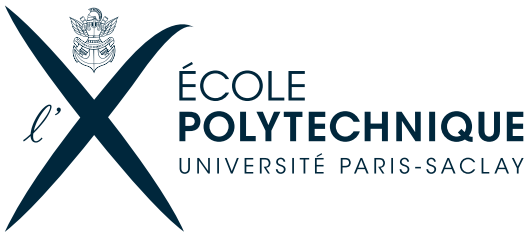
Transient Behavior of Distributed Algorithms and Digital Circuit Models

Thomas Nowak



Transient Behavior of Distributed Algorithms and Digital Circuit Models

Thèse présentée pour obtenir le grade de

DOCTEUR DE L'ÉCOLE POLYTECHNIQUE

par

Thomas Nowak

Soutenue le 5 septembre 2014 devant le jury composé de :

Bernadette Charron-Bost	CNRS, École polytechnique	directrice de thèse
Bernd Heidergott	Vrije Universiteit Amsterdam	rapporteur
Yoram Moses	Technion	rapporteur
François Baccelli	University of Texas at Austin	examinateur
Peter A. Beerel	University of Southern California	examinateur
Stéphane Gaubert	Inria, École polytechnique	examinateur
Ulrich Schmid	Technische Universität Wien	examinateur

Für Sabine
In Liebe und Erinnerung

Contents

Thesis Summary

The overall theme of the thesis is the transient behavior of certain distributed systems. The results can be grouped into three different categories: Transients of max-plus matrices and linear systems, convergence of asymptotic consensus systems, and glitch modeling in digital circuits.

For max-plus algebra, the results are upper bounds on the transient (coupling time) of max-plus matrices and systems. They strictly improve all existing transience bounds. An account of the impact of these bounds in applications is given. The proofs mainly consist of walk reduction and completion procedures. For critical indices, sharper bounds are possible. In fact, they turn out to be independent of the specific weights, and to only depend on the structure of the matrix's digraph and its critical digraph. They are also strict generalizations of the Boolean transience bounds in non-weighted digraphs by the likes of Wielandt or Dulmage and Mendelsohn.

For asymptotic consensus, i.e., a set of agents possessing a real value each and repeatedly updating it by forming weighted averages of its neighbors' values, the thesis strengthens certain upper bounds on the rate of convergence and shows new convergence results for the case of non self-confidence, i.e., agents possibly disregarding their own value. Asymptotic consensus can be described by a non time-homogeneous linear system in classical algebra. The results here are typically in completely dynamic networks. The thesis also presents a worst-case example that shows that exponentially large convergence time is possible even in static networks; meaning that the worst case convergence time in large classes of dynamic networks is actually achieved with a completely static one.

The last part of the thesis is about glitch propagation in digital circuits. More specifically, it is about discrete-value continuous-time models for digital circuits. These models are used in hardware design tool chains because they are much faster than numerically solving the differential equations for timing simulations. However, as is shown in the thesis, none of the existing discrete-value models can correctly predict the occurrence of glitches (short pulses) in the output signal of circuits. Moreover, the thesis proposes a new discrete-value model and proves analytically that it does not share the same characteristics with the existing models that prevented them to correctly predict glitches.

Acknowledgments

First and foremost I would like to express my respect and gratitude to my thesis supervisor, Bernadette Charron-Bost, whose firm scientific standards and constant self-reflection were an inspiration of highest importance to my growth both as an aspiring researcher and as a person. On numerous occasions she showed me the way, not by straight out telling me what to do, but by encouraging a thorough process of self-evaluation. I feel unable to formulate the many ways in which her patience, her encouragement, and the opportunity to witness her deep scientific skills enabled my growth during the years I had the pleasure to work under her supervision.

I would also like to thank Ulrich Schmid of the Vienna University of Technology, who already guided me during my master's thesis and who sparked my interest to pursue a path in scientific research in the following years. I cannot begin to describe the full impact his constant reassurance and hands-on research advice had on my development and the results of this thesis.

My deepest thanks go to Bernd Heidergott, Yoram Moses, François Baccelli, Peter Beerel, and Stéphane Gaubert, who accepted to be part of my thesis committee, and for all of whom I have the greatest of respect.

My work would not have been possible without the many great personalities I had the pleasure to work with, in particular my coauthors Bernadette, Matthias Függer, Alexander Kößler, Glenn Merlet, Robert Najvirt, Ulrich, Hans Schneider, Sergeĭ Sergeev, and Martin Zeiner.

Biggest thanks go out to all the wonderful people I met during my time at the Vienna University of Technology, in particular to Andi, Alex, Hans, Heinrich, Heinz, Jakob, Joschi, Josef, Martin, Motzi, Peter, Robert, Thomas, and Traude. I am grateful to my colleagues at École polytechnique for making my time there a positive and fun experience; in particular I would like to thank my office mates Arnaud, Cécile, Chantal, Claire, Enrico, Jonas, and Victor, as well as Julien Cervelle, Philippe Chassignet, and Thomas Clausen, whom I had the pleasure to teach with. My warmest greetings are addressed to my colleagues at ENS Paris, in particular to Ana, Anne, Bartek, Christelle, Eugène, Florian, Kumar, Marc, Mathieu, Miodrag, Mustafa, Paul, Pierre, Rui, and Seyoung.

I would like to thank my family, who supported me through the years, and through all my life really, in particular Günther, Matthias, Dani, Michi, and Robert. Big shout-outs go to all my friends, the smallest but not least important subset of whom consists of Clemens, Lukas, Markus, and Rafal. My deepest feelings go to Doris, whose love and support I only begin to fathom.

Chapter 1

Introduction

The evolution of a distributed system is determined by the algorithm that the processes run and the environment they are in. The environment essentially governs the order in which the computing steps and communications take place. Oftentimes, this order is not known to the algorithm designer a priori, but certain properties may be known to hold. An example where the order *is* completely known a priori is the synchronous message passing model without failures: All processes take their steps simultaneously at all integral times $0, 1, 2, \ldots$ and all message delays are equal and strictly less than 1, say $1/2$. Here, the algorithm designer can be sure that every message that is sent by a process at time t is received by all other processes at their next computing step at time $t + 1$. If the designer knows that the algorithm will run in such an environment, then their control over the behavior of the final system is high. In a variant of this model, one that allows link failures, the designer does not have this complete a priori knowledge, and hence their control over the system's behavior is reduced. It was shown that a particular variant of the synchronous message passing with link failures, the Heard-Of model [27], captures the computational power of a wide range of message passing models, including completely asynchronous models with crash faults.

In this thesis, we do not study distributed algorithm design. Rather, we fix particularly simple algorithms and study them in different environments. This study is first and foremost one of time complexity and timing constraints. We unite both under the term *transient behavior*. With a *terminating* algorithm, the system comes to a halt after the termination time. Considering its behavior in an unbounded time horizon, the actual execution of the algorithm appears as an initial, transient, time interval. Not all distributed algorithms terminate, nor are they designed to terminate: *Reactive* systems provide an ongoing service, such as clock synchronization or group membership. Their long-term behavior should not be trivial, but a certain regularity is often observed and desirable. Of course, such a regularity depends on the environment satisfying certain regularity properties itself. However, even in a completely regular and predictable environment, the algorithm may need some initial, transient, set-up time before the regular behavior can ensue.

A particular and simple algorithm is the MaxFlood algorithm, in which every process has an initial numeric value and communicates its value to all its neighbors, updating its own value to the maximum of the received values. The algorithm is known to solve the terminating consensus problem [63] in a large variety of environments. The terminating consensus problem requires all processes to agree on a single value, to locally decide on this value, and to halt afterwards. To ensure a non-trivial behavior, the agreed value has to be one of the pro-

cesses' initial values. In a variant of the terminating consensus problem, the processes need to locally decide on values that all are in an ε-neighborhood of each other and that are all in the convex hull of the set of initial values. This is known as the ε-agreement problem. It is solvable in a larger set of environments than terminating consensus. Both problems require a terminating algorithm. A non-terminating variant of ε-agreement is asymptotic consensus. It only requires that the processes' values converge to a common value. Again, this common value is required to be in the convex hull of the set of initial values. Evidently, a time-invariant environment in which ε-agreement is solvable for every $\varepsilon > 0$ allows to solve asymptotic consensus. On the other hand, one can solve ε-agreement by modifying an algorithm for asymptotic consensus if an upper bound on the speed of convergence is available.

There is a very simple algorithm for asymptotic consensus that works in a large class of environments: In every computation step of a process, it updates its value to some average of all values it has received, and then sends out its new value. This simple algorithm has two remarkable properties: Firstly, it is very simple and yet manages to solve asymptotic consensus in a surprisingly large number of different environments. Secondly, it is an algorithm that can be *observed* in nature. More specifically, it serves as a widely accepted model in biology, physics, and sociology to explain various phenomena such as bird flocking, synchronization of coupled oscillators, and opinion spreading. It thus stands to reason to expect the algorithm to have a certain robustness against adverse environments. Of course, one can think of using it to attain approximate agreement in man-made, engineered, systems. And indeed, it is actually used, for example in sensor fusion. For engineered systems, the viewpoint is not one of observing and explaining a given system, but of *analyzing* it for prediction of its future behavior or for assessing the need to improve the system. The speed of convergence in the context of asymptotic consensus is a measure for the stabilization time, or the transient phase, of the system. Obviously, the sharper the analysis of the system and its performance, the tighter it can be integrated into the timing constraints of a larger system, and hence the larger the potential performance of the larger system.

The first part of this thesis is therefore devoted to the analysis of the transient behavior of simple classes of distributed algorithms. By simplicity we mean *linearity*, i.e., systems whose behavior can be described by a linear system. The simple algorithm for asymptotic consensus that we discussed is an instance of a linear distributed algorithm because the update rules for the local values are averages, which are linear functions of the values of the other processes. Another notion of linearity is that of *max-plus* linearity. It means linearity when replacing the addition by the maximum operation and the multiplication by ordinary addition. It encompasses a large number of systems with a time synchronization primitive, as well as the above mentioned MaxFlood algorithm for terminating consensus.

In the second part of the thesis, we look at transient behavior and stabilization times in a domain in which they are of even more direct significance for performance of computing systems, including distributed systems, namely *digital circuits*. The ability to bound the stabilization time of signals in digital circuits is intimately tied to the ability to run the circuit at a higher clock frequency. Failure to predict the stabilization time correctly can lead to glitches which can lead the circuit into a corrupted or even metastable state. Figure 1.1 depicts such a glitch. The occurrence of glitches becomes more common and more critical with higher clock frequencies and lower voltage swings. Glitches are even more critical in clockless and asynchronous circuits. Of course, there are elaborate and very exact physical models for analyzing and simulating digital circuits. They rely on differential equations and can be numerically solved with the help of the widely used Spice simulator. The drawback of these models is that

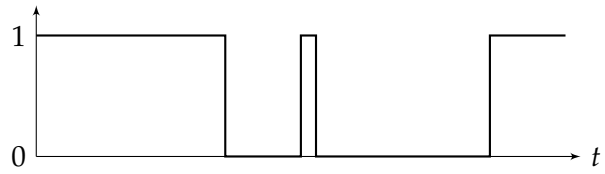


Figure 1.1: A glitch (short pulse) in a digital signal

the time needed to simulate the execution of a circuit can be prohibitively large. Therefore, a number of value discrete event-driven models were developed for the timing analysis of digital circuits. Naturally, they are less accurate than the differential equation models, but their execution time can be much lower. They decompose the circuit into a set of Boolean gates and a set of interconnecting channels. An easy timing model, which is nonetheless widely use, especially in VHDL or Verilog simulators, is to assign every channel a constant delay with which it propagates incoming transitions. More complex channel models include the *inertial delay*, which propagates two (opposing) transitions only if their time distance exceeds a certain threshold. With these channel delays, it is easy to rule out the occurrence of short pulses. However, in physical circuits, short pulses do appear and a faithful model for digital circuits should correctly predict them, and not rule them out by design.

1.1 Structure of the Thesis and Contribution

The first part of the thesis is devoted to linear distributed algorithms. We first treat max-plus linearity and then linearity in the classical sense. Certain graph-theoretic techniques transfer from the max-plus case to the classical case.

In Chapter 2, we give an introduction to max-plus algebra, its applications, and the transient behavior of such systems. We also survey the state of the art for assessing the length of the transient phase. In Chapter 3, we start to go beyond the current state of the art when giving direct generalizations of almost all transience bounds known for Boolean matrices, or equivalently digraphs, to the general max-plus setting. This corresponds to the passage from unweighted to weighted digraphs. Although certain transience bounds in max-plus algebra are already known, none of them had the property that they reduce to classical bounds when restricting them to Boolean matrices. We then go on to use a different approach whose core are walk reductions via cycle removal to develop a general proof strategy for proving max-plus transience bounds in Chapter 4. With this strategy, we are able to improve *all* currently known transience bounds in max-plus algebra.

We present the case of linearity in classical algebra in form of a linear asymptotic consensus algorithm in Chapter 5. We discuss the issues of hypothesis under which the system reaches asymptotic consensus, and the speed of this convergence. It includes a detailed account of the current state of the art. We also present novel results when studying the rate of convergence in systems with a constant system matrix. This corresponds to bounding the spectral gap of the matrix. While certain bounds are known for reversible matrices, we extend the analysis to non-reversible matrices by considering their singular values. Another new contribution in this chapter is the extension of the results from constant matrices to time-dependent matrices whose Perron vector, i.e., their stationary distribution when considered as a Markov chain, remains constant. We also give an example of a constant matrix whose rate

of convergence is exponentially small and hence in the same order of magnitude than that of a large class of time-dependent matrices. This shows that the seemingly easier case of a constant matrix is already the worst case in terms of rate of convergence in this class. We then deal with convergence of asymptotic consensus with more general time-dependent matrices. It starts with a discussion of an appropriate tool for proving this convergence, using a coefficient of ergodicity. Our contribution in this chapter is firstly to explicitly bound the rate of convergence in many cases, and secondly to extend the known results to models where agents do not necessarily have self-confidence. We do this by introducing the notion of an aperiodic core of a time-dependent matrix. One application of this notion are the convergence proofs in certain non-synchronous systems.

The second part of the thesis deals with glitch propagation in digital circuit models. We show that none of the existing binary circuit models faithfully captures glitch propagation, and we propose an alternative model not sharing the same deficiency.

In Chapter 6, we introduce the problem of glitch propagation in digital circuits. After giving the state of the art, we then go on to define what we think is the essence of glitch propagation, namely the Short Pulse Filtration (SPF) problem. After that, we prove the unsolvability of bounded SPF and the solvability of unbounded SPF in physical circuits using a well-accepted differential equation model. In Chapter 7, we present a binary circuit that encompasses all existing binary models to date. They differ in the way channels propagate incoming transitions. In this framework, we also define a new class of models, the involution channel models, and a specific instance based on a first-order model of a physical model, called the exp-channel model. Chapters 8 and 9 prove the insufficiency of the existing models by showing that they either fall into a class where not even unbounded SPF is solvable or into one where even bounded is solvable. We also prove that our newly defined involution channels, and the exp-channels in particular, do not have the same property by showing that unbounded SPF is solvable while bounded SPF is not.

Chapter 10 gives a short summary of the results of the thesis and discusses their consequences. It also includes an outlook on possible future work and perspectives.

Part I

Transience of Distributed Algorithms

Chapter 2

Max-Plus Linear Algorithms

2.1 Introduction

The behavior of certain distributed systems can be described by a sequence of n-dimensional vectors $x(t)$ in $\mathbb{R}^n$ that satisfy a recurrence relation of the form

$$\forall t \geq 1 \;\; \forall i \in [n] \;\; : \;\; x_i(t) = \max_{j \in \mathbf{N}_i} \left(x_j(t-1) + A_{i,j} \right) \tag{2.1}$$

where the $A_{i,j}$ are real numbers, and the $\mathbf{N}_i$ are subsets of $[n] = \{1, 2, \ldots, n\}$. For example, $x_i(t)$ may represent the time of the t^{th} occurrence of a certain event i and $A_{i,j}$ the required time lag between the $(t-1)^{\text{th}}$ occurrence of j and the t^{th} occurrence of i. Notable examples are transportation and automated manufacturing systems [49, 32, 39], network synchronizers [65, 41], and cyclic scheduling [52]. Charron-Bost et al. [25, 26] have shown that it also encompasses the behavior of an important class of distributed algorithms, namely *link reversal algorithms* [47], which can be used to solve a variety of problems [90] like routing [47] or resource allocation [23].

If one allows the $A_{i,j}$ to be $-\infty$, then we can choose $\mathbf{N}_i = [n]$ for all $i \in [n]$ in (2.1). Hence the collection of the $A_{i,j}$ can be seen as a matrix in $\mathbb{R} \cup \{-\infty\}$. This translates into

$$\forall t \geq 1 \;\; \forall i \in [n] \;\; : \;\; x_i(t) = \max_{1 \leq j \leq n} \left(x_j(t-1) + A_{i,j} \right) \tag{2.2}$$

as the governing recurrence for $x(t)$.

Recurrences of the form (2.1) are linear in the *max-plus algebra* (e.g., [56]). The fundamental theorem in max-plus linear algebra—an analog of the Perron-Frobenius theorem in classical algebra—states that the sequence of powers of an irreducible matrix A becomes periodic after a finite index called the *transient* of the matrix. A matrix is irreducible if its corresponding digraph is strongly connected. As an immediate corollary, any max-plus linear system with an irreducible matrix is periodic from some index transient of the system, which clearly is at most equal to the transient of the system's matrix.

The exact sense in which periodicity manifests itself for linear systems is that there is some transient T, a period p, and a real number α such that

$$\forall t \geq T \colon \; \forall i \in [n] \colon \quad x_i(t+p) = x_i(t) + \alpha \;\; , \tag{2.3}$$

i.e., after the transient, during one period, every vector entry changes by a common additive constant. Setting $\varrho = \alpha / p$, we can rewrite (2.3) as

$$\forall t \geq T\colon\ \forall i \in [n]\colon \quad x_i(t+p) = x_i(t) + p \cdot \varrho \ . \tag{2.4}$$

This representation justifies to dub this form of periodicity as having a *linear defect* with ratio ϱ. Clearly, for a given system $x(t)$ periodic in the sense of (2.4), the ratio ϱ is a fundamental performance parameter. If $x(t)$ satisfies recurrence (2.2), then its ratio can be determined with the digraph defined by matrix A: On the set of nodes $[n]$, it is defined as containing an edge (i,j) if and only if $A_{i,j} \neq -\infty$, i.e., $A_{i,j} \in \mathbb{R}$. In view of the event system interpretation, it contains an edge from i to j if the time of the $(t+1)^{\text{th}}$ occurrence of event i may depend on the time of the t^{th} occurrence of event j. We denote this digraph by $G(A)$. Assigning every edge (i,j) in $G(A)$ the weight $A_{i,j}$, the ratio of system $x(t)$ is equal to the largest mean weight of cycles in the digraph $G(A)$. In particular, it is independent of the initial vector $x(0)$. We will further explore and exploit the digraph representation.

