+\sigma_B(\theta))}\end{aligned}$$

Hence $A \oplus B$ is bounded by $(\sigma_A(\theta) + \sigma_B(\theta), \rho_A(\theta) + \rho_B(\theta))$.

The last lemma raises the question what to do, when the flows A and B are not independent. The key for this problem is to take care of expressions of the form $\mathbb{E}(XY)$. The next lemma gives us a way to deal with that.

Lemma 2.22. *Let X and Y be two non-negative real random variables with finite second moment (i.e. $\mathbb{E}(X^2), \mathbb{E}(Y^2) < \infty$). Then holds the* Cauchy-Schwartz inequality*:*

$$\mathbb{E}(XY) \leq (\mathbb{E}(X^2))^{1/2}(\mathbb{E}(Y^2))^{1/2}$$

Let $p, q \in \mathbb{R}^+$ such that $\frac{1}{p} + \frac{1}{q} = 1$ and assume the p-th moment of X and the q-th moment of Y to be finite. Then holds Hölder's inequality*:*

$$\mathbb{E}(XY) \leq (\mathbb{E}(X^p))^{1/p}(\mathbb{E}(Y^q))^{1/q}$$

For the proof we need Lebesgue-integration, which would lead us too far away from our course. Hence a proof is omitted (see for example [**?**]). Let us resume now to the situation of example 2.21 with the difference, that the two flows are not stochastically independent. Using the above lemma we can achieve that $A \oplus B$ is $(\sigma_A(p\theta) + \sigma_B(q\theta), \rho_A(p\theta) + \rho_B(q\theta))$-bounded. But note that we need the conditional assumptions of A being $(\sigma_A(p\theta), \rho_A(p\theta))$-bounded and B being $(\sigma_B(q\theta), \rho_B(q\theta))$-bounded.

Exercises

Exercise 2.23. Prove corollary 2.9.

Exercise 2.24. Let A be some flow, which has a deterministic arrival curve α. What tail-bound and MGF-bound can be given? Let now A be some flow, which has a token bucket arrival curve $\alpha(n) = r \cdot n + B$, show that A is $(\sigma(\theta), \rho(\theta))$-bounded and determine $\sigma(\theta)$ and $\rho(\theta)$.

Exercise 2.25. Assume a flow A has the following properties: All increments $a(n)$ with $n \geq 2$ are i.i.d. uniformly distributed on the interval $[0, b]$ (this means, it has the density $f(x) = \frac{1}{b}$ for all $x \in [0, b]$ and $f(x) = 0$ elsewhere). The first increment $a(1)$ however is exponentially distributed with parameter λ. Give a $(\sigma(\theta), \rho(\theta))$-bound for A.

(Hint: Calculate first the MGF of $A(n) - A(m)$ for $m \neq 0$ and then for $m = 0$. Find then a bound which covers both cases)

[7] You can convince yourself easily that $A \oplus B$ is a flow.

Exercise 2.26. Assume a flow A is $(\sigma(\theta), \rho(\theta))$-bounded for all $\theta \in [0, b)$ with $\sigma(\theta) = \sigma$ and $\rho(\theta) = \rho$ for all θ and some $b > 0$. Show that the expected arrivals in some time interval $[m+1, n]$ is upper bounded by $\rho \cdot (n-m) + \sigma$, i.e.:

$$\mathbb{E}(A(n) - A(m)) \leq \rho \cdot (n-m) + \sigma \qquad \forall\, n \geq m \geq 0$$

(Hint: Show first that $\phi_{A(n)-A(m)}(0) = \lim_{\theta \searrow 0} e^{\theta\rho(\theta)(n-m)+\theta\sigma(\theta)} = 1$, then conclude from the $(\sigma(\theta), \rho(\theta))$-boundedness that the first derivative of $\phi_{A(n)-A(m)}$ at point 0 must be smaller or equal to $\frac{d}{d\theta} e^{\theta\rho(\theta)(n-m)+\theta\sigma(\theta)}\big|_{\theta=0}$ (why?). Then use exercise 1.42)

Exercise 2.27. Assume a flow A is $(\sigma(\theta), \rho(\theta))$-bounded for all $\theta \in [0, b)$ and some $b > 0$. Further assume the following conditions on the bound:

$$\begin{aligned} \theta\rho(\theta) &\xrightarrow{\theta\to\infty} 0 \\ \theta\sigma(\theta) &\xrightarrow{\theta\to\infty} 0 \\ \theta\frac{d}{d\theta}\rho(\theta) &\xrightarrow{\theta\to\infty} 0 \\ \theta\frac{d}{d\theta}\sigma(\theta) &\xrightarrow{\theta\to\infty} 0 \end{aligned}$$

Show that $\mathbb{E}(A(n) - A(m)) \leq \rho(0)(n-m) + \sigma(0)$ (Hint: Proceed as in the previous exercise).

Prove that the conditions are met if the increments are i.i.d. exponentially distributed with parameter λ. Apply this knowledge to recover $\mathbb{E}(A(n) - A(m)) \leq \left(\frac{1}{\lambda}\right)^{n-m}$.

Exercise 2.28. Let there be $J \in \mathbb{N}$ stochastically independent flows A_j $(j = 1, \ldots, J)$, all of them having i.i.d. Bernoulli distributed increments $a_j(i)$ with the same parameter p. One can easily see, that the number of flows sending a paket (i.e. $a_j(n) = 1$) at a certain time n follows a binomial distribution, with parameters p and J:

$$\mathbb{P}\left(\sum_{j=1}^{J} a_j(n) = k\right) = \binom{J}{k} p^k (1-p)^{J-k} \qquad \forall\, n \in \mathbb{N}$$

Hence we can think of the multiplexed flow $A(n) := \sum_{j=1}^{J} a_j(n)$ to be binomially distributed. Give a $(\sigma(\theta), \rho(\theta))$-bound for the multiplexed flow using that it is binomially distributed.

Now use instead the multiplexing property of $(\sigma(\theta), \rho(\theta))$-bounds, as seen in example 2.21.

Exercise 2.29. Next assume two flows A and B. We have that the increments of the first flow $a(i)$ are Bernoulli distributed with parameter p, further we assume that for the increments of B holds $b(i) = 1 - a(i)$. Convince yourself, that the increments of B are again Bernoulli distributed, but with the parameter $1 - p$. Of what form is the flow $A \oplus B$? Give a $(\sigma(\theta), \rho(\theta))$-bound for $A \oplus B$ (Hint: Use exercise 2.24)! Next use the multiplexing property of $(\sigma(\theta), \rho(\theta))$-bounds (with Hölder) to bound the flow $A \oplus B$. Which bound is better? (Do not try to solve this analytically!)

Exercise 2.30. Remember the Cauchy-Schwarz inequality, as you know it from inner product spaces. An inner product space consists of a vector space V (usually $\mathbb{R}^d$ or $\mathbb{C}^d$) a field of scalars F (usually $\mathbb{R}$ or $\mathbb{C}$) and an inner product (dt.: Skalarprodukt)

$$\langle \cdot, \cdot \rangle : V \times V \to F$$

fulfilling:

- (Conjugate) symmetry: $\langle x, y \rangle = \overline{\langle x, y \rangle}$
- Linearity in the first argument: $\langle a \cdot x, y \rangle = a \cdot \langle x, y \rangle$ and $\langle x + y, z \rangle = \langle x, z \rangle + \langle y, z \rangle$
- Positive-definiteness: $\langle x, x \rangle \geq 0$ with equality only for $x = 0$.

Check that $\langle X, Y \rangle = \mathbb{E}(X \cdot Y)$ defines an inner product "on the space of one dimensional real random variables"

For inner product spaces one can define a norm by:

$$\|x\| := \sqrt{\langle x, x \rangle}$$

Convince yourself that the Cauchy-Schwarz inequality as given in lemma 2.22 is in fact the Cauchy-Schwarz inequality for inner product spaces:

$$|\langle x, y \rangle|^2 \leq \|x\| \cdot \|y\|$$

(with equality if x and y are linearly(!) independent)

Exercise 2.31. Prove the bound for stochastically dependent multiplexed flows (as given after lemma 2.22)!

Exercise 2.32. We now investigate the parameter θ in the $(\sigma(\theta), \rho(\theta))$-bounds, which we will call *acuity* later on. Generate some (in the order of thousands) exponentially distributed random variables with parameter $\lambda = 10$ (you might reuse your code from exercise 1.20). Sort your realisations by magnitude and plot them. Now apply the function $f(x) = e^{\theta x}$ to your realisations with varying parameter $\theta \in [0, 10)$. How does the plot change for different values of θ? You should be able to see, that for large θ one has a few very large results, compared to the majority of smaller results. For smaller θ, however, one gets a more balanced picture.

We ask now how many of the realisations are larger than the expected value of one of the realisations $\mathbb{E}(e^{\theta X})$. For this we already know from example 1.25 that $\mathbb{E}(e^{\theta X}) = \frac{\lambda}{\lambda - \theta}$. Write a procedure, which counts for a fixed θ the number of realisations being larger $\frac{\lambda}{\lambda - \theta}$. What percentage of your realisations are larger for a small θ? What is it for a large θ? Explain the difference with the help of your previously produced plots!

2.2 Stochastic Service

In this chapter, we first discuss different alternatives for defining service guarantees in a determinstic context. Then we introduce the definition of stochastic service models before presenting the important results on how to concatenate stochastic servers in a network setting.

2.2.1 The Different Notions of Deterministic Service Curves

In this section, we make a short excursion into deterministic service curves and strict service curves. We talk about the differences between those two and generalize the concept of a service curve to fit our needs. For a more detailed survey on the zoo of different service guarantees [**?**] is a great point to start from.

We motivate stochastic service similarly to the stochastic arrivals and see service elements, which offer certainly *some* amount of service, but only achieve a deterministic lower bound equal to *zero*.

Throughout the rest of this chapter we consider a node working on an arrival flow A and denote the departure flow by D. Remember that in a backlogged period $[m, n]$ it holds that $A(k) > D(k)$ for all $k \in [m, n]$.

Definition 2.33. A service element offers a *strict service curve* $\beta : \mathbb{N}_0 \to \mathbb{R}_0^+$, if for all backlogged periods $[m, n]$ and all input flows A it holds that:

$$D(n) - D(m) \geq \beta(n - m).$$

Another definition for strict service curves, which incorporates the arrival flow A, is given by the following lemma:

Lemma 2.34. *Let $m \in \mathbb{N}_0$ be arbitrary and denote by $n^*(m)$ the beginning of the next(!) backlogged period. The service element offers a strict service curve β if and only if:*

$$D(n) \geq \beta(n - m) + D(m) \wedge A(n) \tag{2.1}$$

for all $n < n^*(m)$ and all input flows A.