The sense in which (2.2) is *linear* is in the max-plus algebra. It is adapted to describe synchronizing discrete event systems. The domain is the one-side extended real line $\mathbb{R}_{\max} = \mathbb{R} \cup \{-\infty\}$. Its additive operation is $a \oplus b = \max\{a,b\}$ and its multiplicative operation is $a \otimes b = a + b$. With these two operations, $\mathbb{R}_{\max}$ is a commutative semi-ring. Its zero element is $-\infty$ and its unity is 0. Every nonzero element has a multiplicative inverse, but the only element having an additive inverse is the zero element. Mimicking the definition of the matrix product in classical algebra, the product of two max-plus matrices $A \in \mathbb{R}_{\max}^{m\times n}$ and $B \in \mathbb{R}_{\max}^{n\times p}$ is the $m \times p$ max-plus matrix defined by

$$(A \otimes B)_{i,j} = \bigoplus_{k=1}^{n} A_{i,k} \otimes B_{k,j} = \max_{1\leq k\leq n} \left(A_{i,k} + B_{k,j}\right) \ . \tag{2.5}$$

A special case is the application of a $m \times n$ max-plus matrix A to a vector $v \in \mathbb{R}_{\max}^{n}$ defined as

$$(A \otimes v)_i = \max_{1\leq j\leq n} \left(A_{i,j} + v_j\right) \ . \tag{2.6}$$

Hence the sequence $x(t)$ defined by (2.2) fulfills the max-plus linear recurrence $x(t+1) = A \otimes x(t)$. Using the associativity of the max-plus matrix product, we hence have $x(t) = A^{\otimes t} \otimes x(0)$ where $A^{\otimes t}$ denotes the t^{th} max-plus power of matrix A. The entries of the matrix power $A^{\otimes t}$ can be characterized in terms of its digraph $G(A)$. The $(i,j)^{\text{th}}$ entry of $A^{\otimes t}$ is equal to the maximum weight of a walk of length t from node i to j in the digraph. This interpretation gives the intuition that, if A is irreducible, i.e., $G(A)$ is strongly connected, then every entry of the matrix powers $A^{\otimes t}$ are eventually periodic with linear defect, and that the ratio is equal to the maximum cycle mean in $G(A)$: As t grows, the largest part of walks of length t is composed of cycles, and the maximum mean cycles have the best weight-to-length ratio. These maximum mean cycles are therefore called *critical* cycles and they eventually govern the maximum weight walks of length t. Indeed, it turns out that the sequence of powers of every irreducible max-plus matrix A is eventually periodic with linear defect, i.e.,

$$\forall t \geq T : \quad A^{\otimes t+p} = A^{\otimes t} + p \cdot \lambda \tag{2.7}$$

for some transient T, some period p, λ being the maximum mean weight of cycles in $G(A)$, and the addition in (2.7) is understood to be component-wise. Hence, in particular, every

max-plus linear system with system matrix A is eventually periodic with linear defect λ and transient at most T.

For all the above mentioned applications, the study of the transient plays a key role in characterizing the system performance: For example, in the case of link reversal routing, the system transient corresponds to the time complexity of the routing algorithm. Besides that, understanding matrix and system transients is of interest on its own for the theory of max-plus algebra.

2.2 Applications

In this section we present how performance analysis of various distributed systems relate to studying the transient of the max-plus linear systems that model them.

2.2.1 Synchronizers

One obvious application of max-plus linear systems in distributed computing are network synchronizers. Their task is to simulate a completely synchronous round structure on top of a not completely synchronous system. We present a particular instance of a classical synchronizer in a particular type of distributed system, and explain how it is modeled in max-plus algebra.

Consider a system of n processes with identifiers $1, 2, \ldots, n$, interconnected by a message-passing system whose network digraph is strongly connected. We assume that all processes take their steps synchronously at all integral times $t = 0, 1, 2, 3, \ldots$, but that there are non-uniform message delays in the system. Assume further that the message delay from process i to process j is constant and equal to $\delta(i, j)$. They run the following algorithm: Every process has a local variable R_i and a local array with a value $L_i[j]$ for every other process j. Every variable is initialized to 0. At every time step, if all values $L_i[j]$ in the array are equal, then the process increments R_i and sends a message including the new value of R_i to all its neighbors. The use of this algorithm is that R_i is a global round number that advances only if the process has received the message of round $R_i - 1$ from all its neighbors. Hence, one can piggy-back to the messages the payload of a simulated algorithm designed for completely synchronized rounds.

The behavior of this synchronization algorithm can be modeled by a max-plus linear system with initial vector $x(0)$ whose all entries are 0 and system matrix

$$A_{i,j} = \begin{cases} \delta(j,i) & \text{if } (j,i) \text{ is a link in the network} \\ -\infty & \text{else .} \end{cases} \tag{2.8}$$

Then, process i starts round t, i.e., sets R_i to t for the first time, at time $x_i(t)$.

Even and Rajsbaum [41] studied the transience of such a network synchronizer in a system with constant integer communication delays. They actually considered a variant of the classical α-synchronizer [6] in a centrally clocked distributed system of n processes that communicate by message passing over a strongly connected network digraph G. Each link has constant transmission delay, specified in terms of central clock ticks. Processes execute the α-synchronizer after an initial boot-up phase: Even and Rajsbaum showed that the synchronizer becomes periodic by time $O(\Delta n^3)$, where Δ is the maximum delay.

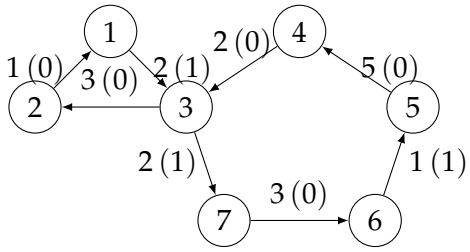


Figure 2.1: Uniform graph G^u. Edges are labeled with processing times, and heights in parentheses

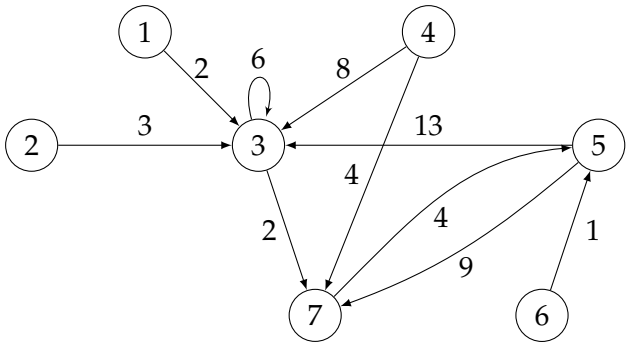


Figure 2.2: Digraph $G(A)$

2.2.2 Cyclic Scheduling

Cohen et al. [30] observed that, in cyclic scheduling, the class of earliest schedules can be described in terms of a max-plus linear systems.

If a finite set $\mathcal{T}$ of tasks (each of which calculates a certain function) is to be scheduled repeatedly on different processes, precedence restrictions are implied by the data flow. These restrictions are of the form that task i may start its number n execution only after task j has finished its number $n-h$ execution. A *schedule* σ maps a pair $(i,n) \in \mathcal{T} \times \mathbb{N}_0$ to a nonnegative integer $\sigma(i,t)$, the t^{th} execution of task i is started. Formally, if P_i denotes the processing time of task i, then a *restriction* R between two tasks i and j is an inequality of the form

$$\forall t \geq h_R : \ \sigma(i,t) \geq \sigma(j, t-h_R) + P_j \tag{2.9}$$

where h_R is called the *height* of restriction R and P_j is its *weight*.

A *uniform graph* [52] describes a set of tasks and restrictions. Formally, it is a quadruple $G^u = (\mathcal{T}, E, p, h)$ such that $(\mathcal{T}, E)$ is a directed (multi-)graph, and $p : E \to \mathbb{N}_0^*$ and $h : E \to \mathbb{N}_0$ are two functions, the *weight* and *height* function, respectively. For a walk W in G^u, let $p(W)$ be the sum of the weights of its edges and $h(W)$ the sum of the heights of its edges. An edge from i to j corresponds to a restriction R between i and j of the form (2.9). All incoming edges of a node j in $\mathcal{T}$ have the same weight, namely P_j. An example of a uniform graph is Figure 2.1.

Call G^u *well-formed* if it is strongly connected and does not contain a nonempty closed walk of height 0. Call a schedule σ an *earliest schedule* if it satisfies all restrictions specified by G^u and it is minimal with respect to the point-wise partial order on schedules. Denote

the maximum height in G^u by $\hat{h}$. Cohen et al. [30] showed that the earliest schedule σ for well-formed G^u is unique and fulfills

$$\sigma(i,t) = (A^{\otimes t} \otimes v)_i \tag{2.10}$$

for all $i \in \mathcal{T}$ and $t \geq 0$, where v is a suitably chosen $(\hat{h} \cdot |\mathcal{T}|)$-dimensional max-plus vector and A a suitably chosen $(\hat{h} \cdot |\mathcal{T}|) \times (\hat{h} \cdot |\mathcal{T}|)$ max-plus matrix. In case heights in G^u are binary, i.e., either 0 or 1, as in our example in Figure 2.2, A and v are obtained as follows: For all $i, j \in \mathcal{T}$, $A_{i,j}$ is the maximum weight of nonempty walks W from i to j in G^u, where all of W's edges have height 0, except for the last edge, which has height 1. In case no such walk exists, $A_{i,j} = -\infty$. For all $i \in \mathcal{T}$, v_i is the maximum weight of walks W from i in G^u, where all of W's edges have height 0. As an example the digraph $G(A)$ for the uniform graph in Figure 2.1 is depicted in Figure 2.2. For this example we obtain the initial vector $v = (0, 1, 4, 6, 11, 0, 3)$.

2.2.3 Full Reversal Routing and Scheduling

Link reversal is a versatile algorithm design paradigm, which was, in particular, successfully applied to routing [47] and scheduling [8]. Charron-Bost et al. [26] showed that the analysis of a general class of link reversal algorithms can be reduced to the analysis of Full Reversal, a particularly simple algorithm on digraphs.

The Full Reversal algorithm comprises only a single rule: Each sink reverses all its (incoming) edges. Given a weakly connected initial digraph G_0 without anti-parallel edges, we consider a *greedy* execution of Full Reversal as a sequence $(G_t)_{t\geq 0}$ of digraphs, where G_{t+1} is obtained from G_t by reversing the edges of *all* sinks in G_t. As no two sinks in G_t can be adjacent, G_{t+1} is well-defined. For each $t \geq 0$ we define the *work vector* $W(t)$ by setting $W_i(t)$ to the number of reversals of node i until iteration t, i.e., the number of times node i is a sink in the execution prefix $G_0, \ldots, G_{t-1}$.

Charron-Bost et al. [25] have shown that the sequence of work vectors can be described as a *min-plus* linear dynamical system. Min-plus algebra is a variant of max-plus algebra, using min instead of max. Denoting by $\otimes'$ the matrix multiplication in min-plus algebra, Charron-Bost et al. established that $W(0) = 0$ and $W(t+1) = A \otimes' W(t)$, where $A_{i,j} = 1$ and $A_{j,i} = 0$ if (i, j) is an edge of the initial digraph G_0; otherwise $A_{i,j} = +\infty$. Observe that the latter min-plus recurrence is equivalent to $-W(t+1) = (-A) \otimes (-W(t))$ where $-A$ is a max-plus matrix, i.e., has entries in $\mathbb{R} \cup \{-\infty\}$.

In the routing case, the initial digraph G_0 contains a nonempty set of *destination nodes*, which are characterized by having a self-loop. The initial digraph without these self-loops is required to be weakly connected and acyclic [25, 47]. It was shown that for such initial digraphs, the execution terminates (eventually all G_t are equal), and after termination, the digraph is destination-oriented, i.e., every node has a walk to some destination node. The set of critical nodes is equal to the set of destination nodes and each strongly connected component of the subgraph defined by the critical cycles consists of a single node.

When using the Full Reversal algorithm for scheduling, the undirected support of the weakly connected initial digraph G_0 is interpreted as a conflict digraph: nodes model processes and an edge between two processes signifies the existence of a shared resource whose access is mutually exclusive. The direction of an edge signifies which process is allowed to use the resource next. A process waits until it is allowed to use all its resources—that is, it waits until it is a sink—and then performs a step, that is, reverses all edges to release its resources. To guarantee liveness, the initial digraph G_0 is required to be acyclic: Cycles remain

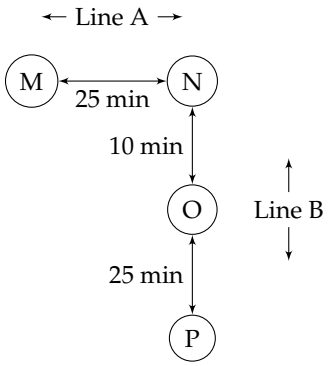


Figure 2.3: A simple train network

constant during every execution of Full Reversal. If there are cycles in the initial digraph, the execution halts it a state in which every node has a path to some node on an (initial) cycle. On the other hand, no new cycles are formed during a step of the Full Reversal algorithm. Hence if the initial digraph is acyclic, then every digraph in every execution is acyclic.

2.2.4 Timed Petri Nets and Transportation Systems

Max-plus systems exactly correspond to a subclass of time Petri nets, called timed event graphs, timed marked graphs, or timed decision-free Petri nets. Petri nets are bipartitioned into two types of nodes: places and transitions. Event graphs are those Petri nets whose places have exactly one incoming and one outgoing transition. The qualifies "timed" refers to the fact that every place has a corresponding holding time for tokens that determines how long a token has to reside in the place before it can be consumed by the subsequent transition. One can identify transitions with nodes and places with edges to get an edge-weighted digraph, which corresponds to a max-plus matrix A. If all places initially have one token and there is at most one place between all pairs of transitions, then $x(t+1) = A \otimes x(t)$ where $x_i(t)$ denotes the time that transition i fires for the t^{th} time. We assume that the initial firing times $x(0)$ are given. Hence we can directly apply our transience bounds to strongly connected timed event graphs that initially have one token at every place. If the number of tokens is different, then one nonetheless identify the Petri net with a max-plus linear system, but the dimension will be higher and it is not necessarily irreducible, even if the original Petri net is strongly connected.

A particular instance of timed event graphs that has attracted a lot of attention are train networks [56, Chapter 8]. The Petri net description includes the travel time between stops, the desired connecting train relations, as well as the time required to change trains. It has one transition for every outgoing track at every station. The places and edges are defined by the line's trajectory and the connecting train relation. Consider the simple example network in Figure 2.3. It includes four stations, named M, N, O, and P. It has two train lines: Line A going from M to N and back; and Line B going from N to P via O and back. Each of the railway segments MN, NO, and OP possesses a certain travel time that the trains need to cover it; namely 25, 10, and 25 minutes, respectively. If one assumes that lines A and B should wait for each other at station N and that initially there is one train leaving every station in every possible direction, then the Petri net yields an irreducible max-plus matrix describing the

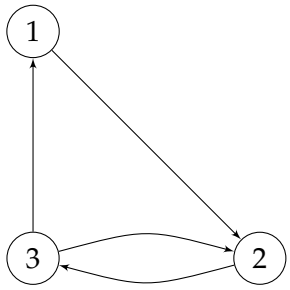


Figure 2.4: A primitive digraph

necessary travel and changeover times in the train system. That is, the earliest times that the trains can leave is described by a max-plus linear system $x(t) = A^{\otimes t} \otimes x(0)$ with irreducible system matrix A.

However, trains do not always depart at the earliest possible time because they need to satisfy a pre-assigned schedule. Schedules are commonly purely periodic, i.e., of the form $y_i(t) = y_i(0) + t \cdot \Delta$ where Δ is the schedule's temporal period. The actual departure times in a system respecting the schedule are then given by $z(t) = \max\{x(t), y(t)\}$.

One interesting performance parameter is the robustness of the system against delays. This can be modeled by the fact that the initial vector $x(0)$ contains the incurred delays and one seeks to know the smallest t at which the system will be on schedule again, i.e., $z(t) = y(t)$ or equivalently $y(t) \geq x(t)$. This smallest t is called the *recovery time*. Clearly, the system can only recover if $\lambda = \lambda(A) < \Delta$, i.e., if the schedule leaves some headroom. In this case, however, one can bound the recovery time in terms of the system's transient T by

$$\frac{T \cdot (A_{\max} - \lambda) + A_{\max} \cdot (\gamma_c - 1) + \max_i x_i(0)}{\Delta - \lambda} \tag{2.11}$$

where $A_{\max}$ is the maximum entry of A and γ_c is the cyclicity of the critical digraph of A. Note that the bounds on T can become prohibitively big in case of very large-scale train networks. In these cases, brute-force simulation can be a viable alternative.

2.2.5 Walks in Digraphs

Powers of a max-plus matrices also correspond to maximum weight walks in edge-weighted digraphs between two fixed nodes and of fixed length. In the particular case of non-weighted digraphs, they contain the information whether there exists a walk between two given nodes with a given length.

The periodicity of walk lengths in digraphs has been established and studied extensively, in particular for applications such as automata theory [74], but also for their general graph-theoretic interest [18]. The case of primitive matrices has attracted particular interest. Primitive digraphs are those that are connected and whose greatest common divisor of their cycle lengths is 1. Figure 2.4 shows an example of a primitive digraph; it would not be primitive without the edge from node 3 to node 2. An alternative definition for primitive digraphs is that some power of their adjacency matrix is strictly positive. In more concrete terms, this characterization reads:

Theorem 2.1 ([18, Theorem 3.4.4]). *A digraph $G = (V, E)$ is primitive if and only if there exists a nonnegative integer T such that*

$$\forall t \geq T \;\; \forall i, j \in V \;\; \exists W\colon \quad W \text{ is a walk from } i \text{ to } j \text{ with } \ell(W) = t \;\; . \tag{2.12}$$

Given a digraph $G = (V, E)$, denote by G^t the digraph with node set V containing an edge (i, j) if and only if there is a walk from i to j of length t in G. In particular, G^0's edges are the self-loops at all nodes. Condition (2.12) in Theorem 2.1 is equivalent to the fact that there exists some T such that G^T, and therefore all G^t with $t \geq T$, is the complete digraph. The smallest T with this property is commonly called the exponent of G. The exponent of the digraph in Figure 2.4 is equal to 5. Hence, if G is primitive, then the sequence of digraphs $G^0, G^1 = G, G^2, \ldots$ is eventually constant, and equal to the complete digraph, after its exponent. Another way of saying this is that the sequence is eventually periodic with period 1 and transient T equal to its exponent, i.e., $G^t = G^{t+1}$ for all $t \geq T$. In fact, even for non-primitive digraphs, this sequence is eventually periodic. However, the period is not necessarily 1, nor is it eventually equal to the complete digraph if it is.

Theorem 2.2 (Schwarz [81, Theorem 4.3]). *Let G be a digraph. Then the sequence of digraphs G^t is eventually periodic.*

We will call the transient of the sequence of digraphs $G^0, G^1, G^2, \ldots$ the *index of convergence* and denote it by $\mathrm{ind}(G)$. It is exactly equal to the transient of the sequence of matrix powers $A^{\otimes t}$ where matrix A is defined by setting $A_{i,j} = 0$ if (i, j) is an edge of G and $A_{i,j} = -\infty$ else. Matrices whose entries are in the two-element set $\{-\infty, 0\}$ are called Boolean matrices. Every product, and hence in particular every power, of Boolean matrices is Boolean. The study of transients of max-plus matrices, i.e., the study of maximum weight walks in weighted digraphs includes the study of Boolean matrices, i.e., unweighted digraphs, when setting all edge weights to 0.