Proof. Let us first assume S offers a strict service curve β and $D(m) = A(m)$, then for all $n < n^*(m)$ it also holds that $D(n) = A(n)$, and thus:

$$D(n) \geq A(n) \geq \beta(n - m) + D(m) \wedge A(n)$$

Now assume $D(m) < A(m)$, i.e. we start in a backlogged period. If also $D(n) < A(n)$ holds, then $[m, n]$ is a backlogged period and we get by our assumption $D(n) - D(m) \geq \beta(n - m)$. The case $D(n) = A(n)$ was already treated, again we have:

$$D(n) \geq A(n) \geq \beta(n - m) + D(m) \wedge A(n).$$

Assume now the service element fulfills equation (2.1), we need to show β is a strict service curve. Let $[m, n]$ be an arbitrary backlogged period, then surely $n < n^*(m)$. Assume $\beta(n - m) + D(m) > A(n)$, then by using equation (2.1) we would obtain:

$$D(n) \geq \beta(n - m) + D(m) \wedge A(n) = A(n),$$

which would imply $D(n) = A(n)$, which is a contradiction to our assumption that $[m, n]$ is a backlogged period. Hence $\beta(n - m) + D(m) \leq A(n)$ and again by equation (2.1):

$$D(n) \geq \beta(n - m) + D(m) \wedge A(n) = \beta(n - m) + D(m)$$

from which it follows that:

$$D(n) - D(m) \geq \beta(n - m).$$

□

There exists also the more general (non-strict) service curve[8].

Definition 2.35. A service element offers a *service curve* $\beta : \mathbb{N}_0 \to \mathbb{R}_0^+$ if for all $n \geq m \in \mathbb{N}_0$ and all input flows A it holds that:

$$D(n) \geq \min_{0 \leq k \leq n} \{A(k) + \beta(n - k)\}.$$

One might ask: What is the difference between a service curve and a strict service curve? It is easy to find examples, in which the departures of some node differ, depending if one uses a service curve or a strict service curve (you are asked to construct one in the exercises). However, we want to shed a bit more light on the differences between service curve definitions. Therefore we introduce Lindley's equation.

Definition 2.36. A service element fulfills *Lindley's equation* if for all $q(n) := A(n) - D(n)$ and all input flows A it holds that:

$$q(n+1) = [q(n) + a(n+1) - s(n+1)]^+,$$

where $s(n)$ is the amount of data the service element can process at time n.

Lindley's equation states that the amount of data, which is queued at the service element, is the amount of data, which has been queued the time step before plus the amount of data which arrives in the meantime minus the amount of data the service element can process in that timestep. Of course, it is possible that the service element could serve more data than have been queued and newly arrived, hence we have to take the maximum with zero in Lindley's equation. This behaviour basically means that our service element is not "lazy" and always works as much as it can, i.e., it is work-conserving. The next lemma shows, that under this behaviour, a service curve becomes a strict service curve (and more).

Lemma 2.37. *Let a service element fulfill Lindley's equation. The following three items are equivalent:*

- *It offers a service curve β.*
- *It holds $S(n) - S(k) \geq \beta(n - k)$ for all $n \geq k \in \mathbb{N}_0$.*
- *It offers a strict service curve β.*

[8] The advantage of service curves over strict service curves is, that we can convolute service elements offering service curves into a single service element offering again a service curve. Such a property does in general not hold for strict service curves (see exercises).

Here $S(n)$ is defined by $S(n) := \sum_{k=0}^{n} s(k)$ with $s(k)$ from Lindley's equation.

Proof. We first show that from the first item follows the second, from which in turn follows the third. In the last step we show, that the third item implies again the first.

(1)⇒(2):

First we claim that

$$q(n) = \max_{0 \le k \le n} \{A(n) - A(k) - S(n) + S(k)\} \tag{2.2}$$

and prove this by induction over n: For $n = 1$, we have by Lindley's equation (assuming $q(0) = a(0) = s(0) = 0$):

$$q(1) = [a(1) - s(1)]^+ = \max_{0 \le k \le 1} \{A(1) - A(k) - S(1) + S(k)\}$$

Now assume (2.2) holds for n, then we have again using Lindley's equation:

$$\begin{aligned} q(n+1) &= [q(n) + a(n+1) - s(n+1)]^+ \\ &= [\max_{0 \le k \le n} \{A(n) - A(k) - S(n) + S(k)\} + a(n+1) - s(n+1)]^+ \\ &= [\max_{0 \le k \le n} \{A(n+1) - A(k) - S(n+1) + S(k)\}]^+ \\ &= \max_{0 \le k \le n+1} \{A(n+1) - A(k) - S(n+1) + S(k)\}. \end{aligned}$$

Assume now S offers a service curve β and let $n, k \in \mathbb{N}_0$ be arbitrary such that $n \ge k$. From (2.2) we know already that for all flows A

$$q(n) = A(n) - D(n) \ge A(n) - A(k) - S(n) + S(k).$$

And by the definition of service curves:

$$D(n) \ge \min_{0 \le k' \le n} \{A(k') + \beta(n - k')\}$$

Combining these, we obtain:

$$\begin{aligned} S(n) - S(k) &\ge D(n) - A(k) \\ &\ge \min_{0 \le k' \le n} \{A(k') - A(k) + \beta(n - k')\} && (2.3) \\ &= \min_{0 \le k' \le k} \{A(k') - A(k) + \beta(n - k')\} \wedge && (2.4) \\ &\quad \min_{k < k' \le n} \{A(k') - A(k) + \beta(n - k')\} \end{aligned}$$

Now use the flow A^*, defined by:

$$\begin{aligned} a^*(1) &\in \mathbb{R}_0^+ \\ a^*(2) = a^*(3) &= \ldots = a^*(k) = 0 \end{aligned}$$

$$\begin{aligned} a^*(k+1) &= \beta(n-k) - \beta(n-k-1) \\ a^*(k+2) &= \beta(n-k-1) - \beta(n-k-2) \\ a^*(k+3) &= \beta(n-k-2) - \beta(n-k-3) \\ &\vdots \\ a^*(n) &= \beta(1) - \beta(0) = \beta(1) \end{aligned}$$

Note that by this construction we have $A^*(k') = A^*(k)$ for all $k' \leq k$. Inserting that flow into (2.4) we eventually have:

$$\begin{aligned} S(n) - S(k) &\geq \min_{0 \leq k' \leq k} \{0 + \beta(n-k')\} \wedge \min_{k < k' \leq n} \left\{ \beta(n-k') + \sum_{l=k+1}^{k'} a^*(l) \right\} \\ &= \beta(n-k) \wedge \min_{k < k' \leq n} \left\{ \beta(n-k') + \sum_{l=k+1}^{k'} \beta(n-l+1) - \beta(n-l) \right\} \\ &= \beta(n-k) \wedge \min_{k < k' \leq n} \{\beta(n-k') + \beta(n-k) - \beta(n-k'\} = \beta(n-k) \end{aligned}$$

for all $n \geq k \in \mathbb{N}_0$.

Hence the first item implies second item.

(2)⇒(3)

If we assume the second item, we have by Lindley's equation for each k in an arbitrary backlogged period $[m, n]$

$$\begin{aligned} A(k) - D(k) = q(k) &= [q(k-1) + a(k) - s(k)]^+ \\ &= q(k-1) + a(k) - s(k) = \\ &= A(k-1) - D(k-1) + a(k) - s(k) \\ &= A(k) - D(k) + d(k) - s(k) \end{aligned}$$

and hence $d(k) = s(k)$ for all $k \in [m, n]$. Using the second item, it follows:

$$D(n) - D(m) = \sum_{k=m+1}^{n} d(k) = \sum_{k=m+1}^{n} s(k) = S(n) - S(m) \geq \beta(n-m)$$

Since we have chosen $[m, n]$ to be an arbitrary backlogged period, we have shown that β is a strict service curve. It is left to show, that from the third item follows again the first item.

(3)⇒(1)

Let n be arbitrary and consider first the case of $D(n) < A(n)$, i.e. n lies in a backlog period and define m^* by the last time before n with $A(m^*) = D(m^*)$. It holds then:

$$\beta(n - m^*) \leq D(n) - D(m^*) = D(n) - A(m^*)$$

Hence:

$$D(n) \geq \beta(n - m^*) + A(m^*) \geq \min_{0 \leq k \leq n} \{A(k) + \beta(n - k)\}$$

Consider now the case of $D(n) = A(n)$, then follows immediately:

$$D(n) = A(n) \geq \min_{0 \leq k \leq n} \{A(k) + \beta(n - k)\}$$

□

Notice that a strict service curve still gives the service element some space to be "lazy", since it only has to work as much as needed to fulfill the constraint given by β. If there is more service available even the strict service curve does not necessarily claim it. Following this train of thoughts one can sort servers by their "lazyness" in the following increasing order:

- A server fulfills Lindley's equation.
- A server offers a strict service curve.
- A server offers a service curve.

However keep in mind, that Lindley's equation does not hold any information about *how much* data the service element processes. In a certain sense one can see the sum of Lindley's equation and a (strict) service curve, as the "strictest" service guarantee possible. Since it gives us information about how much service is available at least ($S(n) - S(m) \geq \beta(n - m)$) and also states that the service element is never lazy, i.e., we always use the whole available service (as seen in the above proof it holds under Lindley's equation: $d(k) = s(k)$ in each backlogged period). This is exactly the second item in the previous lemma. What you should keep in mind is, that if some service element fulfills Lindley's equation, the three kind of service curves immediately fall together.

2.2.2 Definition of Stochastic Service Models

As announced we encounter the same problems with stochastic service, as for stochastic arrivals. As an example consider a service element which offers a constant rate of service (denoted by c) and Lindley's equation. There are two flows entering, which we call A_1 and A_2 and we set A_1 as prioritized flow. This means, that for the second flow only the amount of service, which is leftover by A_1 is offered. We further assume, that A_1 does not queue. This means if A_1 sends in one time step an amount of data larger c, we drop all exceeding data. If we consider one time step and denote by $s_l(n)$ the amount of service A_2 receives in this time step we have $s_l(n) = [c - a_1(n)]^+$(Note: If A_1 queued we would have $s_l(n) = [c - a_1(n) - q_1(n-1)]^+$, where $q_1(n-1)$ would be the queue length

of A_1 at time $n-1$). Now we ask for some service curve β, which the server might offer for A_2. Depending on the arrivals, this might not be possible. If we take for example i.i.d. exponentially distributed increments $a_1(n)$ we get:

$$\mathbb{P}(s_l(n) = 0) = \mathbb{P}([c - a_1(n)]^+ = 0) = \mathbb{P}(a_i(n) \geq c) = e^{-\lambda c} > 0$$

Hence the best service curve we can expect is $\beta(n) = 0$ for all $n \in \mathbb{N}_0$.

Before we give a method to bound stochastic service elements, we need a generalisation of the above concepts.

Definition 2.38. Define the index set $\Lambda(\mathbb{N}_0) := \{(i,j) \in \mathbb{N}_0 \times \mathbb{N}_0 : i \leq j\}$ triangular subset of the lattice $\mathbb{N}_0 \times \mathbb{N}_0$. We call a set indexed by $\Lambda(\mathbb{N}_0)$ a *triangular array* (over $\mathbb{R}$) and denote it to be an element of $\mathcal{D}$.