2.3 State of the Art

2.3.1 Definitions and Preliminaries

A digraph is a pair $G = (V, E)$ of a nonempty set V of nodes and a set $E \subseteq V \times V$ of edges. A walk W in G is a finite sequence of nodes $i_0, i_1, \ldots, i_\ell$ such that $(i_{r-1}, i_r) \in E$ for all $1 \leq r \leq \ell$. We write $\ell(W) = \ell$ for its length. It is empty if $\ell = 0$ and closed if $i_\ell = i_0$. A walk in the digraph is a *path* if every node occurs only once. A closed walk is a *cycle* if it is nonempty and only the start and end node occurs twice.

The length of the shortest cycle in a digraph G is called the *girth* of G. If a digraph is strongly connected, the greatest common divisor of its cycle lengths is called its *cyclicity*. The cyclicity of a (possibly not strongly connected) digraph is the least common multiple of the cyclicities of its strongly connected components. The following two lemmas explicit the role and definition of the cyclicity of strongly connected digraphs.

Lemma 2.3. *Let G be a strongly connected digraph with cyclicity γ. Then the relation on the set of nodes of G defined by*

$$i \sim j \iff \text{there is a walk } W \text{ from } i \text{ to } j \text{ with } \ell(W) \equiv 0 \pmod{\gamma} \tag{2.13}$$

is an equivalence relation. It has exactly γ equivalence classes.

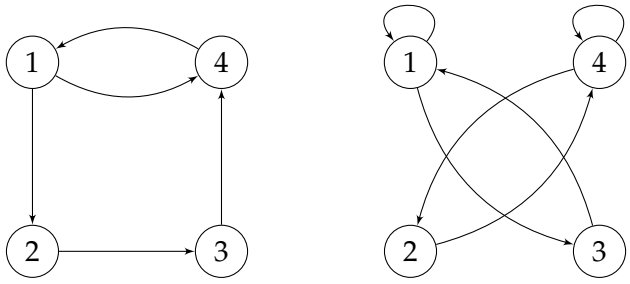


Figure 2.5: A digraph with cyclicity 2 and its completely reducible square

Lemma 2.4. *Let G be a strongly connected digraph with cyclicity γ. Then, for all nodes i and j and all walks W_1, W_2 from i to j, we have the congruence of lengths $\ell(W_1) \equiv \ell(W_2) \pmod{\gamma}$.*

A particular application of this notion is the existence of closed walks whose length are an arbitrary, sufficiently large, multiple of the cyclicity. Recall that $\mathrm{ind}(G)$ denotes the index of convergence of G.

Lemma 2.5. *Let G be a strongly connected digraph with cyclicity γ. Then for all nodes i and all multiples t of γ with $t \geq \mathrm{ind}(G)$, there exists a closed walk at i of length t.*

The cyclicity has the property that it is an exponent for which a strongly connected digraph is *completely reducible*, i.e., there are no edges between distinct strongly connected components. In a completely reducible digraph, the strongly and weakly connected components coincide.

For a digraph G and a nonnegative integer m, denote by G^m the mth Boolean power of G, i.e., the digraph that contains an edge (i, j) if and only if there exists a walk from i to j of length m in G.

Theorem 2.6 ([18, Theorem 3.4.5])**.** *Let G be a strongly connected digraph with cyclicity $\gamma = \gamma(G)$. Then G^γ is completely reducible and its components have cyclicity 1.*

To every $n \times n$ max-plus matrix A corresponds a digraph $G(A)$ with node set $V = [n] = \{1, 2, \ldots, n\}$ containing an edge (i, j) if and only if $A_{i,j} \neq -\infty$. We refer to $A_{i,j}$ as the weight of edge (i, j). Matrix A is *irreducible* if $G(A)$ is strongly connected. If W is a walk in $G(A)$, we define its weight $A(W)$ as the sum of the weights of its edges. The entry $A^{\otimes t}_{i,j}$ is the maximum weight of walks from i to j of length t. We follow the convention that $\max \emptyset = -\infty$. If v is a max-plus column vector of size n, then the entry $\left(A^{\otimes t} \otimes v\right)_i$ is the maximum of the values $A(W) + v_j$ where the maximum is formed over all nodes j and all walks W from i to j of length t.

Denote by $\lambda(A)$ the maximum mean weight $A(Z)/\ell(Z)$ of cycles in $G(A)$. We call *critical* every cycle with maximum mean weight. The sub-digraph of $G(A)$ induced by edges on critical cycles is called its *critical digraph*.

2.3.2 Eventually Periodic Sequences

Let $p \geq 1$ and $\varrho \in \mathbb{R}$. A sequence $f : \mathbb{N}_0 \to \mathbb{R}_{\max}$ is *eventually periodic with period p and ratio ϱ* if there exists a $T \in \mathbb{N}_0$ such that

$$\forall t \geq T \,:\, f(t+p) = f(t) + p \cdot \varrho \ . \tag{2.14}$$

Obviously, if q is a multiple of p, then f is also eventually periodic with period q and ratio ϱ. Hence there always exists a common period of two eventually periodic sequences.

If an eventually periodic sequence f is not eventually equal to $-\infty$, then it has a unique ratio ϱ. Otherwise, f is eventually periodic with respect to all p and all ϱ. In both cases, for a given period, the set of T that satisfy (2.14) is independent of ϱ. In fact it does not depend on p either, as is stated in the next lemma. We call the minimal $T \in \mathbb{N}_0$ satisfying (2.14) the *transient* of f.

Lemma 2.7. *Let $f : \mathbb{N}_0 \to \mathbb{R}_{\max}$ be eventually periodic and let $\varrho \in \mathbb{R}$. Then the set of $T \in \mathbb{N}_0$ that satisfy* (2.14) *is independent of p.*

The notion of eventual periodicity naturally extends to matrices (and thus also vectors): A sequence of matrices $S(t)$ is *eventually periodic with period p and ratio ϱ* if each entry-wise sequence $S_{i,j}(t)$ is eventually periodic with period p and ratio ϱ. Its *transient* is the maximum transient of the $S_{i,j}(t)$.

Cohen et al. [31] showed that the sequence of powers of an irreducible max-plus matrix, and hence of all systems with irreducible matrix, are eventually periodic. Denote by $\gamma_c(A)$ the cyclicity of the critical digraph of $G(A)$.

Theorem 2.8 (Cohen et al. [31])**.** *The sequence of powers $A^{\otimes t}$ of an irreducible square max-plus matrix A is eventually periodic with ratio $\lambda(A)$ and period $\gamma_c(A)$.*

Theorem 2.8 is based on the fact that maximum weight walk eventually include in the majority critical cycles. To give an explicit upper bound on when they visit at least one critical cycle, several authors defined what they considered to be the "second most significant" cycle mean. This can be done in a number of ways, and it depends on their use which one is the most appropriate. One possibility is to consider the second largest cycle mean $\lambda_2(A)$. Another possibility is the largest cycle mean disjoint to all critical cycles, which we denote by $\lambda_{nc}(A)$. Hartmann and Arguelles [53] introduced a third parameter, $\lambda_{\mathrm{ha}}(A)$, which is defined in terms of the max-balancing [78] of $G(A)$. We do not yet formally define all three parameters, but give their relative ordering, also with respect to $\lambda(A)$:

$$\lambda(A) > \lambda_2(A) \geq \lambda_{nc}(A) \geq \lambda_{\mathrm{ha}}(A) \tag{2.15}$$

We denote by $\|A\|$ the difference between the greatest and smallest finite entry in A.

2.3.3 Boolean Matrices

The first bound on the index of convergence of digraphs was given by Wielandt [91] for the case of primitive digraphs, i.e., digraphs whose cyclicity is equal to 1. It was proved for general digraphs by Schwarz.

Theorem 2.9 (Schwarz [81, Theorem 4.3])**.** *The transient of an $n \times n$ Boolean matrix is at most*

$$\mathrm{Wi}(n) = \begin{cases} (n-1)^2 + 1 & \text{if } n \geq 1 \\ 0 & \text{if } n = 0 \ . \end{cases} \tag{2.16}$$

The bound of $\mathrm{Wi}(n)$ was refined independently by Dulmage and Mendelsohn [40] and by Denardo [34] in terms of the digraph's girth g. They arrived at the same bound, which is in the order of $O(g \cdot n)$. This suggests that the lower the girth, the lower the transient.

Theorem 2.10 (Dulmage and Mendelsohn [40]). *The transient of an irreducible primitive $n \times n$ Boolean matrix A is at most $g \cdot (n-2) + n$ where g denotes the girth of $G(A)$.*

Schwarz [80] extended Theorem 2.9 to non-primitive irreducible matrices in a different direction. He showed that the bound of $(n-1)^2 + 1$ remains true and that even a lower upper bound holds, which in the order of $O(n^2/\gamma)$ where γ denotes the cyclicity. This suggests that the higher the cyclicity, the lower the transient. For instance, Schwarz's bound shows that the transients of bi-partite graphs, for which γ is a multiple of 2, are at most $n^2/2 - 2n + 5$.

Theorem 2.11 (Schwarz [80]). *The transient of an irreducible $n \times n$ Boolean matrix A is at most $\gamma \cdot \mathrm{Wi}(\lfloor n/\gamma \rfloor) + (n \bmod \gamma)$ where γ is the cyclicity of $G(A)$.*

Because the girth of a strongly connected digraph is always greater or equal to the cyclicity, the two results suggest a necessary trade-off between the two parameters for attaining a small transient. For instance, the two parameters need to be equal for attaining the minimal transient of 0.

Kim [59] showed an upper bound involving *both* the girth and the cyclicity.

Theorem 2.12 (Kim [59]). *The transient of an irreducible $n \times n$ Boolean matrix A is at most $g \cdot \big(\lfloor n/\gamma \rfloor - 2\big) + n$ where γ is the cyclicity and g the girth of $G(A)$.*

The *factor rank* of an $m \times n$ max-plus matrix A is the least number $r = \mathrm{rk}(A)$ for which there exist vectors $v_1, v_2, \ldots, v_r \in \mathbb{R}_{\max}^m$ and $w_1, w_2, \ldots, w_r \in \mathbb{R}_{\max}^n$ such that

$$A = \bigoplus_{\alpha=1}^{r} v_\alpha \otimes {}^t w_\alpha \ . \tag{2.17}$$

It is also sometimes called the Boolean rank or the Schein rank. The factor rank of an $n \times n$ max-plus matrix is at most n because one can choose w_α to be the α^{th} max-plus unit vector and $(v_\alpha)_i = A_{i,\alpha}$.

Two bounds involving the factor rank were proved for primitive Boolean matrices:

Theorem 2.13 (Gregory-Kirkland-Pullman [50]). *The transient of a primitive irreducible Boolean matrix with factor rank r is at most* $\mathrm{Wi}(r) + 1$.

Theorem 2.14 (Kim [59]). *The transient of a primitive irreducible Boolean matrix A with factor rank r is at most $g \cdot (r-2) + r$ where g is the girth of $G(A)$.*

2.3.4 Nachtigall Decomposition

A significant step in the direction of a transience bound for non-Boolean matrices was done by Nachtigall [72]. While he did not prove a bound on the transient, he showed that the sequence of matrix powers can be written as a maximum of eventually periodic sequences with bounded transients. Such a decomposition in the form of a maximum, by itself, does not yield a bound on the transient of the original sequence; it does not even imply that it is eventually periodic. As a matter of fact, Nachtigall shows the existence of such a decomposition not only for irreducible matrices, but for general square max-plus matrices, for which the sequence of powers is not necessarily eventually periodic.

Theorem 2.15 (Nachtigall [72]). *Let A be an $n \times n$ max-plus matrix. Then there exist eventually periodic matrix sequences $A_1(t), A_2(t), \ldots, A_n(t)$ with transients at most $3n^2$ such that for all $t \geq 0$:*

$$A^{\otimes t} = A_1(t) \oplus A_2(t) \oplus \cdots \oplus A_n(t) \tag{2.18}$$

Nachtigall proved Theorem 2.15 by recursively picking a cycle Z with maximum weight-to-length ratio $A(Z)/\ell(Z)$ and by partitioning the sets of walks in $G(A)$ into the sets of walks that do and do not visit cycle Z. Walks that do not visit Z are walks in the sub-digraph of $G(A)$ that has all edges incident to Z removed. This sub-digraph is the digraph of the matrix obtained from A by setting to $-\infty$ all rows and columns corresponding to nodes in Z; its effective size is strictly smaller than the size of A, which enables a recursive descent. If no cycle exists in $G(A)$ at all, then the transient of A is at most n since in this case, $A_{i,j}^{\otimes t} = -\infty$ for all i, j and all $t \geq n$.

2.3.5 Bound by Hartmann and Arguelles

Hartmann and Arguelles [53] gave the first general transience bound for arbitrary irreducible max-plus matrices. Their proof is purely graph-theoretic.

When analyzing their proof, one can extract a global proof strategy, variants of which are also found in later proofs of transience bounds [84]. In order to prove that some number B is an upper bound on the transient of the sequence $A^{\otimes t}$ for an irreducible matrix A, do the following:

1. Show that one can assume $\lambda(A) = 0$, i.e., the sequence $A^{\otimes t}$ is eventually periodic with ratio 0.

2. Fix two nodes i and j, and a congruence class C_0 modulo some period p of the sequence $A^{\otimes t}$.

3. The assumption $\lambda = 0$ guarantees that the maximum $\max_{k \in M} A_{i,j}^{\otimes t}$ formed over an arbitrary nonempty set M of nonnegative integers exists. We choose the set M to consist of those elements of class C_0 that are greater or equal to B. Since the maximum exists, there exists a walk W from i to j with length in M that attains it. If B is indeed an upper bound on the transient, the values $A_{i,j}^{\otimes t}$ with $t \in M$ will all be equal.

4. Show that, whenever the length of W is greater or equal to some "critical bound" $B_c \leq B$, then it necessarily shares a node with a critical cycle Z.

5. Show that one can reduce walk W by removing subcycles such that it is possible to attain all lengths in M greater or equal to some "pumping bound" $B_p \leq B$ by adding critical cycles. The assumption $\lambda = 0$ implies that all subcycles have weight at most 0 and critical cycles have weight equal to 0. Thus the weights of walks obtained in this way cannot be lower than that of W; hence they are equal to that of W.

6. We then have shown, because the choice of C_0 was arbitrary, that the transient of $A_{i,j}^{\otimes t}$ is at most $B \geq \max\{B_c, B_p\}$.

Hartmann and Arguelles used $p = \gamma_c(A)$ in step (2). For step (5), they described a walk reduction based on the following basic application of the pigeonhole principle:

Lemma 2.16. *Let d be a positive integer. Every collection of at least d integers has a nonempty subcollection whose sum is divisible by d.*

They used this lemma to reduce walk W in step (5). After their reduction their walk could be disconnected, but they showed that adding a copy of (critical) cycle Z reestablishes connection [53, Theorem 4]:

Lemma 2.17 (Hartmann and Arguelles [53])**.** *Let W be a walk that shares a node with some cycle Z and let t be an integer such that $t \equiv \ell(W) \pmod{\ell(Z)}$ and $t \geq n^2$ where n denotes the number of nodes in the graph. Then there exists a walk $\tilde{W}$ obtained from W by removing cycles and possibly adding copies of Z such that $\ell(\tilde{W}) = t$.*

To pump the walk length after the walk reduction, they used a result by Brauer [17] on the Frobenius problem to combine critical cycles to attain a multiple of $\gamma_c(A)$. The use of Brauer's theorem introduces a term that is necessarily quadratic in n to the transience bound. We want to note at this point that this use of Brauer's theorem can be avoided by considering a period in step (2) different from the critical digraph's cyclicity because of Lemma 2.7. Hartmann and Arguelles actually prove Lemma 2.7 later in the paper [53, Lemma 11], but do not use it in the proof of their transience bound.

The same strategy as described above can be adapted to show transience bounds for systems $A^{\otimes t} \otimes v$. In the case that all entries of v are finite, it is possible to show a sharper bound because the walks under consideration do not have both the start and the end node fixed, but only the start node. This allows to circumvent the necessity of showing the existence of walks of prescribed length between two fixed nodes (see Section 2.3.3).

Theorem 2.18 (Hartmann and Arguelles [53])**.** *Let A be an irreducible $n \times n$ max-plus matrix. Then the transient of the sequence of powers $A^{\otimes t}$ is at most*

$$\max\left\{2n^2\,,\ \frac{2n^2\|A\|}{\lambda(A) - \lambda_{\mathrm{ha}}(A)}\right\}\ . \tag{2.19}$$

If, additionally, v is a column vector of size n with only finite entries, then the transient of the system $A^{\otimes t} \otimes v$ is at most

$$\max\left\{2n^2\,,\ \frac{\|v\| + n\|A\|}{\lambda(A) - \lambda_{\mathrm{ha}}(A)}\right\}\ . \tag{2.20}$$

Hartmann and Arguelles also proved a form of asymptotic tightness of their transience bound for matrices. They gave, for every n of the form $n = 3m - 1$ and all positive reals λ and λ_{ha} with $\lambda > \lambda_{\mathrm{ha}}$, an irreducible $n \times n$ max-plus matrix A with $\lambda(A) = \lambda$ and $\lambda_{\mathrm{ha}}(A) = \lambda_{\mathrm{ha}}$ (see [53, Figure 1]). Their example has the property that $\lambda_{\mathrm{ha}}(A) = \lambda_{nc}(A) = \lambda_2(A)$ and $\|A\| = \lambda$. They showed by explicit calculation that A's transient is at least $3 + m(m-2)\lambda/(\lambda - \lambda_{\mathrm{ha}})$.

We can generalize their example to arbitrary n by inserting additional nodes that do not change the transient. This then shows that, even if one can prescribe all the other parameters in the matrix bound of Theorem 2.18, it is asymptotically tight when n tends to infinity:

Theorem 2.19. *Let D_n and M_n be two sequences of positive real numbers such that $D_n \leq M_n$. Then there exists a sequence of irreducible $n \times n$ max-plus matrices A_n such that $\lambda(A_n) - \lambda_2(A_n) = D_n$, $\|A_n\| = M_n$, and the transient of the sequence of matrix powers $A_n^{\otimes t}$ is*

$$\Omega\left(\frac{n^2\|A_n\|}{\lambda(A_n) - \lambda_2(A_n)}\right)\ . \tag{2.21}$$

Because $\lambda_2 = \lambda_{nc} = \lambda_{\text{ha}}$ in Hartmann and Arguelles' example, Theorem 2.19 also holds with either λ_{nc} or λ_{ha} replacing λ_2.

2.3.6 A Bound for Primitive Matrices

A certain class of graph-theoretic arguments has been developed for the case that the matrix is *primitive*, i.e., if its critical digraph has a cyclicity equal to 1. This definition is consistent with the definition of primitivity for Boolean matrices (Section 2.3.3) because all cycles are critical in the Boolean case. This class of arguments was used by both Akian et al. [1, Remark 7.14] and Bouillard and Gaujal [14]. The resulting bound is at most quadratic in the number of nodes, and is in general incomparable with the bound of Hartmann and Arguelles.

Bouillard and Gaujal explained how to extend their result to the case of non-primitive matrices: If A's critical digraph has cyclicity γ_c, then $A^{\otimes\gamma_c}$ is primitive. It is not necessarily irreducible, but it is guaranteed to be *completely reducible*, i.e., permutation similar to a block-wise diagonal matrix whose diagonal blocks are irreducible. Also, every irreducible block contains at least one critical cycle, i.e., their eigenvalues are equal, which implies that the sequence of powers is eventually periodic. If T is the transient of the sequence $A^{\otimes k\gamma_c}$, then the transient of $A^{\otimes k}$ is at most $T\gamma_c$.

Unfortunately, the cyclicity γ_c can be exponential in the size n of the matrix. This was shown by Malka et al. [64, Theorem 4] who constructed matrices whose critical digraphs are disjoint unions of cycles of prime lengths. Using the Prime Number Theorem, one sees that it is possible to construct a critical digraph with cyclicity $\gamma_c = e^{\Omega(\sqrt{n})}$. Malka et al. improved this observation by showing that even the *minimal* period can be in the same order:

Theorem 2.20 (Malka, Moran, and Zaks [64])**.** *There exists a sequence of irreducible $n \times n$ max-plus matrices A_n such that the minimal period of the sequence of matrix powers $A_n^{\otimes t}$ is $\exp\big(\Omega(\sqrt{n})\big)$.*

2.3.7 When All Entries Are Finite

Soto y Koelemeijer [84, Theorem 3.5.12] established a transience bound in the case that all matrix entries are finite, i.e., the corresponding digraph is the complete digraph. His approach is similar to that of Hartmann and Arguelles, but the assumption of existence of all edges in the corresponding digraph allows to construct shorter walks. Utilizing this fact, he arrived at a bound that can be lower than that of Hartmann and Arguelles (first part of Theorem 2.18). But, due to the restriction to all-finite matrices, it is incomparable with the bounds of Hartmann and Arguelles and Bouillard and Gaujal.

Theorem 2.21 (Soto y Koelemeijer [84])**.** *Let A be an $n \times n$ max-plus matrix with only finite entries. Then the transient of the sequence of powers $A^{\otimes t}$ is at most*

$$\max\left\{2n^2\,,\ \left\lceil \frac{2\|A\|}{\lambda(A) - \lambda_2(A)} \right\rceil + n - 1 \right\}\ . \tag{2.22}$$

2.4 Comparison of the Existing Boolean Bounds

We begin by comparing the existing bounds in the Boolean case. It is obvious that Schwarz' bound (Theorem 2.11) is always less or equal to Wielandt's bound (Theorem 2.9) and that Kim's bound (Theorem 2.12) is is always less or equal to Dulmage and Mendelsohn's bound

(Theorem 2.10). The following theorem shows that the bound of Kim in Theorem 2.12 is tighter than the bounds of Wielandt, Schwarz, and Dulmage and Mendelsohn.

Theorem 2.22. *Let A be an irreducible $n \times n$ Boolean matrix with cyclicity γ and girth g. Then*

$$g \cdot \left(\lfloor n/\gamma \rfloor - 2\right) + n \leq \begin{cases} \mathrm{Wi}(n) \\ \gamma \cdot \mathrm{Wi}\left(\lfloor n/\gamma \rfloor\right) + (n \bmod \gamma) \\ g \cdot (n-2) + n \ . \end{cases} \tag{2.23}$$

Proof. The inequality $g \cdot \left(\lfloor n/\gamma \rfloor - 2\right) + n \leq g \cdot (n-2) + n$ is trivial.

We now show that $\gamma \cdot \mathrm{Wi}\left(\lfloor n/\gamma \rfloor\right) + (n \bmod \gamma) \leq \mathrm{Wi}(n)$. If $\gamma > n/2$, then $\lfloor n/\gamma \rfloor = 1$ and thus $\gamma \cdot \mathrm{Wi}\left(\lfloor n/\gamma \rfloor\right) + (n \bmod \gamma) = n \bmod \gamma \leq \gamma - 1 \leq n - 1 \leq \mathrm{Wi}(n)$. We hence assume $\gamma \leq n/2$, in particular $n \geq 2$, in the rest of the argument. In any case, we have

$$\gamma \cdot \mathrm{Wi}\left(\lfloor n/\gamma \rfloor\right) + (n \bmod \gamma) \leq \gamma \cdot \left((n/\gamma - 1)^2 + 1\right) + \gamma - 1 \ . \tag{2.24}$$

Setting $f(\gamma) = \gamma \cdot \left((n/\gamma - 1)^2 + 1\right) + \gamma - 1$, we have

$$\begin{aligned} \frac{d^2}{d\gamma^2} f(\gamma) &= \frac{d^2}{d\gamma^2} \gamma \cdot (n^2/\gamma^2 - 2n/\gamma + 2) - 1 \\ &= \frac{d^2}{d\gamma^2} n^2/\gamma - 2n + 2\gamma - 1 = 2n^2/\gamma^3 > 0 \end{aligned} \tag{2.25}$$

for all $\gamma > 0$. The function f is hence convex in the domain of positive reals. In particular, $f(\gamma) \leq \max\{f(1), f(n/2)\}$ for all $1 \leq \gamma \leq n/2$. It is $f(1) = \mathrm{Wi}(n)$. Because $d/d\gamma\ f(\gamma) = -n^2/\gamma^2 + 2 < 0$ for all $0 < \gamma < n/\sqrt{2}$, function f is nonincreasing in the interval $[1, n/\sqrt{2}]$. In particular, $f(n/2) \leq f(1) = \mathrm{Wi}(n)$, i.e., $f(\gamma) \leq \mathrm{Wi}(n)$ for all $1 \leq \gamma \leq n/2$. We have thus shown $\gamma \cdot \mathrm{Wi}\left(\lfloor n/\gamma \rfloor\right) + (n \bmod \gamma) \leq \mathrm{Wi}(n)$.

To prove the lemma, it hence remains to prove

$$g \cdot \left(\lfloor n/\gamma \rfloor - 2\right) + n \leq \gamma \cdot \mathrm{Wi}(\lfloor n/\gamma \rfloor) + (n \bmod \gamma) \ . \tag{2.26}$$

Because $g \leq n$ and γ divides g, we have $g/\gamma \leq \lfloor n/\gamma \rfloor$. We distinguish two cases:

1. $g \leq \gamma \cdot \left(\lfloor n/\gamma \rfloor - 1\right)$
2. $g = \gamma \cdot \lfloor n/\gamma \rfloor$

In Case 1, because $g \geq 1$, we have $\lfloor n/\gamma \rfloor \geq 2$. Thus,

$$\begin{aligned} g \cdot \left(\lfloor n/\gamma \rfloor - 2\right) + n &\leq \gamma \cdot \left(\lfloor n/\gamma \rfloor - 1\right) \cdot \left(\lfloor n/\gamma \rfloor - 2\right) + n \\ &= \gamma \cdot \left(\lfloor n/\gamma \rfloor - 1\right)^2 + \gamma + (n \bmod \gamma) \\ &= \gamma \cdot \mathrm{Wi}(\lfloor n/\gamma \rfloor) + (n \bmod \gamma) \end{aligned} \tag{2.27}$$

because $n = \gamma \cdot \lfloor n/\gamma \rfloor + (n \bmod \gamma)$.

In Case 2, denote by h the longest cycle length in G. It is a multiple of γ and satisfies $g \leq h \leq n$, which implies $g/\gamma \leq h/\gamma \leq \lfloor n/\gamma \rfloor = g/\gamma$, i.e., $h = g$. Hence all cycles in G have length g. This implies $\gamma = g$ and thus $\lfloor n/\gamma \rfloor = 1$. But in this case, both sides of (2.26) are equal to $(n \bmod \gamma)$ because $n = \gamma \cdot 1 + (n \bmod \gamma)$. This concludes the proof. □

Since the factor rank is upper-bounded by the number of nodes, the bounds involving the factor rank, i.e., Theorem 2.13 and Theorem 2.14, are at most 1 greater than the bounds of Wielandt (Theorem 2.9) and Dulmage and Mendelsohn (Theorem 2.10).

One can also specialize the more general max-plus transience bounds to Boolean matrices: The bound of Hartmann and Arguelles gives an upper bound of $2n^2$ on the transients of irreducible Boolean $n \times n$ matrices, which is strictly larger than Wielandt's bound of $(n-1)^2$. When specializing the bound of Bouillard and Gaujal, one gets an upper bound of $g \cdot (n-2) + 2n - 1$ on the transients of primitive Boolean $n \times n$ matrices whose digraph has girth g. The bound $g \cdot (n-2) + n$ of Dulmage and Mendelsohn is strictly lower. Finally, the bound of Soto y Koelemeijer gives $2n^2$ for the Boolean $n \times n$ matrix whose entries are all 0. The true transient of this matrix is equal to zero.

Chapter 3

Transients of Critical Nodes

In a max-plus linear system $x(t)$, we know that all entries $x_i(t)$ are eventually periodic with the same linear defect λ if the system matrix is irreducible. The general goal for max-plus systems in this thesis is to upper-bound the systems' transient. The existing bounds in the literature focused on the maximum of the entry-wise transients, i.e., the minimal T such that $x_i(t+p) = x_i(t) + p \cdot \lambda$ for all indices $i \in [n]$. We will also adopt this viewpoint when improving all published bounds in Chapter 4. In this chapter, however, we look at the possible discrepancy between the minimal and the maximal entrywise transient.

To do this, we upper-bound the entry-wise transients of critical nodes, i.e., the transients of those sequences $x_i(t)$ where i is a critical node in the digraph of the system matrix A. We show that these transients are at most quadratic in the number of nodes. Our bounds are independent of the exact values of the matrix entries and depend only on the matrix's digraph and its critical digraph. This is to be seen in contrast to the worst-case lower bound by Hartmann and Arguelles (Theorem 2.19), which shows that the global system transient can be arbitrarily large if $\lambda_2(A)$ is arbitrarily close to $\lambda(A)$. Hence the difference between the smallest and the largest entry-wise transient can be arbitrarily large. In applications, it is often of interest which nodes have small entry-wise transients. For example, if one can choose a location to build a home or a factory in a train network, one may prefer those locations where the recovery time in case of delays is small. A small transient gives rise to a small recovery time. Hence preferring locations on critical cycles is a heuristic for minimizing one's local recovery time.

To prove our results, we study the transients of sequences of the form $A_{i,j}^{\otimes t}$ where either i or j is a critical node. It is clear that the transient of the sequence

$$x_i(t) = \bigoplus_{j=1}^{n} A_{i,j}^{\otimes t} \otimes x_j(0) \tag{3.1}$$

is upper-bounded by the maximum of the transients of the $A_{i,j}^{\otimes t}$ with $j \in [n]$. Further, if i is critical, then we have an upper bound on the transients of $A_{i,j}^{\otimes t}$, and hence of $x_i(t)$.

Interestingly, we are able to give exact generalizations of classical transience bounds for Boolean matrices. In fact, we generalize *all* bounds for Boolean matrices recalled in Section 2.3.3. Such generalizations do not exist in the literature to date. Being exact generalizations, the generalized bounds are at least as tight as the original Boolean bounds.

We now present the bounds that we will prove in the rest of the chapter. Denote the transient of the sequence $\left(A_{i,j}^{\otimes t}\right)_t$ by $T_{i,j}(A)$.

In the Boolean case, we showed in Theorem 2.22 that Kim's bound is always lower than Wielandt's, Dulmage and Mendelsohn's, and Schwarz's. For the generalizations we present here, this ordering is not true. In fact, the generalization of Kim's bound is, in general, not lower than any of the other generalizations.

Theorem 3.1 (Weighted Kim)**.** *Let A be an irreducible $n \times n$ max-plus matrix and let $i, j \in [n]$. Let H be a critical component of $G(A)$. If either i or j is in H, then $T_{i,j}(A) \leq g(H) \cdot \left(\lfloor n/\gamma \rfloor - 2\right) + n$ where γ is the cyclicity of $G(A)$.*

We can see that the original bound by Kim, Theorem 2.12, is an immediate consequence of Theorem 3.1. It suffices to note that all edges are critical in the Boolean case and hence there is a single critical component H, which is equal to the whole matrix's digraph. Because the bound is decreasing in γ, the following weighted analog of Dulmage and Mendelsohn's bound is often larger than the bound of Theorem 3.1. However, it can be strictly lower if $\gamma = 1$ and $|H| < n$. In our proof, we first prove Theorem 3.2 and use it to prove Theorem 3.1. The original Boolean bound by Dulmage and Medelsohn, Theorem 2.10, is an immediate consequence of Theorem 3.2.

Theorem 3.2 (Weighted Dulmage-Mendelsohn)**.** *Let A be an irreducible $n \times n$ max-plus matrix and let $i, j \in [n]$. Let H be a critical component of $G(A)$. If either i or j is in H, then $T_{i,j}(A) \leq g(H) \cdot (n-2) + |H|$.*

In the Boolean case, the bound of Wielandt is a consequence of Dulmage and Mendelsohn's bound. For the weighted generalizations, this ordering is also not true. The reason for the ordering in the Boolean case is that a digraph with girth $g = n$ consists of a single Hamiltonian cycle and hence the corresponding matrix's transient is zero. In the remaining cases, i.e., $g \leq n-1$, the bound of Dulmage and Mendelsohn is at most $(n-1) \cdot (n-2) + n = (n-1)^2 + 1 = \mathrm{Wi}(n)$ if $n \geq 2$. Because our weighted generalizations have the critical component's girth instead of the whole digraphs girth, this argument is no longer valid. In fact, a max-plus matrix's digraph can very well contain a critical Hamiltonian cycle without the matrix's transient being zero.

As in the Boolean case, the weighted generalization of Schwarz's bound is lower than that of Wielandt's bound. However, because we first prove the generalization of Wielandt's bound and use it to prove the generalization of Schwarz's bound, we give them both. The Boolean bounds of Schwarz and Wielandt are again easy corollaries.

Theorem 3.3 (Weighted Schwarz)**.** *Let A be an irreducible $n \times n$ max-plus matrix and let $i, j \in [n]$. If either i or j is critical, then $T_{i,j}(A) \leq \gamma \cdot \mathrm{Wi}(\lfloor n/\gamma \rfloor) + (n \bmod \gamma)$ where γ is the cyclicity of $G(A)$.*

Theorem 3.4 (Weighted Wielandt)**.** *Let A be an irreducible $n \times n$ max-plus matrix and let $i, j \in [n]$. If either i or j is critical, then $T_{i,j}(A) \leq \mathrm{Wi}(n)$.*

We also generalize the bounds involving the factor rank by Gregory, Kirkland, and Pullman (Theorem 2.13) and Kim (Theorem 2.14). These generalizations include the digraph's cyclicity, in contrast to the original bounds. They both have the property that they are maximized when $\gamma = 1$. Among themselves, they are not comparable in general.

Theorem 3.5. *Let A be an irreducible $n \times n$ max-plus matrix and let $i, j \in [n]$. Let H be a critical component of $G(A)$. If either i or j is in H, then*

$$T_{i,j}(A) \leq \begin{cases} \gamma \cdot \mathrm{Wi}\left(\lfloor r/\gamma \rfloor\right) + (r \bmod \gamma) + 1 \\ g(H) \cdot \left(\lfloor r/\gamma \rfloor - 2\right) + r + 1 \end{cases} \tag{3.2}$$

where $r = \mathrm{rk}(A)$ is the factor rank of A and γ the cyclicity of $G(A)$.

Specialized to Boolean matrices, we not only recover Theorems 2.13 and 2.14, but also deduce new bounds for Boolean transients. They are factor rank analogues of the bounds of Schwarz and Kim:

Corollary 3.6. *The transient of an irreducible Boolean matrix with factor rank r is less or equal to both*

$$\gamma \cdot \mathrm{Wi}\left(\lfloor r/\gamma \rfloor\right) + (r \bmod \gamma) + 1 \tag{3.3}$$

and

$$g \cdot \left(\lfloor r/\gamma \rfloor - 2\right) + r + 1 \tag{3.4}$$

where γ is the cyclicity and g is the girth of $G(A)$.

3.1 Proof of Weighted Dulmage-Mendelsohn Transience Bound

We begin the proof of Theorem 3.2 by recalling an elementary result of Nachtigall on the transient of power entries if one indices is a node that lies on a critical cycle in terms of the cycle's length.

Lemma 3.7 (Nachtigall [72, Lemma 3.2])**.** *Let A be an $n \times n$ max-plus matrix and let $i, j \in [n]$. Let C be a critical cycle of $G(A)$. If either i or j is a node of C, then $T_{i,j}(A) \leq \ell(C) \cdot (n-1)$.*

To facilitate the proof, we introduce the notion of a visualized matrix and argue that we can assume our matrices to be visualized without loss of generality. We will reuse this notion in subsequent proofs. A max-plus matrix A is *visualized* if it fulfills one of the following equivalent conditions:

1. *Cycle cover:* For every edge (i, j) in $G(A)$, there exists a cycle Z in $G(A)$ containing (i, j) such that the A-weight of every edge of Z is greater or equal to that of (i, j).
2. *Max balancing:* For every set $M \subseteq [n]$, we have $\max_{i \in M, j \notin M} A_{i,j} = \max_{i \notin M, j \in M} A_{i,j}$.

Every visualized matrix with $\lambda(A) \neq -\infty$ has the property that the maximum A-weight is equal to $\lambda(A)$ and that an edge (i, j) is in the critical digraph if and only if $A_{i,j} = \lambda(A)$, as can be seen with the cycle cover condition.

Schneider and Schneider [78] showed how to transform any irreducible max-plus matrix into a visualized one. For that, they used a scaling relative to some potential, i.e., some real vector $v \in \mathbb{R}^n$. The scaling of A relative to v is defined as

$$B = \mathrm{diag}(v)^{\otimes(-1)} \cdot A \cdot \mathrm{diag}(v) = \left(A_{i,j} - v_i + v_j\right)_{i,j} . \tag{3.5}$$

In this case, the digraphs of A and B are equal and the A-weight of all closed walks is equal to their B-weight. Hence also their critical digraphs are equal. Note, however, that in general $\|B\| \neq \|A\|$.

Theorem 3.8 (Schneider and Schneider [78])**.** *For every irreducible max-plus matrix exists a potential vector whose relative scaling is visualized.*

The following lemma is a main tool for transferring transience bounds for certain critical indices to others in the same critical component. We will use it throughout the chapter.

Lemma 3.9. *Let A be an $n \times n$ max-plus matrix and let $i_1, i_2, j \in [n]$. If there exists a walk of length d from i_1 to i_2 in the critical digraph $G_c(A)$, then $T_{i_1,j}(A) \leq T_{i_2,j}(A) + d$.*

Proof. As neither the transients nor the critical digraph changes when passing to the normalization of A and its visualization, we assume without loss of generality that A is normalized and visualized.