We can list the indices of $\Lambda(\mathbb{N}_0)$ by:

$$\begin{array}{llll} (0,0); & & & \\ (0,1); & (1,1); & & \\ (0,2); & (1,2); & (2,2); & \\ (0,3); & \cdots & \cdots & (3,3); \\ \vdots & & & \ddots \end{array}$$

The set of such doubly indexed stochastic processes is denoted by $\mathcal{S}_2$.

Definition 2.39. If A is a flow we define the doubly indexed stochastic process $A(\cdot,\cdot) : \Lambda(\mathbb{N}_0) \to \mathbb{R}_0^+$ by:

$$A(m,n) := A(n) - A(m) \qquad \forall\, (m,n) \in \Lambda(\mathbb{N}_0)$$

The set of such doubly indexed stochastic processes *resulting from flows* is denoted by $\mathcal{F}_2$.

Clearly $\mathcal{F}_2 \subsetneq \mathcal{S}_2 \subset \mathcal{D}$. It is important to note the difference between $A(m,n)$ and $A(n-m)$. In the former case we talk about the arrivals in the interval $[m+1, n]$ (an interval of length $n-m$), in contrast to the latter case, in which we consider the arrivals up to time $n-m$. We also need a min-plus-convolution and a min-plus-deconvolution for the bivariate case. Remember and compare the univariate operations:

$$A \otimes B(n) = \min_{0 \leq k \leq n} \{A(k) + B(n-k)\}$$
$$A \oslash B(n) = \max_{k \geq 0} \{A(n+k) - B(k)\}.$$

Definition 2.40. Let $A, B \in \mathcal{D}$. Then the *bivariate convolution* $\otimes : \mathcal{D} \times \mathcal{D} \to \mathcal{D}$ is defined by:

$$A \otimes B(m,n) := \min_{m \leq k \leq n} \{A(m,k) + B(k,n)\}$$

Further the *bivariate deconvolution* $\oslash : \mathcal{D} \times \mathcal{D} \to \mathcal{D}$ is defined by:

$$A \oslash B(m,n) := \max_{0 \leq k \leq m} \{A(k,n) - B(k,m)\}$$

One should mention that this bivariate operations are no longer commutative. You are further asked in the exercises, to show that the bivariate definitions are a generalization of the univariate case. We now present the bivariate "service curve".

Definition 2.41. Assume a service element has a flow A as input and the output is denoted by D. Let $S \in \mathcal{S}_2$ be a doubly indexed stochastic process with

$$S(m,n) \leq S(m,n') \quad \forall\, (m,n),\, (m,n'),\, (n,n') \in \Lambda(\mathbb{N}_0)\,\text{a.s.}$$

and

$$S(n,n) = 0 \qquad \forall\, n \in \mathbb{N}_0\ \text{a.s.}$$

The service element is a *dynamic S-server*[9] if for all $n \geq 0$ it holds that:

$$D(0,n) \geq A \otimes S(0,n) \qquad (\text{a.s.})$$

Note that the process $S(m,n)$ does generally *not* lie in $\mathcal{F}_2$ and hence we cannot assume $S(m,n) = S(n) - S(m)$, although we encounter situations in which this equation is met.

Example 2.42. We first consider a deterministic rate-latency server, with service curve $\beta_{R,N}(n) = R \cdot (n-N)$ if $n > N$ and equal to zero otherwise. For $S(k,n) := \beta_{R,N}(n-k)$ we have that a dynamic S-server offers $\beta_{R,N}$ as service curve in the univariate sense:

$$D(n) \geq A \otimes S(0,n) = \min_{0 \leq k \leq n} \{A(k) + S(k,n)\} = \min_{0 \leq k \leq n} \{A(k) + \beta_{R,N}(n-k)\} = A \otimes \beta_{R,N}(n)$$

Example 2.43. We return to our example of a service element satisfying Lindley's equation and serving two flows A_1 and A_2, of which the first is prioritized. But we now allow that flow A_1 can queue. As before the service element offers a constant rate of c. In this case, the service element is a dynamic S-server for A_2 with:

$$S(m,n) := [c \cdot (n-m) - A_1(m,n)]^+$$

To see this assume $n \geq 0$ arbitrary and choose $m \leq n$ maximal such that $A_1(m) = D_1(m)$ and $A_2(m) = D_2(m)$ (such an m exists since all flows are zero for $m = 0$). Our server satisfies:

$$D_1(n) + D_2(n) = D_1(m) + D_2(m) + c \cdot (n-m) = A_1(m) + D_2(m) + c \cdot (n-m)$$

Since $D_1(n) \leq A_1(n)$ we can continue with:

$$D_2(n) \geq D_2(m) + c \cdot (n-m) - (A_1(0,n) - A_1(0,m))$$

We have for sure that $D_2(n) \geq D_2(m) = D_2(m) + 0$ and therefore obtain:

$$\begin{aligned} D_2(n) &\geq D_2(m) + [c \cdot (n-m) - A_1(m,n)]^+ = A_2(m) + [c \cdot (n-m) - A_1(m,n)]^+ \\ &\geq \min_{0 \leq k \leq n} \{A_2(k) + S(k,n)\} = A_2 \otimes S(0,n) \end{aligned}$$

[9] *In literature there is often the additional assumption of $S(m,n) \geq 0$ for all m and n. This assumption is not necessary and in some situations burdens us with unwanted restrictions. Note that, in fact, the case $S(m,n) < 0$ causes no trouble, since $D(0,n) \geq A \otimes S(0,n)$ represents a lower bound. The worst thing, which can happen here, is that the lower bound is useless (nevertheless correct): $D(0,n) \geq A \otimes S(0,n)$ with $A \otimes S(0,n) < 0$.*

Similar to the previous section, we next present two different ways to bound dynamic S-servers stochastically. Again, we have a tail-bound and an MGF-bound. There is a slight difference to the way we bounded arrivals, though, and we are going to investigate it in some detail. This time we start with the MGF-bound, which is straightforward:

Definition 2.44. A dynamic S-server is $(\sigma(\theta), \rho(\theta))$-*bounded* for some $\theta > 0$, if $\phi_{S(m,n)}(-\theta)$ exists and

$$\phi_{S(m,n)}(-\theta) \leq e^{\theta\rho(\theta)(n-m)+\theta\sigma(\theta)} \qquad \forall\,(m,n) \in \Lambda(\mathbb{N}_0).$$

A dynamic S-server is $f(\theta, n)$-*bounded* for some $\theta > 0$, if $\phi_{S(m,n)}(-\theta)$ exists and

$$\phi_{S(m,n)}(-\theta) \leq f(\theta, n-m) \qquad \forall\,(m,n) \in \Lambda(\mathbb{N}_0).$$

Note that usually $\rho(\theta)$ is a negative quantity. This basically parallels the definition of the MGF-bound for arrivals.

Next we present the tail-bound for dynamic S-servers:

Definition 2.45. A dynamic S-server is tail-bounded by envelope $\beta \geq 0$ with error ζ, if for all arrival flows A and all $n \in \mathbb{N}_0$, $\varepsilon \in \mathbb{R}^+$ holds:

$$\mathbb{P}(D(0,n) < A \otimes \beta(\varepsilon)(0,n)) < \zeta(n,\varepsilon) \tag{2.5}$$

In the above definition the expression $A \otimes \beta(\varepsilon)(0,n)$ has to be interpreted as: $\min_{0\leq k\leq n}\{A(k)+\beta(n-k,\varepsilon)\}$. The condition $\beta \geq 0$ is to be understood pointwisely. As mentioned, the above definition looks a bit different from the MGF-bound. Instead of bounding the function S directly, rather the output D of the dynamic S-server is bounded - and that for all possible inputs A.

Example 2.46. Consider the same situation as in example 2.43. We have already seen that this service element is an S_l-server with $S_l(m,n) = [c\cdot(n-m) - A_1(m,n)]^+$ with respect to A_2. If we assume the increments of the prioritized flow to be i.i.d. exponentially distributed with parameter λ, then we have for all $\theta \in (0,\lambda)$:

$$\begin{aligned}\phi_{S_l(m,n)}(-\theta) &= \left(\mathbb{E}(e^{-\theta[c\cdot(n-m)-A_1(m,n)]^+}\right) = \left(\mathbb{E}(e^{\min\{0,\theta A_1(m,n)-\theta c\cdot(n-m)\}})\right) \\ &\leq \left(\mathbb{E}(e^{\theta A_1(m,n)-\theta c\cdot(n-m)})\right) = e^{-\theta c(n-m)}\left(\frac{\lambda}{\lambda-\theta}\right)^{n-m}.\end{aligned}$$

Hence we have a dynamic S_l-Server, which is bounded by $f(\theta,n) = e^{-\theta cn}\left(\frac{\lambda}{\lambda-\theta}\right)^n$. We can also express this as a $(\sigma(\theta),\rho(\theta))$-bound with $\sigma(\theta) = 0$ and $\rho(\theta) = -c + 1/\theta \log\left(\frac{\lambda}{\lambda-\theta}\right)$.

Can we also find a tail-bound, given, that A_1 is tail-bounded by the envelope α? We need to find a fitting envelope β and some error function to establish (2.5) for all flows A_2. Inspired by the MGF-bound the envelope could look like $\beta(n,\varepsilon) = [cn - \alpha(n,\varepsilon)]^+$, which is in fact true. To see this choose $n \in \mathbb{N}$ and $\varepsilon > 0$ arbitrary. We know:

$$D_2(0,n) \geq A_2 \otimes S_l(0,n) = \min_{0\leq k\leq n}\{A_2(k) + [c(n-k) - A_1(k,n)]^+\}$$

Now assume for a while that $A_1(k,n) \leq \alpha(n-k,\varepsilon)$ holds for every $k \leq n$. Then the above can be continued with:

$$\begin{aligned} D_2(0,n) &\geq \min_{0\leq k\leq n} \{A_2(k) + [c(n-k) - \alpha(n-k,\varepsilon)]^+\} \\ &= A_2 \otimes \beta(\varepsilon)(0,n) \end{aligned}$$

So we have proven the following statement:

$$A_1(k,n) \leq \alpha(n-k,\varepsilon) \quad \forall\, k \leq n \quad \Rightarrow \quad D_2(0,n) \geq A_2 \otimes \beta(\varepsilon)(0,n)$$

Or more precisely, based on definition 2.1, we have:

$$\{\omega \in \Omega \,|\, \bigcap_{k=0}^{n} A_1(k,n) \leq \alpha(n-k,\varepsilon)\} \subset \{\omega \in \Omega \,|\, D_2(0,n) \geq A \otimes \beta(\varepsilon)(0,n)\}$$

and hence:

$$\mathbb{P}\left(\bigcap_{k=0}^{n} A_1(k,n) \leq \alpha(n-k,\varepsilon)\right) \leq \mathbb{P}(D_2(0,n) \geq A_2 \otimes \beta(\varepsilon)(0,n)$$

From which we eventually obtain

$$\mathbb{P}(D_2(0,n) < A \otimes \beta(\varepsilon)(0,n)) \leq \mathbb{P}(\sup_{0\leq k\leq n} A_1(k,n) - \alpha(n-k,\varepsilon) > 0) \leq \sum_{k=0}^{n} \eta(n-k,\varepsilon) =: \zeta(n,\varepsilon)$$

in the last inequality we have used $\mathbb{P}(A \cup B) \leq \mathbb{P}(A) + \mathbb{P}(B)$.