The lemma's hypothesis is then equivalent to $A^{\otimes d}_{i_1,i_2} = 0$. Since $G_c(A)$ is completely reducible without isolated nodes, there exists a nonempty walk in $G_c(A)$ from i_2 to i_1; denote its length by e. It is $A^{\otimes e}_{i_2,i_1} = 0$. For all $t \geq 0$, we have

$$A^{\otimes t+d+e}_{i_2,j} \geq A^{\otimes e}_{i_2,i_1} + A^{\otimes t+d}_{i_1,j} = A^{\otimes t+d}_{i_1,j} \geq A^{\otimes d}_{i_1,i_2} + A^{\otimes t}_{i_2,j} = A^{\otimes t}_{i_2,j} \ . \tag{3.6}$$

Thus $A^{\otimes t+d+e}_{i_2,j} = A^{\otimes t}_{i_2,j}$ for all $t \geq T_{i_2,j}(A)$. There is hence equality in (3.6) for all $t \geq T_{i_2,j}(A)$. In particular,

$$\forall t \geq T_{i_2,j}(A) : \ A^{\otimes t+d}_{i_1,j} = A^{\otimes t}_{i_2,j} \ , \tag{3.7}$$

from which the lemma follows. □

Proof of Theorem 3.2. We prove the theorem only for the case that i is in H. The case that j is in H follows from the first case by passing to the transpose of A.

Denote by C a cycle in H of minimal length, i.e., $\ell(C) = g(H)$. By Lemma 3.7,

$$T_{k,j}(A) \leq g(H) \cdot (n-1) \tag{3.8}$$

for all nodes k of C.

Because i is a node of H and there are at most $|H| - g(H)$ nodes in H not on C, there exists a path in H of length at most $|H| - g(H)$ from i to some node k of C. Lemma 3.9 hence implies

$$T_{i,j}(A) \leq T_{k,j}(A) + |H| - g(H) \ . \tag{3.9}$$

Combination of (3.9) and (3.8) concludes the proof. □

3.2 Proof of Weighted Wielandt Transience Bound

We complete the proof in Section 3.2.2, but we first need some technical machinery to formalize walk reductions and what it means to remove subcycles from a walk. We chose the multigraph approach to do this. As this is not standard, we develop it in Section 3.2.1.

3.2.1 Walk-Multigraph Correspondence

To prove Theorem 3.4, we need to speak about removing and adding cycles to a given walk. For this, we choose the formalism of the multigraph corresponding to a walk, which we develop in this subsection. We will use this formalism also in the next chapter. A multigraph has a multiset of edges instead of a set. That is, it can contain the same edge more than once.

Definition 3.10 (Almost even multigraphs, induced multigraphs)**.** Let $M = (V, E)$ be a multigraph. We call M *even* if every node's in-degree is equal to its out-degree, i.e., $d_M^-(k) = d_M^+(k)$ for all $k \in V$. Let $i, j \in V$. We say that M is (i, j)*-almost even* if either $i = j$ and M is even, or

- for every node $k \in V \setminus \{i, j\}$, the in-degree of k is equal to its out-degree, i.e., $d_M^-(k) = d_M^+(k)$,
- the in-degree of i is one less than its out-degree, i.e., $d_M^-(i) = d_M^+(i) - 1$, and
- the in-degree of j is one more than its out-degree, i.e., $d_M^-(j) = d_M^+(j) + 1$.

We call M *almost even* if there exists $i, j \in V$ such that M is (i, j)-almost even.

If W is a walk in some digraph, then define its *induced multigraph* $M(W) = (V, E)$ by setting V to be the set of nodes of W and E the multiset of edges in W, counted with multiplicity. If $\mathcal{W}$ is a multiset of walks in digraphs, then define its induced multigraph $M(\mathcal{W})$ as the union multigraph $\bigcup_{W \in \mathcal{W}} M(W)$.

Lemma 3.11. *A multigraph is even without isolated nodes if and only if it is induced by a nonempty multiset of cycles.*

Proof. Let $\mathcal{C}$ be a nonempty multiset of cycles and let $M(\mathcal{C}) = (V, E)$ be the multigraph induced by $\mathcal{C}$. If $|\mathcal{C}| = 1$, i.e., $\mathcal{C} = \{C\}$ for some cycle C, then $d_M^-(k) = d_M^+(k) = 1$ for all nodes $k \in V$. Because the union of even multigraphs is even and the union of multigraphs without isolated nodes does not have isolated nodes, we have proved the first part of the lemma.

Now let $M = (V, E)$ be an even multigraph without isolated nodes. We prove by induction on $|E|$ that M is induced by a multiset of cycles. If $|E| = 1$, then V contains a single node i and the sole element of E is the self-loop (i, i). Hence M is induced by a cycle of length 1. Let now $|E| > 1$. Because there are no isolated nodes in M, every node in M has either an incoming or an outgoing edge. But because M is also even, every node in M has at least one outgoing edge. For every $k \in V$, let $f(k)$ be an outgoing neighbor of k, i.e., let $f : V \to V$ such that $(k, f(k)) \in E$ for all $k \in V$. Now pick any node $k \in V$. As V is finite, there exist positive integers $s < t$ such that $f^s(k) = f^t(k)$. Choose s and t such that $t - s$ is minimal. The walk C with edges $(f^r(k), f^{r+1}(k))$ where $s \leq r < t$ is a cycle because $t - s$ is minimal. The multigraph $M(C)$ induced by C is an even sub-multigraph of M. If $M(C) = M$, then we are done. If not, then the multigraph $M' = (V', E')$ where

$$\begin{aligned} E' &= E \setminus \{(f^r(k), f^{r+1}(k)) \mid s \leq r < t\} \\ V' &= \{i \in V \mid d^+_{(V,E')}(i) > 0\} \end{aligned} \tag{3.10}$$

is even, has no isolated nodes, and satisfies $M = M' \cup M(C)$. Application of the induction hypothesis then concludes the proof. □

Lemma 3.12. *Let W be a nonempty walk in a digraph. Denote by i the start node and by j the end node of W. Then $M(W)$ is (i, j)-almost even, has no isolated nodes, and is weakly connected.*

Proof. Let $i_0, i_1, \ldots, i_\ell$ be the sequence of nodes of W. We have, for all nodes k of $M(W)$,

$$\begin{aligned} d^+_{M(W)}(k) &= |\{0 \le m < \ell \mid i_m = k\}| = |\{0 \le m \le \ell \mid i_m = k\}| - \delta(j,k) \\ d^-_{M(W)}(k) &= |\{0 < m \le \ell \mid i_m = k\}| = |\{0 \le m \le \ell \mid i_m = k\}| - \delta(i,k) \end{aligned} \tag{3.11}$$

where $\delta(a,b)$ denotes the Kronecker delta, i.e., $\delta(a,b) = 1$ if $a = b$ and $\delta(a,b) = 0$ otherwise. We distinguish the two cases $i = j$ and $i \neq j$.

If $i = j$, then $\delta(i,k) = \delta(j,k)$ for all k and hence $d^+_{M(W)}(k) = d^-_{M(W)}(k)$ for all nodes k of $M(W)$ by Equation (3.11), i.e., $M(W)$ is even.

If $i \neq j$, then we have $d^+_{M(W)}(k) = d^-_{M(W)}(k)$ only for all nodes $k \notin \{i,j\}$. For $k = i$, we get

$$d^+_{M(W)}(i) = d^+_{M(W)}(i) + \delta(j,i) = d^-_{M(W)}(i) + \delta(i,i) = d^-_{M(W)}(i) + 1 \tag{3.12}$$

from Equation (3.11), and similarly for $k = j$, we get $d^-_{M(W)}(j) = d^+_{M(W)}(j) + 1$. Hence $M(W)$ is (i,j)-almost even in both cases. □

Lemma 3.13. *Let $M = (V,E)$ be a multigraph and let $i, j \in V$. If M is (i,j)-almost even, has no isolated nodes, and is weakly connected, then there exists a walk W from i to j in M's underlying digraph such that $M = M(W)$.*

Moreover, there exists a path P from i to j and a multiset $\mathcal{C}$ of cycle in M's underlying digraph such that M is induced by $\{P\} \cup \mathcal{C}$.

Proof. We first prove the case $i = j$, in which P can be chosen to be empty. By Lemma 3.11, there exists a multiset $\mathcal{C}$ of cycles such that $M = M(\mathcal{C})$. Assume that $\mathcal{C}$ has maximum cardinality among all multisets of cycles that induce M. We show by induction on $|\mathcal{C}|$ that there exists a single closed walk W such that $M = M(W)$. If $|\mathcal{C}| = 1$, the claim is trivial.

So let $C \in \mathcal{C}$ such that $\mathcal{C}' = \mathcal{C} \setminus \{C\}$ is nonempty. Let $M_1, M_2, \ldots, M_s$ be the weakly connected components of $M(\mathcal{C}')$. Because all of them are even, there exist multisets $\mathcal{C}'_1, \mathcal{C}'_2, \ldots, \mathcal{C}'_s$ of cycles with maximum cardinality such that $M_r = M(\mathcal{C}'_r)$ for all $1 \le r \le s$. Because M is induced by the multiset $\{C\} \cup \bigcup_{r=1}^{s} \mathcal{C}'_r$ and $\mathcal{C}$ was assumed to have maximum cardinality, we have

$$|\mathcal{C}| \ge 1 + \sum_{r=1}^{s} |\mathcal{C}'_r| \tag{3.13}$$

and hence $|\mathcal{C}'_r| \le |\mathcal{C}| - 1$ for all $1 \le r \le s$. Because the M_r are weakly connected by definition, we can thus apply the induction hypothesis to them and get the existence of closed walks $W_1, W_2, \ldots, W_s$ such that $M_r = M(W_r)$ for all $1 \le r \le s$. The cycle C shares a node with every W_r because otherwise M would not be weakly connected. Write $C = V_0 \cdot V_1 \cdots V_s$ such that the start node i_r of V_r shares a node with W_r for all $1 \le r \le s$. Because the W_r are closed, we can assume without loss of generality that i_r is the start (and end) node of W_r. But then $W = V_0 \cdot W_1 \cdot V_1 \cdots W_r \cdot V_r$ is a closed walk that induces M.

It remains to prove the case $i \neq j$. In this case, there exists a path P in M from i to j: Let N_ℓ be the set of nodes in M of distance at most ℓ from i, i.e.,

$$N_\ell = \left\{k \mid \bigcup_{m=0}^{\ell} \mathcal{W}^m(i \to k) \neq \emptyset\right\} . \tag{3.14}$$

Because of the degree conditions imposed by the (i,j)-almost evenness, $N_{\ell+1} \supsetneq N_\ell$ if $j \notin N_\ell$. There hence exists a path P from i to j in M. Denote by M' the resulting multigraph after

removing the edges of P from M. The nontrivial weakly connected components of M' are all even and we can hence use the lemma in the already proved case of even weakly connected multigraphs. □

Definition 3.14 (Realization, path-cycles decomposition)**.** Let $M = (V, E)$ be an almost even weakly connected multigraph without isolated nodes. Let $i, j \in V$ such that M is (i, j)-almost even. A walk W in M's underlying digraph is called an *(i, j)-realization* of M if $M = M(W)$.

Let W be a walk in a digraph G. A *path-cycles decomposition* of W is a pair $(P, \mathcal{C})$ where P is a path in G and $\mathcal{C}$ is a multiset of cycles in G such that $M(W) = M(\{P\} \cup \mathcal{C})$.

3.2.2 Critical Hamiltonian Cycles

We now prove Theorem 3.4.

Lemma 3.15. *Let d be a positive integer and let $x_1, x_2, \ldots, x_d$ be integers. Then there exists a nonempty subset I of $[d]$ such that*

$$\sum_{i \in I} x_i \equiv 0 \pmod{d} \ . \tag{3.15}$$

Proposition 3.16. *Let G be a digraph with $|G| = n$ that contains a Hamiltonian cycle C_H. Then, for every walk W in G, there exists a walk V obtained from W by removing cycles and adding copies of C_H whose length satisfies*

$$\ell(V) - \mathrm{Wi}(n) \in \{0, 1, \ldots, n-1\} \quad \textit{and} \quad \ell(V) \equiv \ell(W) \pmod{n} \ . \tag{3.16}$$

Proof. Let $(P, \mathcal{C})$ be a path-cycles decomposition of W. Let $\mathcal{C}'$ be a minimal sub-multiset of $\mathcal{C}$ such that

$$\sum_{C \in \mathcal{C}'} \ell(C) \equiv \sum_{C \in \mathcal{C}} \ell(C) \pmod{n} \ . \tag{3.17}$$

By Lemma 3.15, we have $|\mathcal{C}'| \leq n - 1$. Also, $\ell(C) \leq n - 1$ for all $C \in \mathcal{C}'$.

Denote by i the start node and by j the end node of P, i.e., also of W. Let M be the multigraph induced by $\{P\} \cup \mathcal{C}'$. The multigraph M is (i, j)-almost even. We distinguish two cases:

1. M is weakly connected.

2. M is not weakly connected.

In Case 1, we choose W' to be any (i, j)-realization of M. Its length satisfies

$$\begin{aligned} \ell(W') &= \ell(P) + \sum_{C \in \mathcal{C}'} \ell(C) \\ &\leq (n-1) + (n-1) \cdot (n-1) \\ &< \mathrm{Wi}(n) + n - 1 \ . \end{aligned} \tag{3.18}$$

In Case 2, there in particular exists a cycle $\hat{C} \in \mathcal{C}'$ that is node-disjoint to P, i.e., $(\ell(P) + 1) + \ell(\hat{C}) \leq n$, which is equivalent to $\ell(P) + \ell(\hat{C}) \leq n - 1$. Let M' be the multigraph induced

by $\{P\} \cup \mathcal{C}' \cup \{C_H\}$. As C_H is Hamiltonian, M' is connected. Because C_H is a closed walk, M' is (i,j)-almost even. Choose W' to be any (i,j)-realization of M'. Its length satisfies

$$\begin{aligned} \ell\left(W'\right) &= \ell\left(P\right) + \ell(\hat{C}) + \sum_{C \in \mathcal{C}' \setminus \{\hat{C}\}} \ell\left(C\right) + \ell\left(C_H\right) \\ &\leq (n-1) + (n-2) \cdot (n-1) + n \\ &= \mathrm{Wi}(n) + n - 1 \ . \end{aligned} \tag{3.19}$$

In both cases, W' is a walk obtained from W by removing cycles and adding copies of C_H that satisfies $\ell\left(W'\right) - \mathrm{Wi}(n) \leq n-1$. If $\ell\left(W'\right) - \mathrm{Wi}(n) \geq 0$, then set $V = W'$. If not, let V be obtained from W' by adding $\lceil (\mathrm{Wi}(n) - \ell\left(W'\right))/n \rceil$ copies of C_H. We thus arrive at a walk V obtained from W by removing cycles and adding copies of C_H whose length satisfies $0 \leq \ell\left(V\right) - \mathrm{Wi}(n) \leq n-1$ and $\ell\left(V\right) \equiv \ell\left(W\right) \pmod{n}$. □

Lemma 3.17. *Let A be an $n \times n$ max-plus matrix. Let $i \in [n]$ and $T \geq 0$. If there exists a $p > 0$ such that $A_{i,j}^{\otimes T} = A_{i,j}^{\otimes T+p}$ for all $j \in [n]$, then $T_{i,j}\left(A\right) \leq T$ for all $j \in [n]$.*

Proof. For all $j \in [n]$ and all $t \geq T$, we have

$$A_{i,j}^{\otimes t+p} = \max_{k \in [n]} A_{i,k}^{\otimes T+p} + A_{k,j}^{\otimes t-T} = \max_{k \in [n]} A_{i,k}^{\otimes T} + A_{k,j}^{\otimes t-T} = A_{i,j}^{\otimes t} \tag{3.20}$$

by definition of max-plus matrix multiplication and the lemma's hypothesis. This concludes the proof. □

Proof of Theorem 3.4. We prove the theorem only for the case that i is critical. The case that j is critical follows from the first case by passing to the transpose of A. Also, because neither the transients nor the critical digraph changes when passing to the normalization of A and its visualization, we assume without loss of generality that A is normalized and visualized.

If $g(H) \leq n-1$, then the theorem follows from Theorem 3.2. We hence assume $g(H) = n$, that is, there exists a critical Hamiltonian cycle C_H in $G(A)$.

Let W be a walk of maximum weight amongst all walks from i to j whose length is in the set $\mathrm{Wi}(n) + n \cdot \mathbb{N}_0$. Such a walk exists if A is normalized. By Proposition 3.16, there exists a walk V obtained from W by removing cycles and adding copies of C_H such that $\ell\left(V\right) = \mathrm{Wi}(n)$. We have $A(V) \geq A(W)$. But by the choice of W, $A(V) \leq A(W)$, that is, $A(V) = A(W)$.

Let V' be obtained from V by adding a copy of C_H. The length of V' is $\mathrm{Wi}(n) + n$ and the weight of V' is the same as that of V, i.e., $A(V') = A(V) = A(W)$. Because, in particular, V has maximum weight amongst all walks from i to j of length $\mathrm{Wi}(n)$ and V' amongst all of length $\mathrm{Wi}(n) + n$, we have

$$A_{i,j}^{\otimes \mathrm{Wi}(n)} = A(V) = A(V') = A_{i,j}^{\otimes \mathrm{Wi}(n)+n} \tag{3.21}$$

Application of Lemma 3.17 now concludes the proof. □

3.3 Proof of Weighted Schwarz Transience Bound

Call a cyclicity class of a digraph G *small* if its cardinality is minimal.

Lemma 3.18. *Let G be a strongly connected digraph with n nodes. Denote by m the cardinality of small cyclicity classes of G and by γ the cyclicity of G. Then $m \leq \lfloor n/\gamma \rfloor$.*

Moreover, if $m = \lfloor n/\gamma \rfloor$, then the number of small cyclicity classes of G is at least $\gamma - (n \bmod \gamma)$.

Proof. Let $\Gamma_1, \Gamma_2, \ldots, \Gamma_\gamma$ be the cyclicity classes of G and let $\Gamma_1, \Gamma_2, \ldots, \Gamma_p$ be the small classes. Then

$$n = \sum_{q=1}^{\gamma} |\Gamma_l| = p \cdot m + \sum_{q=p+1}^{\gamma} |\Gamma_l| \geq p \cdot m + (\gamma - p) \cdot (m+1) = \gamma \cdot m + \gamma - p \ . \tag{3.22}$$

Because $p \leq \gamma$, (3.22) implies $n \geq \gamma \cdot m$, i.e., $m \leq n/\gamma$, which implies $m \leq \lfloor n/\gamma \rfloor$. This proves the first part of the lemma.

If $m = \lfloor n/\gamma \rfloor$, then (3.22) implies $p \geq \gamma - \left(n - \gamma \cdot \lfloor n/\gamma \rfloor\right) = \gamma - (n \bmod \gamma)$. This proves the second part of the lemma. □

Proof of Theorem 3.3. We prove the theorem only for the case that i is critical. The case that j is critical follows from the first case by passing to the transpose of A. Also, because neither the transients nor the critical digraph changes when passing to the normalization of A and its visualization, we assume without loss of generality that A is normalized and visualized.

Let p be the number and m the cardinality of small cyclicity classes of $G(A)$. Let k be a critical node in a small cyclicity class of $G(A)$. Set $B = A^{\otimes\gamma}$. Then k is a critical node in a strongly connected component of size m in $G(B)$. By Theorem 3.4, we have $T_{k,j}(B) \leq \mathrm{Wi}(m)$. This implies $T_{k,j}(A) \leq \gamma \cdot \mathrm{Wi}(m)$ by Lemma 3.17. There exists a path in $G_c(A)$ of length at most $\gamma - p$ from i to some node k in a small class of $G(A)$. By Lemma 3.9, we hence have

$$T_{i,j}(A) \leq T_{k,j}(A) + \gamma - p \leq \gamma \cdot \mathrm{Wi}(m) + \gamma - p \ . \tag{3.23}$$

We distinguish two cases:

1. $m \leq \lfloor n/\gamma \rfloor - 1$
2. $m = \lfloor n/\gamma \rfloor$

These are all possible cases by the first part of Lemma 3.18.

In Case 1, we have

$$\mathrm{Wi}(m) \leq \mathrm{Wi}\left(\lfloor n/\gamma \rfloor - 1\right) \leq \mathrm{Wi}(\lfloor n/\gamma \rfloor) - 1 \ . \tag{3.24}$$

Hence combination of (3.23) and (3.24) implies

$$T_{i,j}(A) < \gamma \cdot \mathrm{Wi}(\lfloor n/\gamma \rfloor) \leq \gamma \cdot \mathrm{Wi}(\lfloor n/\gamma \rfloor) + (n \bmod \gamma) \ . \tag{3.25}$$

In Case 2, we have $p \geq \gamma - (n \bmod \gamma)$ by the second part of Lemma 3.18. Hence (3.23) implies $T_{i,j}(A) \leq \gamma \cdot \mathrm{Wi}(\lfloor n/\gamma \rfloor) + (n \bmod \gamma)$. □

3.4 Proof of Weighted Kim Transience Bound

We prove the theorem only for the case that i is in H. The case that j is in H follows from the first case by passing to the transpose of A. Also, because neither the transients nor the critical digraph changes when passing to the normalization of A and its visualization, we assume without loss of generality that A is normalized and visualized.

Let p be the number and m the cardinality of small cyclicity classes of $G(A)$. Let k be a critical node in a small cyclicity class of $G(A)$. Set $B = A^{\otimes\gamma}$. Then k is a critical node in a strongly connected component of size m in $G(B)$. By Theorem 3.2, we have $T_{k,j}(B) \leq g(H)/\gamma \cdot (m-2) + m$. This implies

$$T_{k,j}(A) \leq g(H) \cdot (m-2) + \gamma \cdot m \tag{3.26}$$

by Lemma 3.17. There exists a path in H of length at most $\gamma - p$ from i to some node k in a small class of $G(A)$. By Lemma 3.9, we hence have

$$T_{i,j}(A) \leq g(H) \cdot (m-2) + \gamma \cdot m + \gamma - p \ . \tag{3.27}$$

We distinguish two cases:

1. $m \leq \lfloor n/\gamma \rfloor - 1$
2. $m = \lfloor n/\gamma \rfloor$

These are all possible cases by the first part of Lemma 3.18.

In Case 1, we have

$$g(H) \cdot (m-2) \leq g(H) \cdot \big(\lfloor n/\gamma \rfloor - 2\big) - g(H) \leq g(H) \cdot \big(\lfloor n/\gamma \rfloor - 2\big) - \gamma \ . \tag{3.28}$$

Combination of (3.27) and (3.28) implies

$$\begin{aligned} T_{i,j}(A) &\leq g(H) \cdot \big(\lfloor n/\gamma \rfloor - 2\big) + \gamma \cdot m - p \\ &< g(H) \cdot \big(\lfloor n/\gamma \rfloor - 2\big) + n \end{aligned} \tag{3.29}$$

because $\gamma \cdot m \leq n$.

In Case 2, we have $p \geq \gamma - (n \bmod \gamma)$ by the second part of Lemma 3.18. Hence (3.27) implies

$$\begin{aligned} T_{i,j}(A) &\leq g(H) \cdot \big(\lfloor n/\gamma \rfloor - 2\big) + \gamma \cdot \lfloor n/\gamma \rfloor + (n \bmod \gamma) \\ &= g(H) \cdot \big(\lfloor n/\gamma \rfloor - 2\big) + n \ . \end{aligned} \tag{3.30}$$

This concludes the proof.

3.5 Proof of Weighted Transience Bounds Involving the Factor Rank

Define the $r \times r$ matrix B by setting $b_{\alpha,\beta} = \bigoplus_{i=1}^{n} w_{\alpha,i} \cdot v_{\beta,i}$. We will apply the bounds of Theorems 3.4, 3.2, 3.3, and 3.1 to the critical nodes of B and transfer the result to the critical nodes of A.

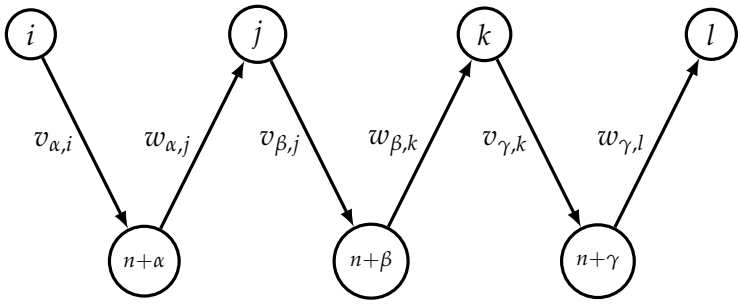


Figure 3.1: A walk in $G(Z)$

Form the following $(n+r)\times(n+r)$ matrix Z: For $i\in[n]$ and $\alpha\in[r]$, set $z_{i,n+\alpha}=v_{\alpha,i}$ and $z_{n+\alpha,i}=w_{\alpha,i}$. All other entries of Z are 0. Figure 3.1 depicts an example of a walk in $G(A)$.

The entries of $Z^{\otimes 2}$ satisfy:

$$\begin{aligned} z^{(2)}_{i,j} &= a_{i,j} && \text{for all } i,j\in[n]\\ z^{(2)}_{n+\alpha,n+\beta} &= b_{\alpha,\beta} && \text{for all } \alpha,\beta\in[r]\\ z^{(2)}_{i,n+\beta} &= 0 && \text{for all } i\in[n] \text{ and } \beta\in[r]\\ z^{(2)}_{n+\alpha,j} &= 0 && \text{for all } \alpha\in[r] \text{ and } j\in[n] \end{aligned} \tag{3.31}$$

Matrix Z is irreducible: Because A is irreducible, by the first equality in (3.31), there exists a walk in $G(Z)$ between every pair of nodes in $[n]$. None of the vectors v_α, w_α is zero by the minimality of r, i.e., every node in $\{n+1,\dots,n+r\}$ has an incoming and an outgoing neighbor in $\{1,\dots,n\}$. Hence there exists a walk between every pair of nodes in $G(Z)$.

Every walk in $G(Z)$ alternates between nodes in $\{1,\dots,n\}$ and nodes in $\{n+1,\dots,n+r\}$. In particular, all walks in $G(Z)$ between two nodes in $\{1,\dots,n\}$ or two nodes in $\{n+1,\dots,n+r\}$ have even length. This implies that also B is irreducible by the second equality in (3.31).

Every nonempty closed walk in $G(Z)$ of length ℓ surjectively corresponds to both a closed walk in $G(A)$ of length $\ell/2$ and a closed walk in $G(B)$ of length $\ell/2$. See Figure 3.2 for an example of this correspondence. In particular, the cyclicities of $G(A)$ and $G(B)$ are equal, i.e., also $G(B)$ has cyclicity d. The correspondence also maps critical closed walks to critical closed walks. That is, k is also a critical node in $G(Z)$ and it has both a critical incoming neighbor $n+\alpha$ and a critical outgoing neighbor $n+\beta$ in $G_c(Z)$. Furthermore, both α and β are critical in $G(B)$ and they are contained in the same component H' of $G_c(B)$. The correspondence also yields the equality $g(H')=g(H)$ of the girths.

By Theorems 3.3 and 3.1, we have $T_\beta(B), T^\alpha(B)\le T$ where $T+1$ is the minimum of the right-hand sides of (3.2). The second equality in (3.31) implies the two inequalities $T_{n+\beta}(Z)\le 2T$ and $T^{n+\alpha}(Z)\le 2T$, which implies $T_k(Z), T^k(Z)\le 2T+1$ by Lemma 3.9. The first equality in (3.31) now implies $T_k(A), T^k(A)\le\lceil(2T+1)/2\rceil=T+1$, which concludes the proof.

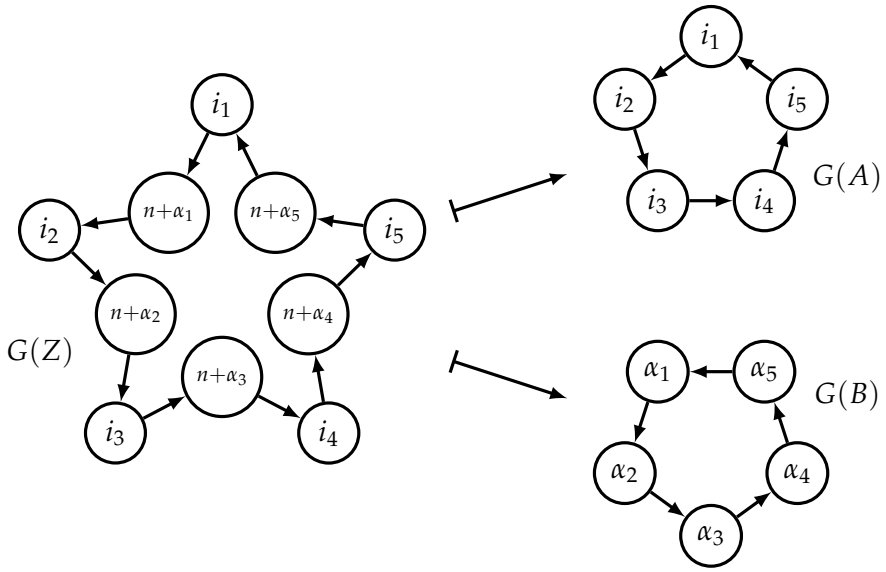


Figure 3.2: Correspondence between closed walks in $G(Z)$, $G(A)$, and $G(B)$

Chapter 4

Global Transience Bounds

While the previous chapter dealt with the entry-wise transients of critical nodes only, the present chapter deals with the global transients of matrices and systems. That is, the maximum of all entry-wise transients. The chapter is logically independent from Chapter 3, but it reuses and further develops some of the techniques used, such as the cycle decomposition based on the multigraph approach. The bounds presented in this chapter improve all published bounds on the transients of matrices and systems. However, not all of them are exact generalizations of known bounds for Boolean matrices, as was the case of the bounds for critical nodes in Chapter 3. Nonetheless, one of our bounds is a generalization of Wielandt's bound and one is a generalization of Dulmage and Mendelsohn's bound. These two generalizations, while generalizing the Boolean bounds, do not generalize their weighted analogues from Chapter 3.

The chapter is organized as follows: Section 4.1 states our transience bounds that we will prove in this chapter. Section 4.2 applies our transience bounds to the applications we introduced in Section 2.2. In Sections 4.3–4.7, we prove our transience bounds for max-plus matrices. We modify our proof strategy to deduce transience bounds for max-plus linear systems in Section 4.8. Section 4.9 discusses the relation between matrix and system transients. Finally, in Section 4.10, we present a different, more algebraic and arguably shorter and simpler, approach for proving transience bounds that uses the Nachtigall decomposition.

4.1 Results

We prove transience bounds for both matrices and systems. The matrix bounds being upper bounds also on the system transients and the system transients being more important in applications. All of them are a maximum of two terms: one to ensure connectivity to the critical digraph and one for the necessary combinatorial constructions starting from a critical node. The bounds put forward in Chapter 3 only contained one term because connectivity to the critical digraph was ensured by assumption. We denote the bound to ensure connectivity to the critical digraph the *critical bound*. This critical bound includes the factor $1/\big(\lambda(A) - \lambda_0(A)\big)$ where $\lambda_0(A)$ is the maximum cycle mean of some sub-matrix of A, i.e., $\lambda_0(A) = \lambda(B)$ where B is a matrix derived from A by setting certain entries to $-\infty$. For all such B, we have $\lambda_0(A) \leq \lambda(A)$. We will furthermore make sure that $\lambda_0(A) < \lambda(A)$. Hence, the smaller $\lambda(B)$, the lower the critical bound. If $G(B)$ does not contain any cycle, then $\lambda_0(A) = \lambda(B) = -\infty$ and hence $1/\big(\lambda(A) - \lambda_0(A)\big) = 0$.

We consider two different ways to define B: the Hartmann-Arguelles scheme and the cycle threshold scheme. The λ_0 in the cycle threshold scheme, which we denote by λ_{ct}, is always smaller than the λ_0 in the Hartmann-Arguelles scheme, which we denote by λ_{ha}. This makes the critical bound smaller. However, it necessitates a larger combinatorial term. A simpler choice for B, and therefore λ_0, that was used in the literature is setting all those $A_{i,j}$ to $-\infty$ where either i or j is a critical node. The next two subsections introduce each of the two schemes we consider and present the corresponding transience bounds we deduce.

Both schemes are defined in terms of sub-digraphs of $G(A)$ that is to be removed. These digraphs G_0 will be super-digraph of the critical digraph $G_c(A)$, which guarantees $\lambda_0(A) < \lambda(A)$. Then let B be the matrix defined by

$$B_{i,j} = \begin{cases} A_{i,j} & \text{if neither } i \text{ nor } j \text{ is a node of } G_0 \\ -\infty & \text{else .} \end{cases} \tag{4.1}$$

4.1.1 Hartmann-Arguelles Scheme

This subsection introduces another choice for G_0, used by Hartmann and Arguelles, and shows how to reduce the criticality threshold of G_c to that of this $G_0 = G_{\text{ha}}$. Its idea is to grow the critical digraph as long as one can be sure that we can exchange edges of walks that visit the enlarged digraph for a closed walk in the enlarged digraph that connects to the critical digraph G_c. This is done in the visualization of the max-plus matrix because we can then use the cycle cover property to decide whether or not to keep an edge solely by looking at its weight.

Let V be the visualization of A. Given $\mu \in \mathbb{R}_{\max}$, define the *Hartmann-Arguelles threshold digraph* $T^{\text{ha}}(\mu)$ induced by all edges (i,j) in $G(A) = G(V)$ with $V_{i,j} \geq \mu$. For $\mu = \lambda(A) = \lambda(V)$ we have $T^{\text{ha}}(\mu) = G_c(A) = G_c(V)$. Let λ_{ha} be the maximum of $\mu \leq \lambda(A)$ such that $T^{\text{ha}}(\mu)$ has a strongly connected component that does not contain any strongly connected component of $G_c(A)$. If no such μ exists, then $\lambda_{\text{ha}} = -\infty$ and $T^{\text{ha}}(\lambda_{\text{ha}}) = G(V)$.

The sub-digraph $G_0 = G_{\text{ha}}$ defining B in the Hartmann-Arguelles scheme is the union of the strongly connected component's of $T^{\text{ha}}(\lambda_{\text{ha}})$ intersecting $G_c(A)$. We denote this matrix B by B_{ha}. Observe that $\lambda(B_{\text{ha}}) = \lambda_{\text{ha}}$ and the digraphs $T^{\text{ha}}(\mu)$, for all μ, are completely reducible due to the max-balancing property of V.

The following theorem is our main theorem for bounds on matrix transients using the Hartmann-Arguelles scheme. It contains four bounds: The first two contain the maximum girth of components of the critical digraph, whereas the latter two contain the cyclicity and the index of convergence. For reasons that become apparent in the proof of the bounds, we dub the first two "repetitive" and the latter two "explorative" bounds. The difference between the first and the second bound, as well as between the third and the fourth, is that the second and the fourth use the maximum path length and the maximum cycles length as a parameter. Both these parameters are NP-hard to compute in general, but they can be useful to include for the case that one has a priori bounds on them. Two trivial a priori bounds are $n-1$ for the length of paths and n for the length of cycles if n is the number of nodes in the digraph. Symbolically, we write $\text{cdd}(G)$ for the *cab driver's diameter*, i.e., the longest path length in G, and $\text{cf}(G)$ for the *circumference*, i.e., the longest cycle length in G.

Theorem 4.1. *Let A be an irreducible $n \times n$ max-plus matrix. Then its transient $T(A)$ is bounded by all of the following terms:*

$$n + \max\left\{\hat{g}(n-2)\,,\ \frac{(\hat{\gamma}+1)(n-1)\cdot\|A\|}{\lambda(A)-\lambda_{\mathrm{ha}}(A)}\right\} \tag{4.2}$$

$$\mathrm{cdd} + \max\left\{(\hat{g}-1)(\mathrm{cf}-1)+\hat{g}\cdot\mathrm{cdd}\,,\ \frac{((\hat{\gamma}-1)\,\mathrm{cf}+(\hat{\gamma}+1)\,\mathrm{cdd})\cdot\|A\|}{\lambda(A)-\lambda_{\mathrm{ha}}(A)}\right\} \tag{4.3}$$

$$n + \max\left\{\hat{\gamma}(n-2)+\hat{\mathrm{ind}}\,,\ \frac{(\hat{\gamma}+1)(n-1)\cdot\|A\|}{\lambda(A)-\lambda_{\mathrm{ha}}(A)}\right\} \tag{4.4}$$

$$\mathrm{cdd} + \max\left\{(\hat{\gamma}-1)(\mathrm{cf}-1)+\hat{\gamma}\cdot\mathrm{cdd}+\hat{\mathrm{ind}}\,,\ \frac{((\hat{\gamma}-1)\,\mathrm{cf}+(\hat{\gamma}+1)\,\mathrm{cdd})\cdot\|A\|}{\lambda(A)-\lambda_{\mathrm{ha}}(A)}\right\} \tag{4.5}$$

where $\hat{g}$ is the maximum girth of critical components, $\hat{\gamma}$ is the maximum cyclicity of critical components, $\mathrm{cf} = \mathrm{cf}(G(A))$ is the circumference of $G(A)$, $\mathrm{cdd} = \mathrm{cdd}(G(A))$ is its cab driver's diameter, $\hat{\mathrm{ind}}$ is the maximum index of convergence of critical components, and n_c is the number of critical nodes.

Note that the first bound in theorem, Equation 4.2, is a direct generalization of Dulmage and Mendelsohn's bound for Boolean matrices (Theorem 2.10), for then $\lambda_{\mathrm{ha}} = -\infty$ and the bound reduces to $n + g\cdot(n-2)$ where g is the girth of $G(A)$.

The following corollary is a bound for primitive irreducible matrices. It is reminiscent of the bound of Bouillard and Gaujal [14]. Except for an explicit critical bound, the bound is a strict improvement of their bound.

Corollary 4.2. *Let A be an irreducible $n \times n$ max-plus matrix whose critical digraph has cyclicity 1. Then its transient is bounded by*

$$T(A) \le n + \max\left\{n + n_c - 2 + (n_c - 2h_2)\hat{g}\,,\ \frac{(2n-2)\|A\|}{\lambda(A)-\lambda_{\mathrm{ha}}(A)}\right\} \tag{4.6}$$

where n_c is the number of critical nodes, $\hat{g}$ is the maximum girth of critical components, and h_2 is the number of critical components of size at least 2.