We see in the above example that constructing the tail-bound for leftover service is a bit harder, than the corresponding MGF-bounds. Fortunately, the generalization of the above example follows pretty much along the same lines, such that most of the work is done here already.

Theorem 2.47. *Assume a dynamic S-server fulfills Lindley's equation and serves two inputs A_1 and A_2, of which the former one is prioritized. Further assume A_1 to be tail-bounded with envelope α and error η. Then the dynamic S-server for A_2 is tail-bounded by envelope $[S-\alpha]^+$ and error $\zeta(n,\varepsilon) := \sum_{k=0}^{n} \eta(k,\varepsilon)$.*

Proof. Exercise 2.53. □

2.2.3 Concatenation of Stochastic Servers

Next we present the concatenation theorem. Imagine two servers in a tandem, i.e. the output of the first service element is fed into the second service element. The concatenation theorem states, that we can abstract the two service elements as a new service element, which describes the system as a whole.

Theorem 2.48. *Let there be two service elements, such that the output of the first serivce element is the input for the second service element. Assume the first element to be a dynamic S-server and the second element to be a dynamic T-server. Then holds:*

- *The whole system is a dynamic $S \otimes T$-server.*
- *If the dynamic servers are stochastically independent and bounded by $(\sigma_S(\theta), \rho_S(\theta))$ and $(\sigma_T(\theta), \rho_T(\theta))$, respectively, with $\rho_T(\theta), \rho_S(\theta) < 0$ and $\rho_S(\theta) \neq \rho_T(\theta)$ for some $\theta > 0$, the whole system is bounded by*

$$(\sigma_S(\theta) + \sigma_T(\theta) + B, \max\{\rho_T(\theta), \rho_S(\theta)\})$$

with:

$$B := -\frac{1}{\theta}\log(1 - e^{-\theta|\rho_S(\theta)-\rho_T(\theta)|})$$

- *If the dynamic servers are tailbounded by envelopes β and γ with errors ζ and κ, respectively. Further assume $\beta(n,\varepsilon)$ and $\gamma(n,\varepsilon)$ are monotone increasing in n. Then the whole system is tailbounded with envelope $\beta \otimes \gamma$ and error $\zeta_{sys}(n,\varepsilon) = \sum_{l=0}^{n} \zeta(l,\varepsilon) + \kappa(n,\varepsilon)$.*

Proof. We prove first that the whole system is a dynamic $S \otimes T$-server. Denote the input flow at the first service element by A, the output of the first service element by I and the output of the second service element by D. We have then for all $n \in \mathbb{N}_0$:

$$\begin{aligned}
D(0,n) &\geq I \otimes T(0,n) \geq (A \otimes S) \otimes T(0,n) = \min_{0\leq k\leq n}\{(A\otimes S)(0,k) + T(k,n)\} \\
&= \min_{0\leq k\leq n}\{\min_{0\leq l\leq k}\{A(0,l) + S(l,k)\} + T(k,n)\} \\
&= \min_{0\leq l\leq k\leq n}\{A(0,l) + S(l,k) + T(k,n)\} \\
&= \min_{0\leq l\leq n}\{A(0,l) + \min_{l\leq k\leq n}\{S(l,k) + T(k,n)\}\} \\
&= \min_{0\leq l\leq n}\{A(0,l) + S\otimes T(l,n)\} = A \otimes (S\otimes T)(0,n)
\end{aligned}$$

Hence the whole system is a dynamic $S \otimes T$-server.

Next we validate the MGF-bound: Assume first $|\rho_T(\theta)| < |\rho_S(\theta)|$. With the use of 3.2 holds for all $(m,n) \in \Lambda(\mathbb{N}_0)$:

$$\begin{aligned}
\phi_{S\otimes T(m,n)}(-\theta) &\leq \sum_{k=m}^{n} e^{\theta\rho_S(\theta)(k-m)+\theta\sigma_S(\theta)} e^{\theta\rho_T(\theta)(n-k)+\theta\sigma_T(\theta)} \\
&= e^{\theta\rho_T(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \sum_{k=m}^{n} e^{\theta\rho_S(\theta)(k-m)} e^{\theta\rho_T(\theta)(m-k)} \\
&= e^{\theta\rho_T(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \sum_{k=m}^{n} e^{\theta(k-m)(\rho_S(\theta)-\rho_T(\theta))} \\
&= e^{\theta\rho_T(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))} \sum_{k'=0}^{n-m} \left(e^{\theta(\rho_S(\theta)-\rho_T(\theta))}\right)^{k'}
\end{aligned}$$

and since we have $e^{\theta(\rho_S(\theta)-\rho_T(\theta))} < 1$ and can approximate the sum by its corresponding geometric series:

$$\phi_{S\otimes T(m,n)}(-\theta) \leq e^{\theta\rho_T(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))}\frac{1}{1-e^{\theta(\rho_S(\theta)-\rho_T(\theta))}}$$

The case of $|\rho_T(\theta)| > |\rho_S(\theta)|$ is very similar:

$$\begin{aligned}
\phi_{S\otimes T(m,n)}(-\theta) &\leq \sum_{k=m}^{n} e^{\theta\rho_S(\theta)(k-m)+\theta\sigma_S(\theta)}e^{\theta\rho_T(\theta)(n-k)+\theta\sigma_T(\theta)} \\
&= e^{\theta\rho_S(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))}\sum_{k=m}^{n} e^{\theta\rho_S(\theta)(k-n)}e^{\theta\rho_T(\theta)(n-k)} \\
&= e^{\theta\rho_S(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))}\sum_{k=m}^{n} e^{\theta(n-k)(\rho_T(\theta)-\rho_S(\theta))} \\
&= e^{\theta\rho_S(\theta)(n-m)+\theta(\sigma_S(\theta)+\sigma_T(\theta))}\sum_{k'=0}^{n-m} (e^{\theta(\rho_T(\theta)-\rho_S(\theta))})^{k'}
\end{aligned}$$

All left to do is to prove the tail-bound. Let $\varepsilon > 0$, $n \in \mathbb{N}_0$ be arbitrary and A some flow. Assume for a while that $I(l) \geq A\otimes\beta(\varepsilon)(0,l)$ for all $l \leq n$ and $D(n) \geq I\otimes\gamma(\varepsilon)(0,n)$. Then follows as above:

$$D(0,n) \geq A \otimes (\beta(\varepsilon) \otimes \gamma(\varepsilon))(0,n)$$

Define $\beta_{sys} = \beta \otimes \gamma$, we have then:

$$\begin{aligned}
&\mathbb{P}(D(0,n) < A \otimes \beta_{sys}(\varepsilon)(0,n) \\
\leq &\mathbb{P}\left(\bigcup_{l=0}^{n}\{I(0,l) < A \otimes \beta(\varepsilon)(0,l)\} \cup \{D(0,n) < I \otimes \gamma(\varepsilon)(0,n)\}\right) \\
\leq &\sum_{l=0}^{n}\mathbb{P}(I(0,n) < A \otimes \beta(\varepsilon)(0,n)) + \mathbb{P}(D(0,n) < I \otimes \gamma(\varepsilon)(0,n)) \\
\leq &\sum_{l=0}^{n}\zeta(l,\varepsilon) + \kappa(n,\varepsilon) =: \zeta_{sys}(n,\varepsilon)
\end{aligned}$$

□

We want to give some notes about the above theorem, since some obstacles may rise at this stage. First: note that for the MGF-bound we need both performance bounds to exist for the same θ. Since the service bounds are usually valid for all θ in some interval $(0,b)$, we can ensure, by intersecting the corresponding intervals, to find an interval on which both performance bounds hold. Second: The assumption of stochastic independence may not be given (see the exercises). The trick is to use again Hölder's inequality, however this may - as in exercise 2.29 - lead to poorer bounds. Third: We used $\rho_S(\theta) \neq \rho_T(\theta)$. If we have $\rho_S(\theta) = \rho_T(\theta)$ there is no way to use adavantage of the

geometric series to achieve a $(\sigma(\theta), \rho(\theta))$-bound. Instead the sum degenerates to $n-m+1$ and we need the more general $f(\theta, n)$-bounds.

The tail-bound seems to cause less trouble, but note that we have chosen the same ε for the two original tailbounds. One can easily generalise the above proof to the case, where one chooses a different ε for each tailbound. How to choose the ε is not clear, like the question if there lies a worthwhile optimization potential in doing so. Another problem, which rises in the tailbound approach is that the error of the convoluted service contains the sum $\sum_{l=0}^{n} \zeta(l, \varepsilon)$, so one needs the errors ζ to be summable. In fact one needs to be quite careful, when it comes to the error of concatenated servers, when using tailbounds. How to solve this problem, however, lies outside the scope of this course.

Exercises

Exercise 2.49. Show that for $S(m, n) = S(n - m)$ we have:

$$A \otimes S(0, n) = A \otimes S(n)$$

where the right handed side is the usual univariate convolution.

Exercise 2.50. Assume $S_i(m, n) = S_i(n) - S_i(m)$ for $i \in \{1, 2\}$. Give an example for S_i such that

$$S(m, n) := S_1 \otimes S_2(m, n) \neq S(0, n) - S(0, m)$$

for some $(m, n) \in \Lambda(\mathbb{N}_0)$.

Exercise 2.51. In this exercise we investigate how *strict* a server can be, after subtracting a cross-flow from it. The answer is simple[10]: After subtracting no strict service can be offered, regardless of how strict our original service had been. To show this two statements must be proven. Throughout this we assume a node serving two arrivals, of which A_1 is prioritized over A_2 (see also example 2.43).

- Assume first the service element is a dynamic S-Server with respect to $A := A_1 + A_2$. Show that it is also a dynamic S_l-Server with respect to A_2 with $S_l(m, n) = S(m, n) - A_1(m, n)$.
- Assume now the service element fulfills Lindley's equation with respect to A (i.e. it is a "strictest" server). Construct input flows A_1, A_2 and a service S such that for some interval $[m, n]$ with $D_2(m - 1) = A_2(m - 1)$ holds $S_l(m, n) > D_2(m, n)$, i.e. S_l is not a strict service curve in sense of definition (2.33).[11] (Hint: You can choose a very simple S)

[10] and might surprise someone who is familiar with deterministic network calculus. In the univariate determinstic calculus it is proven, that you always need a strict service curve, when subtracting some crossflow, otherwise there exists none non-trivial service curve for the leftover service. This is an important difference between stochastic network calculus - as presented here - and deterministic network calculus!

[11] In fact any backlogged interval would work here. But in taking the additional assumption $D_2(m-1) = A_2(m-1)$ into account, we can also exclude that the leftover service could be *weak strict*. A dynamic S-Server is defined to be *weak strict* if for any backlogged period $[m, n]$ with $D(m - 1) = A(m - 1)$ holds: $D(m, n) \geq S(m, n)$.