Proof. We use bound (4.4) and estimate $\hat{\mathrm{ind}}$. Let $H_1, H_2, \ldots, H_h$ be the critical components and $H_1, H_2, \ldots, H_{h_2}$ be the components of size at least 2. Denoting by n_k the number of nodes of H_k and by g_k its girth, we have

$$\begin{aligned}\hat{\mathrm{ind}} &= \max_{1\le k\le h}\mathrm{ind}(H_k) \le \sum_{k=1}^{h}\mathrm{ind}(H_k) = \sum_{k=1}^{h_2}\mathrm{ind}(H_k) \le \sum_{k=1}^{h_2}(n_k+(n_k-2)g_k)\\ &\le \sum_{k=1}^{h_2}(n_k+(n_k-2)\hat{g}) = n_c + (n_c-2h_2)\hat{g}\end{aligned} \tag{4.7}$$

by Theorem 2.10. Plugging this estimate into (4.4) and noting $\hat{\gamma} = 1$ concludes the proof. □

The following theorem contains the analogues of the bounds of Theorem 4.1 for max-plus linear systems. Their critical bound depends on the initial vector and can be lower than that in the matrix bounds.

Theorem 4.3. *Let A be an irreducible $n \times n$ max-plus matrix and let $v \in \mathbb{R}^n$. Then its transient $T_v(A)$ is bounded by all of the following terms:*

$$\max\left\{n + \hat{g}(n-2)\ ,\ \frac{\|v\| + \|A\| \cdot (n-1)}{\lambda(A) - \lambda_{\mathrm{ha}}(A)}\right\} \tag{4.8}$$

$$\max\left\{(\hat{g}-1)(\mathrm{cf}-1) + (\hat{g}+1)\cdot \mathrm{cdd}\ ,\ \frac{\|v\| + \|A\| \cdot (n-1)}{\lambda(A) - \lambda_{\mathrm{ha}}(A)}\right\} \tag{4.9}$$

$$\max\left\{n + \hat{\gamma}(n-2) + \hat{\mathrm{ind}}\ ,\ \frac{\|v\| + \|A\| \cdot (n-1)}{\lambda(A) - \lambda_{\mathrm{ha}}(A)}\right\} \tag{4.10}$$

$$\max\left\{(\hat{\gamma}-1)(\mathrm{cf}-1) + (\hat{\gamma}+1)\cdot \mathrm{cdd} + \hat{\mathrm{ind}}\ ,\ \frac{\|v\| + \|A\| \cdot (n-1)}{\lambda(A) - \lambda_{\mathrm{ha}}(A)}\right\} \tag{4.11}$$

where $\hat{g}$ is the maximum girth of critical components, $\hat{\gamma}$ is the maximum cyclicity of critical components, $\mathrm{cf} = \mathrm{cf}(G(A))$ is the circumference of $G(A)$, $\mathrm{cdd} = \mathrm{cdd}(G(A))$ is its cab driver's diameter, $\hat{\mathrm{ind}}$ is the maximum index of convergence of critical components, and n_c is the number of critical nodes.

4.1.2 Cycle Threshold Scheme

This subsection introduces another, completely novel, choice for a super-digraph G_0 of the critical digraph. We will proof that it is always larger than the Hartmann-Arguelles digraph G_{ha}. Its idea is not to pass via edge weights of the visualization of the matrix and then using the cycle cover property, but to directly define G_0 in terms of cycle weights.

For $\mu \in \mathbb{R}_{\max}$, define the *cycle threshold digraph* $T^{\mathrm{ct}}(\mu)$ induced by all nodes and edges belonging to the cycles in $G(A)$ with mean weight greater or equal to μ. Again, for $\mu = \lambda(A)$ we have $T^{\mathrm{ct}}(\mu) = G_c(A)$. Let λ_{ct} be the maximum of $\mu \leq \lambda(A)$ such that $T^{\mathrm{ct}}(\mu)$ has a strongly connected component that does not contain any strongly connected component of $G_c(A)$. If no such μ exists, then $\lambda_{\mathrm{ct}} = -\infty$ and $T^{\mathrm{ct}}(\lambda_{\mathrm{ct}})$ is equal to $G(A)$.

The sub-digraph $G_0 = G_{\mathrm{ct}}$ defining B in the cycle threshold scheme is the union of the strongly connected component of $T^{\mathrm{ct}}(\lambda_{\mathrm{ct}})$ intersecting $G_c(A)$. This matrix B will be denoted by B_{ct}. We again observe that $\lambda(B_{\mathrm{ct}}) = \lambda_{\mathrm{ct}}$.

One can deduce transience bounds with the cycle threshold scheme and thus λ_{ct} similarly as in Theorem 4.1. The first terms in the maximum of these bounds will be strictly larger than that in Theorem 4.1. We will show a particular case of this requiring a little more work, but showing that the term can be chosen to be the Wielandt number $\mathrm{Wi}(n)$. Even in the case of the Hartmann-Arguelles scheme, this is a nontrivial extension of Theorem 4.1. A particular consequence is that $T(A) \leq \mathrm{Wi}(n)$ if $\lambda_{\mathrm{ct}}(A) = -\infty$. By that, it is a strict generalization of Wielandt's bound (Theorem 2.9).

Theorem 4.4. *Let A be an irreducible $n \times n$ max-plus matrix. Then its transient is bounded by*

$$T(A) \leq \max\left\{\mathrm{Wi}(n)\ ,\ \frac{(n^2 - n + 1)\cdot \|A\|}{\lambda(A) - \lambda_{\mathrm{ct}}(A)} + n - 1\right\} \tag{4.12}$$

We now give more precise bounds in the case that the matrix has all finite entries, which translates into its digraph being the complete digraph on n nodes. Trivially, every such matrix is irreducible. In particular, we strictly improve the bound given by Soto y Koelemeijer for these matrices (Theorem 2.21).

Theorem 4.5. *Let A be an $n \times n$ max-plus matrix with all finite entries, i.e., $A \in \mathbb{R}^{n\times n}$. Then its transient is bounded by*

$$T(A) \leq \max\left\{\mathrm{Wi}(n)\ ,\ \frac{2\|A\|}{\lambda(A)-\lambda_{\mathrm{ct}}(A)}+n-2\right\}\ . \tag{4.13}$$

We also prove an analogue of Theorem 4.4 for linear systems:

Theorem 4.6. *Let A be an irreducible $n \times n$ max-plus matrix and let $v \in \mathbb{R}^n$. Then the transient is bounded by*

$$T_v(A) \leq \max\left\{\mathrm{Wi}(n)\ ,\ \frac{\|v\|+\|A\|\cdot(n-1)}{\lambda(A)-\lambda_{\mathrm{ct}}(A)}\right\} \tag{4.14}$$

4.1.3 Comparison of the Different Schemes

In this section, we compare our two schemes with each other and also with the trivial choice of the critical digraph for B. We start with a computational comparison.

Theorem 4.7. *Let A be an irreducible max-plus matrix. The digraphs G_c and G_{ha} can be computed in polynomial time. The computation of the threshold digraphs $T^{\mathrm{ct}}(0)$ is NP-hard.*

Proof. Denote by n the number of nodes of the digraph $G(A)$.

For the computation of G_c, we can use Karp's algorithm. This takes $O(n^3)$ time.

Concerning G_{ha}, Schneider and Schneider [78] proved that a max-balancing of A can be computed in time $O(n^4)$. The same order of complexity is added if we examine the at most n^2 threshold digraphs (for each of them, the strongly connected components can be found in $O(n^2)$ time).

To show NP-hardness of the computation of $T^{\mathrm{ct}}(\mu)$, we reduce the Longest Path Problem [48, p. 213, ND29] to it. Consider the Longest Path Problem as a decision problem that takes as input an edge-weighted digraph with integer weights, a pair of nodes (i, j) with $i \neq j$ in the digraph, and an integer K. The output is YES if there exists a path of weight at least K from i to j. The output is NO if there is none. Observe that if $i \neq j$, then by inserting the edge (j, i) with weight $-K$, the Longest Path Problem can be polynomially reduced to the problem of calculating $T^{\mathrm{ct}}(0)$ by checking whether the new edge (j, i) belongs to $T^{\mathrm{ct}}(0)$. □

The relation between these schemes is as follows. The cycle threshold scheme is more precise, while the non-critical scheme is the coarser. We measure this in terms of the size of B and the value $\lambda(B)$.

Lemma 4.8. *G_c is a sub-digraph of G_{ha}, which is a sub-digraph of G_{ct}. In particular, $\lambda_{\mathrm{ct}} \leq \lambda_{\mathrm{ha}} \leq \lambda_{\mathrm{nc}}$.*

Proof. Evidently both G_{ct} and G_{ha} are super-digraphs of G_c, which is extracted from all non-critical nodes. This implies that $\lambda(B_{\mathrm{ct}}) \leq \lambda(B_{\mathrm{nc}})$ and $\lambda(B_{\mathrm{ha}}) \leq \lambda(B_{\mathrm{nc}})$.

We show that G_{ha} is a sub-digraph of G_{ct}. For this we can assume that the whole digraph is max-balanced, and notice first that $T^{\mathrm{ha}}(\mu) \subseteq T^{\mathrm{ct}}(\mu)$ for any value of μ. We also have that $T^{\mathrm{ha}}(\mu_1) \supseteq T^{\mathrm{ha}}(\mu_2)$ and $T^{\mathrm{ct}}(\mu_1) \supseteq T^{\mathrm{ct}}(\mu_2)$ for any $\mu_1 \leq \mu_2$. Now consider the value λ_{ct}. The components of $T^{\mathrm{ct}}(\lambda_{\mathrm{ct}})$ which do not contain the components of $G_c(A)$, have the property that any other cycle intersecting with them has a strictly smaller cycle mean. It follows that all edges of these components have cycle mean λ_{ct}. Indeed, suppose that there is a component containing an edge with a different weight. In this component, any cycle that contains this

edge also has an edge with weight strictly greater than λ_{ct}. The cycle cover property implies that there is a cycle containing this edge, where this edge has the smallest weight. The mean of that cycle is strictly greater than λ_{ct}, a contradiction. But then $T^{\mathrm{ha}}(\lambda_{\mathrm{ct}})$ contains these components as its strongly connected component's. In particular they do not contain the components of $G_c(A)$, hence $\lambda_{\mathrm{ct}} \leq \lambda_{\mathrm{ha}}$.

If $\mu = \lambda_{\mathrm{ct}} = \lambda_{\mathrm{ha}}$ then $T^{\mathrm{ha}}(\mu) \subseteq T^{\mathrm{ct}}(\mu)$, while we have shown that the components of $T^{\mathrm{ct}}(\mu)$ not containing the components of $G_c(A)$ are also components of $T^{\mathrm{ha}}(\mu)$. It follows that $G_{\mathrm{ha}} \subseteq G_{\mathrm{ct}}$.

If $\lambda_{\mathrm{ct}} < \lambda_{\mathrm{ha}}$ then we obtain that

$$G_{\mathrm{ct}} \supseteq T^{\mathrm{ct}}(\lambda_{\mathrm{ha}}) \supseteq T^{\mathrm{ha}}(\lambda_{\mathrm{ha}}) \supseteq G_{\mathrm{ha}} \ , \tag{4.15}$$

thus $G_{\mathrm{ha}} \subseteq G_{\mathrm{ct}}$ in any case, hence $G(B_{\mathrm{ct}}) \subset G(B_{\mathrm{ha}})$. □

The following example shows that all three schemes can differ.

Example 4.9. Consider the following matrix A, described by its digraph $G(A)$. The node set N of $G(A)$ is partitioned into $N = N_{\mathrm{c}} \cup N_{\mathrm{nc}} \cup N_{\mathrm{ha}} \cup N_{\mathrm{ct}}$. Further choose parameters $\lambda_{\mathrm{c}} > \lambda_{\mathrm{nc}} > \lambda_{\mathrm{ha}} > \lambda_{\mathrm{ct}} > z$ such that $\lambda_{\mathrm{ha}} \leq (2\lambda_{\mathrm{ct}} + \lambda_{nc})/3$. For example, set $\lambda_{\mathrm{c}} = 0$, $\lambda_{nc} = -1$, $\lambda_{\mathrm{ha}} = -3$, $\lambda_{\mathrm{ct}} = -4$, $z = -5$. The edges in $G(A)$ are as follows: For each $X \in \{\mathrm{c}, \mathrm{nc}, \mathrm{ha}, \mathrm{ct}\}$ the nodes in N_X are connected by one cycle of length $|N_X|$ whose edges all have weight λ_X. Choose an arbitrary node i_X in N_X. There are 6 additional edges in $G(A)$: The two edges $(i_{\mathrm{c}}, i_{\mathrm{nc}})$ and $(i_{\mathrm{nc}}, i_{\mathrm{c}})$ with weight λ_{nc}, the two edges $(i_{\mathrm{nc}}, i_{\mathrm{ha}})$ and $(i_{\mathrm{ha}}, i_{\mathrm{c}})$ with weight λ_{ct}, and the two edges $(i_{\mathrm{ha}}, i_{\mathrm{ct}})$ and $(i_{\mathrm{ct}}, i_{\mathrm{ha}})$ with weight z. Figure 4.1 depicts digraph $G(A)$.

The cycle cover condition holds in $G(A)$, and so A is visualized. Indeed, the cycle cover condition trivially holds for all edges contained in a single node set N_X and for the edges of the cycles of length 2. But it also holds for the two edges $(i_{\mathrm{ha}}, i_{\mathrm{c}})$ and $(i_{\mathrm{nc}}, i_{\mathrm{ha}})$ because $\lambda_{\mathrm{nc}} > \lambda_{\mathrm{ct}}$.

The critical digraph of A is induced by the node set N_{c}. This implies that the edge set of B_{nc} is induced by $N_{\mathrm{nc}} \cup N_{\mathrm{ha}} \cup N_{\mathrm{ct}}$, the edge set of B_{ha} by $N_{\mathrm{ha}} \cup N_{\mathrm{ct}}$, and the edge set of B_{ct} by N_{ct}.

4.2 Applications

In this section, we apply our transience bounds to the applications presented in Section 2.2.

Because the matrices' and vectors' entries are often integers, we take a closer look: If A is an integer matrix, i.e., all finite entries of A are integers, the term $\lambda - \lambda_{\mathrm{ha}}$ or $\lambda - \lambda_{\mathrm{ct}}$ cannot become arbitrarily small: This is obvious when $\lambda_{\mathrm{ha}} = -\infty$ or $\lambda_{\mathrm{ct}} = -\infty$; otherwise, let C_0 be a critical cycle, and let C_1 be a cycle such that $\lambda_0 = A(C_1)/\ell(C_1)$. Then we have

$$\lambda - \lambda_0 = \frac{A(C_0)\ell(C_1) - A(C_1)\ell(C_0)}{\ell(C_0)\ell(C_1)} \ ,$$

and so

$$\frac{1}{\lambda - \lambda_0} \leq (n - n_0) \cdot n_0 \leq \frac{n^2}{4} \ , \tag{4.16}$$

where n_0 denotes the number of non-isolated nodes in $G(B)$. It follows that, in case of integer matrices, the transient is in $O(\|A\| \cdot n^3)$ for a given initial vector.

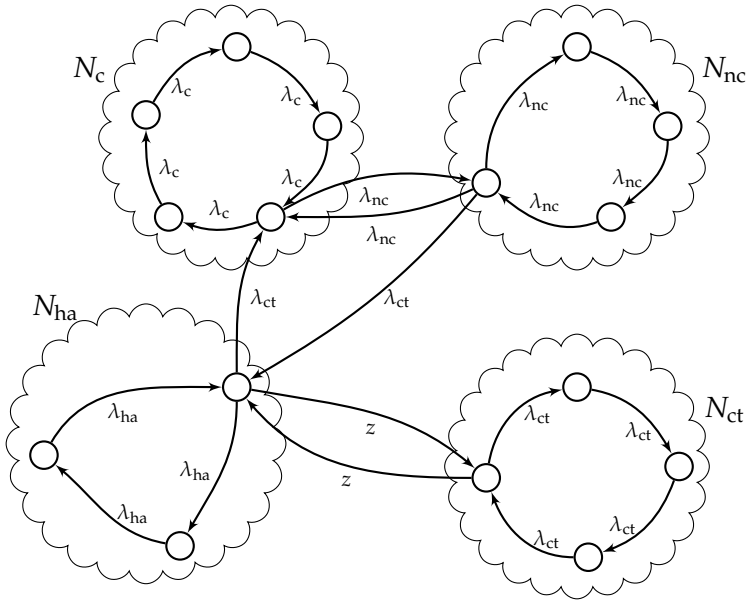


Figure 4.1: The digraph $G(A)$ of Example 4.9

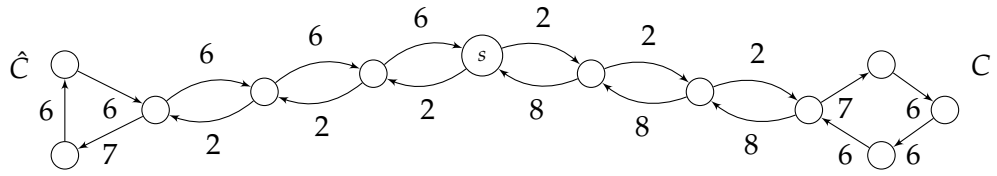


Figure 4.2: Digraph $H_{3,2}$

4.2.1 Synchronizers

As an example, let us consider a network synchronizer in the "ℓ-sized cherry" digraph family $H_{\ell,c}$, with $\ell \geq 2$ and $c \geq 1$, introduced by Even and Rajsbaum [41]. Each weighted digraph $H_{\ell,c}$ contains $n = 4\ell$ nodes and is constructed as follows: Let $\hat{C}$ and C be two cycles of length ℓ and $\ell+1$ respectively, with edge weights $3c$, except for one link per cycle with weight $3c+1$. There exists for both $\hat{C}$ and C a path of length ℓ to a distinct node s, and an anti-parallel path back. Hereby the edges in the path from s to C and from s to $\hat{C}$ have weight c, the edges in the path from $\hat{C}$ to s have weight $3c$, and from C to s, $4c$.

We observe that the nodes of $\hat{C}$ are the critical nodes, $\|A\| = 3c$, $n = 4\ell$ $n_c = \ell$, and $\lambda = 3c + 1/\ell$. Even and Rajsbaum's bound is

$$(112c - 16)\ell^3 + (32 - 12c)\ell^2 + 8\ell - 1 \ , \tag{4.17}$$

resulting in an upper bound of 5711 on the transient in case of $H_{3,2}$. It is $\lambda_{\mathrm{ha}} = 3c + 1/(\ell+1)$. Moreover for the critical sub-digraph G_c, the maximum girth of strongly connected components of G_{c} is $\hat{g} = \ell$. Thereby we may bound the synchronizer's transient with Theorem 4.3 by

$$\max\left\{4\ell^2 + \ell \ , \ 3c(4\ell^3 + 3\ell^2 - \ell\right\} = 12c\ell^3 + 9c\ell^2 - 3c\ell \tag{4.18}$$

resulting in an upper bound of 792 on the transient in case of $H_{3,2}$.