Exercise 2.52. Consider two strict rate-latency server with rate and latency equal to 1

$$S_1(m,n) = S_2(m,n) = \beta(n-m) = [n-m-1]^+$$

Show that the concatenated server $S = S_1 \otimes S_2$ does not offer a strict service curve. In expression construct an input flow A, such that a backlogged period $[m,n]$ emerges from it and $A \otimes S(m,n) < S(m,n)$.

Exercise 2.53. Proof theorem 2.47.

Exercise 2.54. Assume a dynamic S_1-server and a dynamic S_2-server, which is stochastically *dependent* of the first one. Assume further that the S_1-server is $(\sigma_1(p\theta), \rho_1(p\theta))$-bounded and the S_2-server is $(\sigma_2(q\theta), \rho_2(q\theta))$-bounded for some $\theta > 0$, p and q with $\frac{1}{p} + \frac{1}{q} = 1$ and $\rho_1(p\theta) \neq \rho_2(q\theta)$. Show that the concatenation of S_1 and S_2 is $(\sigma_1(p\theta) + \sigma_2(q\theta) + B, \min\{|\rho_1(p\theta)|, |\rho_2(q\theta)|\})$-bounded, with $B = -\frac{1}{\theta}\log(1 - e^{-\theta|\rho_1(p\theta)-\rho_2(q\theta)|})$

Exercise 2.55. Assume the situation as in theorem 2.48. Show that for $\rho_S(\theta) = \rho_T(\theta)$ the dynamic $S \otimes T$-server is $(\sigma(\theta), \rho(\theta))$-bounded with:

$$\rho(\theta) := \rho_S(\theta) + \frac{1}{\theta}$$

and

$$\sigma(\theta) := \sigma_S(\theta) + \sigma_T(\theta)$$

(Hint: Use the simple inequality $n + 1 \leq e^n$ for all $n \in \mathbb{N}_0$)

Exercise 2.56. Let a dynamic S-server be $(\sigma(\theta), \rho(\theta))$-bounded. Consider the following expression:

$$s_S^*(\theta) = \limsup_{m\to\infty} -\frac{1}{\theta m}\log \mathbb{E}(e^{-\theta S(n,n+m)})$$

Show that $s^*(\theta) \leq -\rho(\theta)$. Assume now the situation as in theorem 2.48. Show that for the concatenated server holds:

$$s_{S\otimes T}^*(\theta) \leq -\rho_T(\theta)$$

for all $\rho_T(\theta)$ with $|\rho_T(\theta)| < |\rho_S(\theta)|$. (Hint: Use the geomtric sum, instead of the geometric series: $\sum_{k'=0}^{m} p = \frac{1-p^{m+1}}{1-p}$ for $0 < p < 1$). Next show that also for the case $\rho_T(\theta) = \rho_S(\theta)$ holds:

$$s_{S\otimes T}^*(\theta) \leq -\rho_T(\theta)$$

Exercise 2.57. In this exercise we generalise example 2.46. Consider a dynamic S-server, which is bounded by $(\sigma_S(\theta), \rho_S(\theta))$. Further assume there is a flow A which is $(\sigma_A(\theta), \rho_A(\theta))$-bounded. Give $(\sigma(\theta), \rho(\theta))$-bounds for the leftover Service and the case that

- A is stochastically independent of S.

- A is stochastically dependent of S. What assumptions are further needed for this case?

Exercise 2.58. We consider in this exercise two dynamic server S and T with constant rates s and t, respectively. Each of them gives high priority to a crossflow A_S and A_T, respectively. A low priority flow of interesst enters the dynamic S-server and afterwards traverses the dynamic T-server. Show that the overall system can be described by a dynamic S_{sys}-server, with

$$S_{sys}(m,n) = (s \wedge t)(n-m) - A_S(m,n) - A_T(m,n) + A_S \otimes A_T(m,n)$$

for all $(m,n) \in \Lambda(\mathbb{N}_0)$.

Exercise 2.59. In this programming exercise we want to visualize how differently *service guarantees* (weak strict and normal service elements) can behave, compared to the *real* output of a node. For this we assume again the scenario of a node serving two flows with a fixed rate $c = 1$. The two flows behave very similar, assume them both to be exp(2.5)-distributed up to time $n = 49$. At time $n = 50$, however, the prioritized flow breaks this pattern and produces a large burst of value b. Compute the following values, to compare the different service guarantees for the output D_2 of the second flow and this "bad" behaving prioritized flow:

- (real output) Simulate the output for the second flow and calculate the cumulative output $D_2^{(1)}(n)$ of it.
- (weak strict service element) $D_2^{(2)}(n) = S_l(m,n) + A_2(m)$ and $S_l(m,n) := [c(n-m) - A_1(n) + A_1(m)]^+$, with m being the last time A_2 was not backlogged (i.e. $m = \max_{k \geq 0}\{k : A_2(k) = D_2^{(2)}(k)\}$, to compute m use $D_2^{(2)}(k) = D_2^{(1)}(k)$ for all k).
- (normal service element) $D_2^{(3)}(n) = A_2 \otimes S_l(0,n)$

Compare the three guarantees for different values of $b \in \{1, 5, 10, 50\}$. Analyse the above formulas to discover, why the service guarantees differ from each other and the real output.

3 Stochastic Performance Bounds

Now, as we know how to model stochastic arrivals and service guarantees, we ask in this chapter for performance guarantees. We are interested in bounds on the backlog and the delay of a node, as well as in bounds on the departures of a node. Again the theory here parallels the deterministic approach, with the difference, that the bounds only hold with a high probability. Or stated more positively: The achieved bounds are only violated with very small probabilities.

If not stated otherwise we assume in this chapter a flow A, which is bounded by $(\sigma_A(\theta), \rho_A(\theta))$ and a stochastically independent node S, which is $(\sigma_S(\theta), \rho_S(\theta))$-bounded for the same $\theta > 0$. Further for the tail-bounded case we assume A and S to be tail-bounded by envelopes α and β with errors η and ζ, respectively. The flow A is the input for node S and the corresponding output flow is denoted by D.

3.1 Backlog Bound

Before we can formulate our backlog bound we need another theorem concerning the convolution of MGFs. As a reminder we give the following

Definition 3.1. Let $f, g : \Lambda(\mathbb{N}_0) \to \mathbb{R}$ be two triangular arrays. The *bivariate convolution* $* : \mathcal{D} \times \mathcal{D} \to \mathcal{D}$ of f and g at $(m,n) \in \Lambda(\mathbb{N}_0)$ is defined by:

$$f * g(m,n) = \sum_{k=m}^{n} f(m,k) g(k,n)$$

The *bivariate deconvolution* $\circ : \mathcal{D} \times \mathcal{D} \to \mathcal{D}$ of f and g at $(m,n) \in \Lambda(\mathbb{N}_0)$ is defined by:

$$f \circ g(m,n) = \sum_{k=0}^{m} f(k,n) g(k,m)$$

Theorem 3.2. *Let $X, Y \in \mathcal{S}$ be two stochastically independent random processes. Then it holds that:*

$$\phi_{X \otimes Y(m,n)}(-\theta) \leq \left(\phi_X(-\theta) * \phi_Y(-\theta)\right)(m,n)$$

for all $\theta > 0$ and $(m,n) \in \Lambda(\mathbb{N}_0)$ such that the above MGFs exist.

Further it holds that:

$$\phi_{X \oslash Y(m,n)}(\theta) \leq \left(\phi_X(\theta) \circ \phi_Y(-\theta)\right)(m,n)$$

for all $\theta > 0$ and $(m,n) \in \Lambda(\mathbb{N}_0)$ such that the above MGFs exist.

Proof. We show the first inequality, since the second can be proven in the same fashion:

$$\begin{aligned}\phi_{X\otimes Y}(-\theta) &= \mathbb{E}(e^{-\theta \min_{m\le k\le n}\{X(m,k)+Y(k,n)\}}) = \mathbb{E}(e^{\max_{m\le k\le n}\{-\theta X(m,k)-\theta Y(k,n)\}}) \\ &= \mathbb{E}(\max_{m\le k\le n}\{e^{-\theta X(m,k)}\cdot e^{-\theta Y(k,n)}\}) \le \sum_{k=m}^{n}\mathbb{E}(e^{-\theta X(m,k)}\cdot e^{-\theta Y(k,n)}\}) \\ &= \sum_{k=m}^{n}\phi_{X(m,k)}(-\theta)\phi_{Y(k,n)}(-\theta) = \left(\phi_X(-\theta) * \phi_Y(-\theta)\right)(m,n)\end{aligned}$$

For all $\theta > 0$ such that the above MGFs exist. □

We present the performance bound for the MGF-case first and give a short discussion on it. Thereafter we present the corresponding bound for the tail-bound-case.

Theorem 3.3. *For the backlog $q(n) := A(n) - D(n)$ at time n and all $x \in \mathbb{R}_0^+$ holds:*

$$\mathbb{P}(q(n) > x) \le e^{-\theta x+\theta(\sigma_A(\theta)+\sigma_S(\theta))}\sum_{k=0}^{n} e^{\theta k(\rho_A(\theta)+\rho_S(\theta))}$$

Proof. We know by the previous chapter:

$$q(n) \le A(n) - A\otimes S(0,n) = \max_{0\le k\le n}\{A(k,n) - S(k,n)\} = A \oslash S(n,n)$$

Hence from $q(n) > x$ follows $A \oslash S(n,n) > x$ and using Chernoff's inequality yields:

$$\begin{aligned}\mathbb{P}(q(n) > x) &\le \mathbb{P}(A \oslash S(n,n) > x) \le e^{-\theta x}\mathbb{E}(e^{\theta A \oslash S(n,n)}) \\ &\le e^{-\theta x}\mathbb{E}(e^{\theta A}) \circ \mathbb{E}(e^{-\theta S})(n,n) \\ &= e^{-\theta x}\sum_{k=0}^{n}\mathbb{E}(e^{\theta A(k,n)})\mathbb{E}(e^{-\theta S(k,n)}) \\ &\le e^{-\theta x}\sum_{k=0}^{n} e^{\theta(\sigma_A(\theta)+(n-k)\rho_A(\theta))}e^{\theta(\sigma_S(\theta)+(n-k)\rho_S(\theta))} \\ &\le e^{-\theta x+\theta(\sigma_A(\theta)+\sigma_S(\theta))}\sum_{k'=0}^{n} e^{\theta k'(\rho_A(\theta)+\rho_S(\theta))}\end{aligned}$$

□

There is another way to derive the above bound. Instead of using first Chernoff's inequality and the inqeuality $a + b \ge a \vee b$ thereafter, we can interchange that order:

$$\begin{aligned}\mathbb{P}(\max_{0\le k\le n}\{A(k,n) - S(k,n)\} > x) &= \mathbb{P}\left(\bigcup_{k=0}^{n} A(k,n) - S(k,n) > x\right) \\ &\le \sum_{k=0}^{n}\mathbb{P}(A(k,n) - S(k,n) > x) \\ &\le \sum_{k=0}^{n} e^{-\theta x}\mathbb{E}(e^{\theta(A(k,n)-S(k,n))})\end{aligned}$$

and then proceed as above. The result is the same, even if A and S are not stochastically independent (see the exercises). However one should keep these two different ways in mind, since both of them may contain possible improvements on the bound.

Before we continue, we analyse the structure of the above proof. It consists of the following steps:

1. We expressed the quantity $q(n)$ in terms of A and S.
2. We translate the achieved inequality into an inequality of probabilities.
3. We use Chernoff to bound the new probability and moment generating functions arise.
4. We use the MGF-bounds of A and S.

We will see this structure again in the next two sections.

As a last note we discuss θ. A little assumption was made, by assuming A and S can be bounded with the same parameter θ. This assumption is not very restrictive, since we normally can bound A by $(\sigma_A(\theta'), \rho_A(\theta'))$ for all $0 < \theta' < \theta$ if it is already bounded by $(\sigma_A(\theta), \rho_A(\theta))$. The same holds for S. Hence if the θ in the two bounds differ, we can decrease one of them to the minimal θ.

Nevertheless the parameter θ can and should be optimized in the above bound:

Corollary 3.4. *Assume A is $(\sigma_A(\theta), \rho_A(\theta))$-bounded for all $\theta \in [0, b]$ and S is $(\sigma_S(\theta), \rho_S(\theta))$-bounded for all $\theta \in [0, b]$ and some $b \in \mathbb{R}^+ \cup \{\infty\}$. Then[1]:*

$$\mathbb{P}(q(n) > x) \leq \inf_{\theta \in [0,b]} e^{-\theta x + \theta(\sigma_A(\theta) + \sigma_S(\theta))} \sum_{k'=0}^{n} e^{\theta k'(\rho_A(\theta) + \rho_S(\theta))}$$

We proceed with the tail-bounded version of a backlog bound.

Theorem 3.5. *For all $n \in \mathbb{N}_0$ and $\varepsilon > 0$ holds:*

$$\mathbb{P}(q(n) > \alpha(\varepsilon) \oslash \beta(\varepsilon)(0)) \leq \zeta(n, \varepsilon) + \sum_{m=0}^{n} \eta(m, \varepsilon)$$

Proof. Let $n \in \mathbb{N}_0$ and $\varepsilon > 0$ be arbitrary. Assume for a while that

$$A(m, n) \leq \alpha(n - m, \varepsilon) \qquad \forall\, m \leq n \tag{3.1}$$

and

$$D(0, n) \geq A \otimes \beta(\varepsilon))(0, n) \tag{3.2}$$

would hold. We would have then:

$$\begin{aligned} q(n) = A(0, n) - D(0, n) &\leq \max_{0 \leq k \leq n} \{A(n) - A(k) - \beta(n - k, \varepsilon)\} \\ &\leq \max_{0 \leq k \leq n} \{\alpha(n - k, \varepsilon) - \beta(n - k, \varepsilon)\} \\ &= \alpha(\varepsilon) \oslash \beta(\varepsilon) S(n, n) \end{aligned}$$

[1] We investigate this parameter θ, which we have already called *acuity*, in the exercises.

Since we have followed from (3.1) and (3.2) the above inequality, we have:

$$\begin{aligned}
&\mathbb{P}(q(n) > \alpha(\varepsilon) \oslash \beta(\varepsilon)(n,n)) \\
\leq &\mathbb{P}\left(D(0,n) < A \otimes \beta(\varepsilon)(0,n) \cup \bigcup_{m=0}^{n} A(m,n) > \alpha(n-m,\varepsilon)\right) \\
\leq &\mathbb{P}(D(0,n) < A \otimes \beta(\varepsilon))(0,n)) + \sum_{m=0}^{n} \mathbb{P}(A(m,n) > \alpha(n-m,\varepsilon)) \\
\leq &\zeta(n,\varepsilon) + \sum_{m=0}^{n} \eta(m,\varepsilon)
\end{aligned}$$

□

Let us have a closer look at the above bound: the crucial point here is, that - compared to the MGF-backlog-bound - we do not have a bound for the event $\{q(n) > x\}$with some x choosen by ourselves. Instead $q(n)$ is compared with the rather involved expression $\max_{0\leq k\leq n}\{\alpha(n-k,\varepsilon) - \beta(n-k,\varepsilon)\}$. This, in fact, presents us some challenges. Since we need to solve:

$$\max_{0\leq k\leq n}\{\alpha(n-k,\varepsilon) - \beta(n-k,\varepsilon)\} = x$$

In general this equation might or might not have an unique solution for ε. However, this alters drastically, if we state the more general

Corollary 3.6. *For all $n \in \mathbb{N}_0$ and $\varepsilon, \varepsilon' > 0$ holds:*

$$\mathbb{P}(q(n) > \alpha(\varepsilon) \oslash \beta(\varepsilon')S(n,n)) \leq \zeta(n,\varepsilon') + \sum_{m=0}^{n} \eta(m,\varepsilon)$$

Now the solution to

$$\max_{0\leq k\leq n}\{\alpha(n-k,\varepsilon) - \beta(n-k,\varepsilon)\} = x$$

does not neet to be unique and we are confronted with an optimization problem: How to choose $\varepsilon, \varepsilon'$ such that above equality holds and $\zeta(n,\varepsilon') + \sum_{k=0}^{n} \eta(k,\varepsilon)$ becomes minimal?

Another important note on the tail-bound is, that it works without the assumption of stochastic independence between A and S. This is an important advantage of the tail-bounds.

Exercises

Exercise 3.7. Prove the second part of 3.2!

Exercise 3.8. Use Hölder's inequality to derive a backlog bound for the case that the flow A and the service S are not stochastically independent.

Exercise 3.9. In this exercise we derive a backlog bound for an arrival flow, which is heavy-tailed, i.e. $\phi_{A(m,n)}(\theta)$ does not exist for any choice of $m < n \in \mathbb{N}_0$ and $\theta \in \mathbb{R}^+$. To achieve this we need a deterministic lower bound for the dynamic server S:

$$S(m,n) \geq f(m,n)$$

Use the alternative proof of 3.3 and Markov's inequality instead of Chernoff's inequality, to show:

$$\mathbb{P}(q(n) > x) \leq \sum_{k=0}^{n} \left(\frac{1}{x + f(k,n)} \right)^a \mathbb{E}(A(k,n)^a) \qquad \forall\, a \in \mathbb{N}$$

Use this bound to compute the probability that a node with service $S(m,n) = (n-m)c$ serving an arrival with Pareto distributed increments ($x_{\min} = 0.5$, $\alpha = 3$) has a backlog higher 1 at time n. To achieve the best possible bound optimize the parameter $a \in \{1,2,3\}$.

Exercise 3.10. We want to investigate how a system with increasing number of flows scales. Assume the following scenario: We have one node with a constant service rate $c = 1$ (i.e. $S(m,n) = c(n-m)$ for all $m \leq n \in \mathbb{N}_0$) and N stochastically independent flows arrive at this node. All of the flows share the same distribution for their increments a_i ($i = 1, \ldots, N$). All increments are i.i.d., we consider three different cases of what this distribution looks like:

- $a_i(m) = 1$ with probability $\frac{1}{N}$ and $a_i(m) = 0$ with probability $1 - \frac{1}{N}$
- $a_i(m) = \frac{2}{N}$ with probability $\frac{1}{2}$ and $a_i(m) = 0$ with probability $a_i(m) = \frac{1}{2}$
- $a_i(m) = N$ with probability $\frac{1}{N^2}$ and $a_i(m) = 0$ with probability $1 - \frac{1}{N^2}$

Show that the expected number of arrivals in one time step is in each case equal to 1 (Hence the node has an utilization of 100%).

Give for each case the corresponding $f(\theta,n)$-bound (compare example 2.18) and calculate the backlog bound at time $n = 1$ for a fixed N. How do these bounds evolve for large N? What is

$$\lim_{N \to \infty} \mathbb{P}(q(1) > 2)$$

for each case? Why do they behave differently? Analyse the variance of the increments!

(Hints: For the first case you need that $(1 + \frac{t}{N})^N \xrightarrow{N \to \infty} e^t$, for the second case that $(\frac{1}{2}(e^{\frac{2\theta}{N}} + 1))^N \xrightarrow{N \to \infty} e^\theta$. The variance of a binomial distribution with parameters N and p is given by $\mathrm{Var}(B) = Np - Np^2$)

Exercise 3.11. Construct an arrival A such that the union bound applied in the alternative proof of Theorem 3.3 becomes strict at some time $n \in \mathbb{N}_0$, i.e. holds with equality.

Exercise 3.12. Remember exercise 2.32 in which we have investigated the parameter θ and called it *acuity*. We now check what happens for the backlog bound, when the acuity is altered. For this consider a node, which offers a constant rate service c and an arrival with i.i.d. exponentially distributed increments with parameter $\lambda = 10$. Compute the backlog bound for x at time $n = 1$ and give it in the form

$$\mathbb{P}(q(1) > x) \leq h(x, \theta)(1 + f(\lambda, \theta)g(c, \theta))$$

Here the three functions f, g and h are only dependent on θ and one further parameter. We can identify f and g as the parts of the bound, which result from our arrival bound and the service bound, respectively. Plot for varying $\theta \in [0, 10)$ the functions f, g and the above backlog bound for $x = 1$.

3.2 Delay Bound

The delay bound is achieved similarly to the backlog bound and we follow the same scheme as presented in the previous section to achieve it. First a definition of delay is needed.

Definition 3.13. The *virtual delay* d at time n is defined as:

$$d(n) := \min\{m : A(n) \leq D(n + m)\}$$

The virtual delay can be interpreted as the time D needs to catch up the amount of arrivals.

Theorem 3.14. *For all $N \in \mathbb{N}_0$ holds:*

$$\mathbb{P}(d(n) > N) \leq e^{\theta\rho_S(\theta)N + \theta(\sigma_A(\theta) + \sigma_S(\theta))} \sum_{k=0}^{n+N} e^{\theta(n-k)(\rho_A(\theta) + \rho_S(\theta))}$$

Proof. Assume it holds $d(n) > N$, i.e. $A(n) - D(n + N) > 0$. We have then:

$$\begin{aligned} 0 < A(n) - D(n + N) &\leq A(n) - A \otimes S(0, n + N) \\ &= \max_{0 \leq k \leq n+N} \{A(n) - A(k) - S(k, n + N)\} \end{aligned}$$

Hence the implication

$$d(n) > N \quad \Rightarrow \quad \max_{0 \leq k \leq n+N} \{A(k, n) - S(k, n + N\} > 0$$

holds and therefore:

$$\begin{aligned}
\mathbb{P}(d(n) > N) &\leq \mathbb{P}(\max_{0\leq k\leq n+N}\{A(k,n) - S(k,n+N)\} > 0\}) \\
&\leq \mathbb{E}(e^{\theta \max_{0\leq k\leq n+N}\{A(k,n)-S(k,n+N)\}}) \\
&\leq \mathbb{E}(\max_{0\leq k\leq n+N} e^{\theta(A(k,n)-S(k,n+N))}) \leq \sum_{k=0}^{n+N} \mathbb{E}(e^{\theta(A(k,n)-S(k,n+N))}) \\
&= \sum_{k=0}^{n+N} \mathbb{E}(e^{\theta A(k,n)})\mathbb{E}(e^{-\theta S(k,n+N)}) \\
&\leq \sum_{k=0}^{n+N} e^{\theta\rho_A(\theta)(n-k)+\theta\sigma_A(\theta)} e^{\theta\rho_S(\theta)(N+n-k)+\theta\sigma_S(\theta)} \\
&= e^{\theta\rho_S(\theta)N+\theta(\sigma_A(\theta)+\sigma_S(\theta))} \sum_{k=0}^{n+N} e^{\theta(n-k)(\rho_A(\theta)+\rho_S(\theta))}
\end{aligned}$$

□

Parallel to the backlog bound, the above bound can be achieved on a slightly different way. From $\mathbb{P}(\max_{0\leq k\leq n}\{A(k,n) - S(k,n+N) > 0\})$ one might continue with:

$$\begin{aligned}
\mathbb{P}(d(n) > N) &\leq \mathbb{P}\left(\bigcup_{k=0}^{n} A(k,n) - S(k,n+N) > 0\right) \\
&\leq \sum_{k=0}^{n} \mathbb{P}(A(k,n) - S(k,n+N) > 0) \\
&\leq \sum_{k=0}^{n} \mathbb{E}(e^{\theta(A(k,n)-S(k,n+N))})
\end{aligned}$$

and then proceed as before. Also, we can state the same corollary concerning the parameter θ.

Corollary 3.15. *Assume A is $(\sigma_A(\theta), \rho_A(\theta))$-bounded for all $\theta \in [0, b]$ and S is $(\sigma_S(\theta), \rho_S(\theta))$-bounded for all $\theta \in [0, b]$ and some $b \in \mathbb{R}^+ \cup \{\infty\}$. Then:*

$$\mathbb{P}(d(n) > N) \leq \inf_{\theta\in[0,b]} e^{\theta\rho_S(\theta)N+\theta(\sigma_A(\theta)+\sigma_S(\theta))} \sum_{k=0}^{n+N} e^{\theta(n-k)(\rho_A(\theta)+\rho_S(\theta))}$$

Making a further (but often reasonable) assumption on S the above bound can be improved.

Corollary 3.16. *Assume the situation as in the previous corollary Further let $S(k, n + N) \geq 0$ for all $n + 1 \leq k \leq N + n$. Then:*

$$\mathbb{P}(d(n) > N) \leq \inf_{\theta\in[0,b]} e^{\theta\rho_S(\theta)N+\theta(\sigma_A(\theta)+\sigma_S(\theta))} \sum_{k=0}^{n} e^{\theta k(\rho_A(\theta)+\rho_S(\theta))}$$

Proof. Using $A(k,n) - S(k,n+N) \leq 0$ for all $n+1 \leq k \leq N+n$ we have:

$$\begin{aligned} 0 < A(n) - D(n+N) &\leq A(n) - A \otimes S(0, n+N) \\ &= \max_{0 \leq k \leq n} \{A(n) - A(k) - S(k, n+N)\} \end{aligned}$$

The remainder of the proof looks like above. □

For the tail-bounded version we need a stability condition on the service.

Definition 3.17. Assume S is tail-bounded with envelope β. We say S is *delay-stable* for ε and some envelope α, if there exists for each $k \in \mathbb{N}_0$ a constant N_k, such that:

$$\beta(k+l) - \alpha(k, \varepsilon) \geq 0 \qquad \forall\, l > N_k$$

holds.

The interpretation of delay-stability is that β will exceed the value of $\alpha(k,\varepsilon)$ eventually - no matter which k we choose.

Theorem 3.18. *Assume S is delay stable for some ε and envelope α. Then we have for all $n \in \mathbb{N}_0$:*

$$\begin{aligned} &\mathbb{P}\big(d(n) > \min\{N \geq 0 \,:\, \beta(\varepsilon) \oslash \alpha(\varepsilon)(n, n+N') < 0 \,\forall\, N' \geq N\}\big) \\ \leq &\sum_{k=0}^{n} \eta(k,\varepsilon) + \sum_{k=n+1}^{\infty} \zeta(k,\varepsilon) \end{aligned}$$

Proof. Fix an arbitrary n and assume for a while

$$D(n+l) \geq A \otimes \beta(\varepsilon)(0, n+l) \tag{3.1}$$

for all $l \geq 0$ and

$$A(k,n) \leq \alpha(n-k, \varepsilon) \tag{3.2}$$

for all $k \leq n$. Choose now some arbitrary $l < d(n)$, then we have by the definition of virtual delay and (3.1):

$$A(n) > D(n+l) \geq \min_{0 \leq k \leq n+l} \{A(k) + \beta(n+l-k, \varepsilon)\}$$

From this and (3.2) we can follow

$$\begin{aligned} 0 &< \max_{0 \leq k \leq n+l} \{A(n) - A(k) - \beta(n+l-k, \varepsilon)\} \\ &\leq \max_{0 \leq k \leq n} \{A(n) - A(k) - \beta(n+l-k, \varepsilon)\} \vee 0 \\ &\leq \max_{0 \leq k \leq n} \{\alpha(n-k, \varepsilon) - \beta(n+l-k, \varepsilon)\} \vee 0 \end{aligned} \tag{3.3}$$

where we have used that β is non-negative in the transition to the second line. Writing it shorter, we eventually have derived $0 < \beta(\varepsilon) \oslash \alpha(\varepsilon)(n, n+l)$ for all $l < d(n)$. By the assumption of delay-stability, we know there exists for each $k \in \{1, \dots, n\}$ an N_k, such that

$$\alpha(n-k, \varepsilon) - \beta(n+l-k, \varepsilon) \leq 0$$

for all $l > N_k$. Define $N := \max_{0 \leq k \leq n} N_k$, then we have

$$\beta(\varepsilon) \oslash \alpha(\varepsilon)(n, n+l) \leq 0$$

for all $l > N$. Note that the above equation can not hold, while $l < d(n)$. Hence we implied the following: if $l > N$, we also have $l \geq d(n)$. Collecting all of this, we started with equations (3.1) and (3.2) and derived:

$$d(n) \leq N := \min\{N \geq 0 \,:\, \beta(\varepsilon) \oslash \alpha(\varepsilon)(n, n+l) \leq 0 \,\forall\, l \geq N\}$$

Moving to probabilities yields:

$$\begin{aligned}
&\mathbb{P}(d(n) > N) \\
\leq &\mathbb{P}\left(\bigcup_{k=0}^{n} A(k,n) > \alpha(n-k, \varepsilon) \cup \bigcup_{l \geq 0} D(n+l) < A \oslash \beta(\varepsilon)(0, n+l)\right) \\
\leq &\sum_{k=0}^{n} \eta(k, \varepsilon) + \sum_{k=n}^{\infty} \zeta(k, \varepsilon)
\end{aligned}$$

□

Again, we can make similar comments, to the ones made for the backlog bound: Solving the equation

$$x = \min\{N \geq 0 : \beta(\varepsilon) \oslash \alpha(\varepsilon)(n, n+l) \leq 0 \,\forall\, l \geq N\}$$

is quite involved and the existence of a solution is not guaranteed. Further for the generalized case uniqueness of a solution (if existence) is in general not given. We close with the tail-bound version of a delay bound, which allows optimization over ε and ε':

Corollary 3.19. *Assume S is delay-stable for all $\varepsilon \in I$ (I being some interval). Then we have for all $n \in \mathbb{N}_0$ and $\varepsilon' > 0$ and $\varepsilon \in I$:*

$$\begin{aligned}
&\mathbb{P}(d(n) > \min\{N \geq 0 : \beta(\varepsilon') \oslash \alpha_A(\varepsilon)(n, n+l) \leq 0 \,\forall\, l \geq N\}) \\
\leq &\sum_{k=0}^{n} \eta(k, \varepsilon) + \sum_{k=n}^{\infty} \zeta(k, \varepsilon')
\end{aligned}$$

Exercises

Exercise 3.20. One can also define a backwards oriented delay by:

$$\tilde{d}(n) := \min\{m \,:\, D(n) - A(n-m) > 0\}$$

Show that

$$\mathbb{P}(\tilde{d}(n) > N) \leq e^{\theta\rho_S(\theta)N + \theta(\sigma_A(\theta)+\sigma_S(\theta))} \sum_{k=0}^{n-N} e^{\theta k(\rho_A(\theta)+\rho_S(\theta))}$$

for all $n > N \in \mathbb{N}_0$.

Exercise 3.21. If we compare the backlog bound with the delay bound we notice that they only differ in the leading terms $e^{-\theta x}$ and $e^{\theta\rho_S(\theta)N}$. In fact one can interpret the delay bound as a backlog bound, if there is some guarantee on the service.

Assume that $D(n) - D(m) \geq f(m,n)$ if $[m,n]$ is a backlog period. (i.e. the node offers a deterministic strict service curve). Show that the delay bound, can be expressed as a backlog bound:

$$\mathbb{P}(d(n) > N) \leq \mathbb{P}(q(n) > f(n, n+N))$$

(Hint: Construct first the dynamic S-Server. Then show that from $d(n) > N$ follows $q(n) > f(n, n+N)$)

Exercise 3.22. Assume now that the arrivals have a strict bound, i.e. $A(m,n) \leq f(m,n)$, where f is monotone decreasing in the first variable. Show that the backlog bound can be expressed by the backward delay bound from exercise 3.20:

$$\mathbb{P}(q(n) > x) \leq \mathbb{P}(\tilde{d}(n) > m)$$

with m maximal such that

$$f(m,n) < x$$

is still fulfilled.

Exercise 3.23. Let us investigate the bound from 3.14 in more detail, in expression the term $\sum_{k'=0}^{n} e^{\theta k'(\rho_A(\theta)+\rho_S(\theta))}$. You might remember the geometric series:

$$\sum_{k=0}^{\infty} q^k$$

which converges to the value $\frac{1}{1-q}$ if $|q| < 1$ holds. How behaves the delay bound for $n \to \infty$, in the case $\rho_A(\theta) \geq -\rho_S(\theta)$?

The above shows, that one can think of $\rho_A(\theta) < -\rho_S(\theta)$ as a stability condition. The following counterexample will however show, that it is not a stability condition in the intuitive sense, that the utilization of a node is below 100%. Assume for this a constant rate server with rate c and an arrival with exponentially distributed increments (as in example 2.17). The average rate of arrivals per time step is equal to: $\mathbb{E}(a(n)) = \frac{1}{\lambda}$, where

λ is the parameter of the exponential distribution. This gives for the node an utilization of

$$\frac{\frac{1}{\lambda}}{c} = \frac{1}{\lambda c}$$

Show that there exists a choice of c, λ and θ, sucht that $\frac{1}{\lambda c} < 1$ holds, but also:

$$\rho_A(\theta) \geq -\rho_S(\theta)$$

This means our system is stable by construction (the node's utilization is below 100%), but our delay bound diverges for $n \to \infty$, no matter how large we choose N in 3.14. This makes it blatantly obvious, that the derived delay-bound is really only a *bound*. A bound, which in some situations is far from the real systems behaviour. Hence the right choice of θ is crucial[2].

Exercise 3.24. In this exercise we investigate how different timescales influence the quality of the delay bound. In expression, if one has two descriptions of a system differing only in different definitions of what one timestep is, which of the two descriptions leads to a better delay bound?

To do this assume a flow A which is $(0, \rho_A(\theta))$-bounded and a node, which offers a constant rate c. Introduce a parameter $m \in \mathbb{N}$ called granularity and a corresponding flow A_m such that:

$$a_m(k) := \sum_{l=1}^{m} a(m \cdot (k-1) + l)$$

Give a bound for $\mathbb{E}(e^{\theta A_m(k,n)})$.

Next we define the virtual delay with granularity m. Let T be a multiple of m:

$$d_m(T) := m \cdot \min\{k \,:\, A(T) < D(T + m \cdot k)\}$$

Show that $d(T) = d_m(T)$.

Show that for all N being a multiple of m holds:

$$\mathbb{P}(d_m(T) > N) \leq e^{-\theta c N} \sum_{l=0}^{\frac{T}{m}} (e^{\theta(\rho_A(\theta) - c) \cdot m})^l$$

Show now that for an even N and T the bound for $d_2(T)$ is strictly smaller than the bound for $d(T)$. Using this result, what is the best bound one can achieve for the event, that the delay at (an arbitrary) time $T \in \mathbb{N}$ is smaller than (an arbitrary) $N \in \mathbb{N}$?

[2] To see this even more obvious you can prove the following: For every choice of c and λ with $1/\lambda < c$ exists a θ such that $\rho_A(\theta) \geq -\rho_S(\theta)$. You may proceed like follows: Reformulate $\rho_A(\theta) \geq -\rho_S(\theta)$ as $(1 - \theta/\lambda)e^{\theta c} < 1$. Now interpret the last expression as a function in θ, denoted by f. Calculate the values $f(0)$ and $f(\lambda)$ and apply the intermediate value theorem.

3.3 Output Bound

Again we follow the same strategy as for the backlog bound.

Theorem 3.25. *For $\rho_S(\theta) < -\rho_A(\theta)$ the output is bounded by $(\sigma_A(\theta)+\sigma_S(\theta)+B(\rho_A(\theta),\rho_S(\theta)),\rho_A(\theta))$. With $B(\rho_A(\theta),\rho_S(\theta)) = -\frac{1}{\theta}\log(1-e^{(\theta\rho_A(\theta)+\rho_S(\theta))})$*

Proof. We have:

$$\begin{aligned} D(n) - D(m) &\leq D(n) - \min_{0\leq k\leq m}\{A(k)+S(k,m)\} \\ &\leq A(n) - \min_{0\leq k\leq m}\{A(k)+S(k,m)\} \\ &= \max_{0\leq k\leq m}\{A(n)-A(k)-S(k,m)\} = A \oslash S(m,n) \end{aligned}$$

Hence:

$$\begin{aligned} \mathbb{E}(e^{\theta(D(n)-D(m))}) &\leq \mathbb{E}(e^{\theta(A\oslash S(m,n))}) \leq \mathbb{E}(e^{\theta A}) \circ \mathbb{E}(e^{-\theta S})(m,n) \\ &= \sum_{k=0}^{m} \mathbb{E}(e^{\theta A(k,n)})\mathbb{E}(e^{-\theta S(k,m)}) \\ &\leq \sum_{k=0}^{m} e^{\theta\rho_A(\theta)(n-k)+\theta\sigma_A(\theta)} e^{\theta\rho_S(\theta)(m-k)+\theta\sigma_S(\theta)} \\ &= e^{\theta\rho_A(\theta)(n-m)+\theta\sigma_A(\theta)+\theta\sigma_S(\theta)} \sum_{k=0}^{m} e^{\theta(m-k)(\rho_A(\theta)+\rho_S(\theta))} \\ &\leq e^{\theta\rho_A(\theta)(n-m)+\theta(\sigma_A(\theta)+\sigma_S(\theta)-\frac{1}{\theta}\log(1-\exp(\theta\rho_A(\theta)+\rho_S(\theta)))} \end{aligned}$$

Where the sum was bounded by its corresponding geometric series in the last line ($\sum_{k=0}^{\infty} q^k = \frac{1}{1-q}$ if $0 < |q| < 1$). □

Note that the last step in the above proof is not necessary to get a bound on the output. The only reason for performing the last inequality is to stay inside the class of $(\sigma(\theta),\rho(\theta))$-bounded arrivals. Coming back to this framework allows analyzing whole networks, since the output bound can be used again as input bound in theorems 3.3, 3.14 and 3.25. In this course we sticked to the $(\sigma(\theta),\rho(\theta))$-bounds, since they make formulas more tractable and give us a clearer view on the overall picture.

However all of the presented theorems can be generalized to the usage of $f(\theta,n)$-bounds, avoiding the last step in the previous proof. Leaving it out gets us bounds, which can be significally better (compare exercise 3.23 to see how much one can gain here).

The last thing to do is the tail-bounded version of output bounds. For these we need again a stability condition, similar to the one for delay-bounds:

Definition 3.26. We say S is *output-stable* for envelope α and $\varepsilon > 0$, if for every m the function $\alpha(\varepsilon) \oslash \beta(\varepsilon)(n, n+m)$ is bounded in n; i.e. there exists N_m, such that

$$\alpha(\varepsilon) \oslash \beta(\varepsilon)(n,n+m) \leq \alpha(\varepsilon) \oslash \beta(\varepsilon)(N_m, N_m+m) < \infty \qquad \forall\, n \in \mathbb{N}_0$$

The intuition behind above definition is the following: while the expression on the left-hand side is dependent on both m and n, the expression on the right-hand side is dependent on m only. This allows us to move from a bivariate function to a univariate, by just maximizing over the first variable. Hence S being output-stable just means that the maximization does not cause any trouble, i.e. the maximum is finite and is adopted by some index N_m. As envelopes are always univariate moving to a univariate function is a necessary step, as seen now.

Theorem 3.27. *Assume S is output-stable for some ε and envelope α. Define the following envelope*

$$\alpha_D(m,\varepsilon) := \alpha(\varepsilon) \oslash \beta(\varepsilon)(N_m, N_m + m)$$

where N_m maximizes the set $\alpha(\varepsilon) \oslash \beta(\varepsilon)(n, n+m)$. We have that D is tailbounded by envelope α_D with error

$$\eta_D(m,\varepsilon) := \zeta(0,\varepsilon) + \sum_{k=0}^{\infty} \eta(m+k,\varepsilon)$$

Proof. Let $m, n \in \mathbb{N}_0$ and $\varepsilon > 0$ arbitrary. We consider the expression $D(n+m) - D(n)$. Assume for a while that

$$D(0,n) \geq A \otimes \beta(\varepsilon)(0,n) \tag{3.1}$$

holds, as well as:

$$A(k, n+m) \leq \alpha(n+m-k,\varepsilon) \qquad \forall\, k \leq n \tag{3.2}$$

We have then:

$$\begin{aligned}
D(n+m) - D(n) &\leq D(n+m) - \min_{0\leq k\leq n} \{A(k) + \beta(n-k,\varepsilon)\} \\
&\leq A(n+m) - \min_{0\leq k\leq n} \{A(k) + \beta(n-k,\varepsilon)\} \\
&= \max_{0\leq k\leq n} \{A(n+m) - A(k) - \beta(n-k,\varepsilon)\} \\
&\leq \max_{0\leq k\leq n} \{\alpha(n+m-k,\varepsilon) - \beta(n-k,\varepsilon)\} \\
&= \alpha(\varepsilon) \oslash \beta(\varepsilon)(n, n+m) \\
&\leq \alpha(\varepsilon) \oslash \beta(\varepsilon)(N_m, N_m+m) \qquad (3.3)
\end{aligned}$$

From our initializing assumption we derived henceforth inequality (3.3). Again, moving

to probabilities finishes the proof:

$$\begin{aligned}
&\mathbb{P}(D(n, n+m) > \alpha_D(m, \varepsilon)) \\
\leq &\mathbb{P}\left(D(0,n) < A \otimes \beta(\varepsilon)(0,n) \cup \bigcup_{k=0}^{n} A(k, n+m) > \alpha(n+m-k, \varepsilon)\right) \\
\leq &\zeta(n, \varepsilon) + \sum_{k=0}^{n} \eta(n+m-k, \varepsilon) \\
\leq &\zeta(0, \varepsilon) + \sum_{k'=0}^{n} \eta(m+k', \varepsilon) \\
\leq &\zeta(0, \varepsilon) + \sum_{k'=0}^{\infty} \eta(m+k', \varepsilon)
\end{aligned}$$

□

We have seen two stability conditions needed to derive performance bounds, when we are in the tail-bounded case. To state the theorems as general as possible, we introduced delay-stability, as well as, output-stability. The following definition, brings these two together and allows computation of both bounds at the same time (while being a "too" strict of an assumption, if one is interested in either delay or output):

Definition 3.28. S is *stable* for some $\varepsilon > 0$ and α if it is delay-stable and output-stable.

We conclude this chapter by giving an example for a stable service S.

Example 3.29. Let S be a constant rate server, $S(k, n) = (n-k)c$, serving a crossflow A_1 with envelope $\alpha_1(n, \varepsilon) := r_1 n + \varepsilon$ and error $\eta_1(n, \varepsilon) = Me^{-\theta\varepsilon}$. We recall from example 2.46 that the flow of interest sees a dynamic S-server with envelope $\beta(n, \varepsilon) := [cn - \alpha_1(n, \varepsilon)]^+$. Now let the flow of interest A_2 have envelope $\alpha_2(n, \varepsilon) := r_2 n + \varepsilon$. We have then:

$$\begin{aligned}
&\alpha_2(\varepsilon) \oslash \beta(\varepsilon)(n, n+m) \\
= &\max_{0 \leq k \leq n} \{\alpha_2(n+m-k, \varepsilon) - [c(n-k) - \alpha_1(n-k, \varepsilon)]\} \\
\geq &\max_{0 \leq k \leq n} \{r_2(n+m-k) + \varepsilon - c(n-k) + r_1(n-k) + \varepsilon\} \\
= &\max_{0 \leq k \leq n} \{(n-k)(r_2 + r_1 - c) + r_2 m + 2\varepsilon\}
\end{aligned}$$

and if $c > r_1 + r_2$ we can maximize the expression by setting $k = n$ and the result is only dependent on m:

$$r_2 m + 2\varepsilon < \infty$$

Hence S is output-stable.

For delay-stability we easily check

$$\beta(n, \varepsilon) \geq (c - r_1)n - \varepsilon \geq 0$$

if $c \geq r_1$. Hence for our example β is stable if $c > r_1 + r_2$, this kind of stability condition can be frequently found in the literature (on stochastic network calculus or queueing systems in general).

Bibliography