Since Even and Rajsbaum express transmission delays with respect to a discrete global clock, all weights are integers. All our transience bounds are in $O(\|A\| \cdot n^3)$. The example digraph family shows that this is asymptotically tight since Even and Rajsbaum proved that the transient for digraph $H_{c,\ell}$ is in $\Omega(c \cdot \ell^3) = \Omega(\|A\| \cdot n^3)$.

4.2.2 Cyclic Scheduling

We apply our transience bounds to a naturally arising special case of restrictions, namely those whose direct dependencies in a given iteration reach back at most to the previous one, and not further back (i.e., restrictions with binary heights). For this case, we are able to state explicit upper bounds, and thereby asymptotic upper bounds, on the number of task executions from where on the schedule becomes periodic.

However, we cannot directly apply our transience bounds on the digraph $G(A)$ obtained from G^u, since $G(A)$ is not necessarily strongly connected, as it is the case for the example in Figure 2.2.

However, we present a transformation of G^u yielding a strongly connected digraph $G(A)$ in case of binary heights, and has the same earliest schedule as the original digraph G^u: For every restriction between tasks i and j in G^u one can add the *redundant restriction* $\sigma(i,t) \geq \sigma(j,t-1) + P_j$ without changing the earliest schedule, since $\sigma(j,t) \geq \sigma(j,t-1)$ for all tasks j and $t \geq 1$. With this transformation we obtain:

Proposition 4.10. *If G^u is well-formed, has binary heights, and contains all redundant restrictions, then A is irreducible.*

Proof. It suffices to show that whenever there is an edge from i to j in G^u, then this edge also exists in $G(A)$. Because G^u contains all redundant restrictions, if there exists an edge from i to j, then there also exists an edge of height 1 from i to j. Hence there exists a walk of length 1 from i to j in G^u whose last (and only) edge has height 1. Hence, by definition of A, the entry $A_{i,j}$ is finite. This concludes the proof. □

Figures 4.3 and 4.2.2 depict the transformed digraph G^u of the above example with redundant restrictions and its corresponding weighted digraph $G(A)$. Observe that, in contrast to Figure 2.2, $G(A)$ is strongly connected in Figure 4.2.2.

Because of (2.10) and Proposition 4.10 we may now directly apply our transience bounds to (the strongly connected) digraph $G(A)$, obtaining upper bounds on the transients of the earliest schedule for G^u.

For the given example, $\|v\| = 11$, the critical cycle is from node 7 to 5 and back, $\lambda = 6.5$, $\lambda_{\text{ha}} = 6$, $B_{\max} = 8$, $A_{\min} = 1$, $\hat{g} = 2$, $\hat{\gamma} = 2$, $\hat{\text{ind}} = 0$, and we obtain a critical bound of 106.

Bounds in terms of the parameters of the original uniform graph G^u can be derived as well by relating graph parameters of G^u to parameters of $G = G(A)$. For that purpose, we denote by $\delta(G^u)$ and $\Delta(G^u)$ the minimum and maximum weight of an edge in G^u, respectively. From the definition of max-plus matrix A and initial vector v, it immediately follows that in case of binary heights, $n = |\mathcal{T}|$, $\|v\| \leq (|\mathcal{T}| - 1) \cdot \Delta(G^u)$, $A_{\max} \leq |\mathcal{T}| \cdot \Delta(G^u)$, $A_{\min} \geq \delta(G^u)$,

$$\lambda(G) = \max\{p(C)/h(C) \mid C \text{ is a closed walk in } G^u\} \ , \tag{4.19}$$

$\lambda_{\text{ha}}(A)$ is at most the second largest $A(C)/h(C)$ of closed walks C in G^u, and $\hat{g}$ is at most the number of links with height 1 in closed walks C in G^u with maximum $A(C)/h(C)$. As a

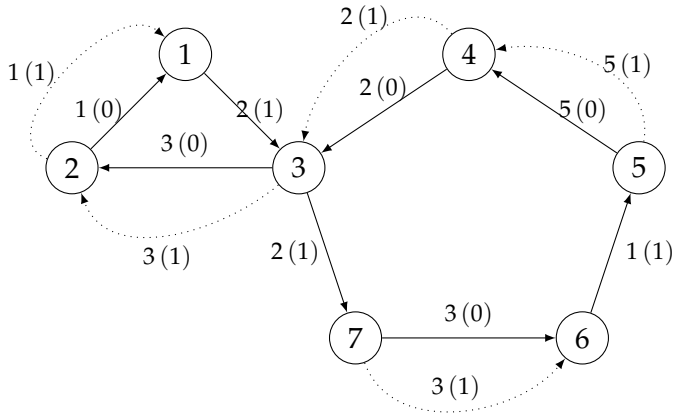


Figure 4.3: Uniform graph G^u with redundant restrictions (dotted)

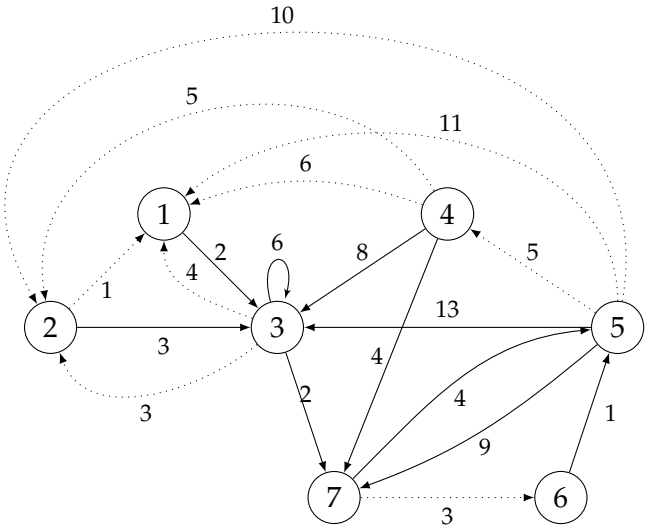


Figure 4.4: Digraph $G(A)$. Edges due to redundant restrictions are dotted

consequence of the above bounds, the transient is in $O(\|A\| \cdot |\mathcal{T}|^3) = O(|\mathcal{T}|^4)$, assuming constantly bounded $\delta(G^u)$ and $\Delta(G^u)$. To the best of our knowledge, this is the first asymptotic bound on the transient of an earliest schedule with tasks $\mathcal{T}$ and binary heights.

4.2.3 Link Reversal

In link reversal systems used for routing, we have $\lambda = 0$ and $\lambda_{\mathrm{ha}} \leq -1/(n - n_c) \leq -1/(n - 1)$. Since $\hat{g} = 1$, we obtain from Theorem 4.1, for $n \geq 3$, that the termination time is at most $(n-1)^2$, which improves on the asymptotic quadratic bound given by Busch and Tirthapura [20].

If the undirected support of initial digraph G_0 without the self-loop at the destination nodes is a *tree*, we can use our bounds to give a new proof that the termination time of Full Reversal routing is linear in n [25, Corollary 5]. In that particular case either $\lambda_{\mathrm{ha}} = -1/2$ or $\lambda_{\mathrm{ha}} = -\infty$. Theorem 4.1 yields the linear bound $2(n-1)$, whereas Hartmann and Arguelles arrive at $2n^2$.

In the scheduling case, the critical components have at least two nodes because there are no self-loops. Malka and Rajsbaum [65, Theorem 6.4] proved by reduction to Timed Marked Graphs that the transient is at most in the order of $O(n^4)$. Theorem 4.6, together with (4.16) shows a transience bound of $n^2 \cdot (n-1)/4 = O(n^3)$. Thus, our bounds allow to improve this asymptotic result by an order of n.

In the case of Full Reversal scheduling on *trees* we again obtain a bound linear in n: In this case it holds that $\lambda = -1/2$, and $\lambda_{\mathrm{ha}} = -\infty$. Thus the critical bound is N. Further, $G_c = G$ and $\hat{g} = 2$. Theorem 4.1 thus imply that $2n - 4$ is an upper bound on the transient of Full Reversal scheduling on trees. By contrast Hartmann and Arguelles again obtain the quadratic bound of $2n^2$.

4.3 Proof Strategy for Matrix Transients

This section describes our graph-based strategy to prove upper bounds on the transient of an irreducible $n \times n$ max-plus matrix A.

We start by defining, for a set $\mathbf{N}$ of nonnegative integers and a pair of nodes (i, j), an $\mathbf{N}$-*realizer* for (i, j) to be any walk of maximum A-weight in the set of walks in $\mathcal{W}(i \to j)$ with length in $\mathbf{N}$. As shown in the next proposition, of particular interest is the case of sets $\mathbf{N}$ of the form

$$\mathbf{N}_{\geq B}^{(t,\pi)} = \{s \in \mathbb{N}_0 \mid s \geq B \ \wedge \ s \equiv t \pmod{\pi}\} \tag{4.20}$$

where B, n, and π are positive integers.

Theorem 4.11. *Let A be an $n \times n$ max-plus matrix and let $i, j \in [n]$. If B and π are positive integers such that for every integer $t \geq B$, either*

- *there exists an $\mathbf{N}_{\geq B}^{(t,\pi)}$-realizer for (i, j) of length t or*
- *there exists no $\mathbf{N}_{\geq B}^{(t,\pi)}$-realizer for (i, j) at all,*

then the sequence $A_{i,j}^{\otimes t}$ is eventually periodic with period π and transient at most B.

Proof. Let $t \geq B$. If there is no $\mathbf{N}_{\geq B}^{(t,\pi)}$ realizer, then clearly $A_{i,j}^{\otimes t} = A_{i,j}^{\otimes t+\pi} = -\infty$.

So let W_t be an $\mathbf{N}_{\geq B}^{(t,\pi)}$-realizer for (i,j) of length t. Denote by $X(t)$ the set of walks W in $\mathcal{W}(i \to j)$ with $\ell(W) \in \mathbf{N}_{\geq B}^{(t,\pi)}$, and let $x(t)$ be the maximum of values $A(W)$ where $W \in X(t)$. It follows that $x(t) = A(W_t)$.

From $t + \pi \equiv t \pmod{\pi}$ follows $X(t+\pi) = X(t)$ and so $x(t+\pi) = x(t)$. Moreover, we have $\mathcal{W}^t(i \to j) \subseteq X(t)$ and $\mathcal{W}^{t+\pi}(i \to j) \subseteq X(t+\pi)$, which implies $A_{i,j}^{\otimes t} \leq x(t)$ and $A_{i,j}^{\otimes t+\pi} \leq x(t+\pi)$. Conversely because $W_t \in \mathcal{W}^t(i \to j)$, we have $A_{i,j}^{\otimes t} \geq A(W_t) = x(t)$. Similarly, $A_{i,j}^{\otimes t+\pi} \geq A(W_{t+\pi}) = x(t+\pi)$. Since $x(t+\pi) = x(t)$, it follows that $A_{i,j}^{\otimes t} = A_{i,j}^{\otimes t+\pi}$. This concludes the proof. □

We say that a max-plus matrix A is *normalized* if $\lambda(A) = 0$. The sequence of powers of a normalized irreducible max-plus matrix is eventually periodic without linear defect. The *normalization* of a matrix A is the matrix $\bar{A}$ defined by $\bar{A}_{i,j} = A_{i,j} - \lambda(A)$. The normalization of a max-plus matrix with $\lambda(A) \neq -\infty$ is normalized. Its transient does not change when passing to the normalization, nor does its critical digraph. Also, $\|\bar{A}\| = \|A\|$.

Based on Theorem 4.11, we now define a strategy for determining upper bounds on system transients. The strategy includes the following parameters:

- a completely reducible sub-digraph H of the critical digraph $G_c(A)$ that contains at least one cycle of each critical component. For a node k of H, denote by H_k its strongly connected component and by $d_k = \gamma(H_k)$ its cyclicity.
- integers B_{crit}^H, $B_{\text{red}}^{H,k}$, and $B_{\text{pump}}^{H,k}$ for all nodes k of H

We choose $\pi = \gamma(H)$ to be the cyclicity of H, i.e., the least common multiple of the d_k. Let B be the maximum of B_{crit}^H and the $B_{\text{red}}^{H,k} + B_{\text{pump}}^{H,k}$.

We want to show that the transient $T(A)$ of A is at most B. Let $t \geq B$ and $i, j \in [n]$.

1. *Normalized matrix.* Because neither our bounds nor the transients change when passing to the normalization of A, we assume A to be normalized. Let W_0 be an $\mathbf{N}_{\geq B}^{(t,\pi)}$-realizer for (i,j) if it exists. If not, there is nothing to show.

2. *Critical bound.* Show that $B \geq B_{\text{crit}}^H$ implies that the realizer W_0 can be chosen to contain at least one node of H. Let k be the first node of H on W_0.

3. *Walk reduction.* Next we show that we can *reduce* W_0, by removing subcycles and adding critical cycles, to arrive at a new walk $\hat{W}_0$ that (a) contains the critical node k, (b) whose length $\ell(\hat{W}_0)$ is in the same residue class modulo d_k as $\ell(W_0)$, and (c) $\ell(\hat{W}_0)$ is upper-bounded by $B_{\text{red}}^{H,k}$.

4. *Pumping in the critical digraph.* In this step, we completely walk $\hat{W}_0$ to length t by adding a critical closed walk in H_k. This yields an $\mathbf{N}_{\geq B}^{(t,\pi)}$-realizer because removing cycles at most increases the weight and adding a critical closed path does not change the weight due to the normalization assumption, i.e., $A(W) \geq A(W_0)$, which is the same as $A(W) = A(W_0)$ because W_0 is a realizer.

   We show that there are closed walks at node k in H of length s for every multiple s of d_k larger or equal to $B_{\text{pump}}^{H,k}$. If we set $s = t - \ell(\hat{W}_0)$, then s is a multiple of d_k and is larger

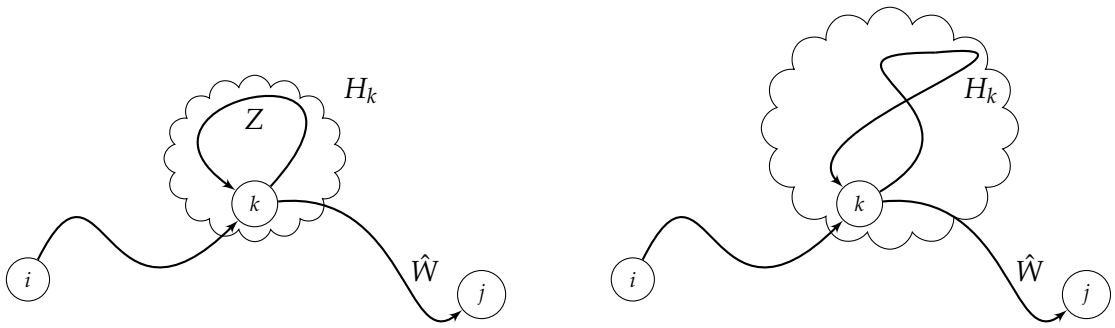


Figure 4.5: Pumping repetitively (left) and exploratively (right)

or equal to $B - B^{H,k}_{\text{red}} \geq B^{H,k}_{\text{pump}}$. We can hence complete $\hat{W}_0$ by adding to it a critical closed walk of length s to obtain a new walk W of length t from i to j.

Theorem 4.11 then shows that B is a bound on the transient.

To deduce the bounds we present in Section 4.7, we make the following choices for the parameter H. The choices for the critical bound, the reduction bound, and the pumping bound are more involved and, in fact, the heart of the proof.

4.4 Pumping in the Critical Digraph

This section is dedicated to step 4 of the proof strategy, i.e., how to pump in the critical digraph. Our ability to pump is limited to find closed walks in the sub-digraph H. We present two possible, reasonable, and fundamentally different choices for H: the repetitive choice and the explorative one.

In the repetitive case, we choose a shortest cycle in every critical component and define H as their induced digraph. The index of convergence of all H_k are zero and we can immediately pump by multiples of the girth $d_k = g(H_k)$ by inserting copies of the chosen cycle.

For the explorative case, we choose $H = G_c(A)$ to be the whole critical digraph. Here, we can pump by arbitrary multiples of the cyclicity of k's critical component, provided they are not smaller than the component's index of convergence.

There is hence a trade-off between the two choices: In the repetitive case the index of convergence of H is smaller, in fact zero, whereas in the explorative case the cyclicity d_k is smaller, which allows for a smaller walk reduction bound. Figure 4.5 illustrates the difference between the two choices.

We are thus led to introduce the parameters $\hat{g}(A)$, the maximum girth of critical components of A, the parameter $\hat{\gamma}(A)$, the maximum cyclicity of critical components of A, and the parameter $\hat{\text{ind}}(A)$, the maximum index of convergence of critical components of A. In fact, $\hat{\text{ind}}(A) = \text{ind}(G_c(A))$.

In any case, we will choose the pumping bound to satisfy

$$B^{H,k}_{\text{pump}} \geq d_k \cdot \left\lceil \frac{\text{ind}(H_k)}{d_k} \right\rceil - d_k + 1 \ , \tag{4.21}$$

which is always upper-bounded by the index of convergence $\text{ind}(H_k)$. In the repetitive case, where $\text{ind}(H_k) = 0$, it is equal to $-d_k + 1$, and hence can even be negative. While this seems

bizarre, it is not counter-intuitive because the pumping bound in step 4 only has to guarantee that one is able to pump in multiples of d_k greater or equal to the bound. Thus, even if $B^{H,k}_{\text{pump}} = -d_k + 1$, every multiple of d_k greater or equal to $B^{H,k}_{\text{pump}}$ is nonnegative. The general reasoning for the choice of the pumping bound in (4.21) is the following elementary lemma.

Lemma 4.12. *If $a \equiv b \pmod d$, then $a \leq b$ if and only if $a - d + 1 \leq b$.*

Hence, by the lemma, every multiple s of d_k greater or equal to $B^{H,k}_{\text{pump}}$ is greater or equal to $\text{ind}(H_k)$ because both $d_k \cdot \lceil \text{ind}(H_k)/d_k \rceil$ and s are congruent 0 modulo d_k.

4.5 Walk Reductions

This section concerns step 3 of the proof strategy, i.e., we give choices for $B^{H,k}_{\text{red}}$. To do this, we deal with walk reductions in a more general context: Given a sub-digraph H of G, we seek to reduce the walks that visits H by removing cycles and inserting cycles of H in such a way that the reduced walk's length is lower than some threshold while satisfying length congruence and connectedness conditions.

We introduce two notions of "walk reduction thresholds" that can be both used to give a possible choice for the reduction bound $B^{H,k}_{\text{red}}$ in the proof strategy. The first one is more general, while the second one is more tailored to our specific proof strategy for transience bounds and makes use of the fact that we assume k to be the *first* node of H on walk W in step 2 of the proof strategy. This amounts to a less natural definition of the second walk reduction that has the upside that we can formally fix node k once and for all in the walk reductions. The fact that we assume that it appears before all other occurrence of nodes of H yields a sharper bound with one of our walk reduction methods. So, while theoretically we only need the second definition to prove our transience bounds for max-plus matrices, the first, more natural, definition, on the other hand, is of its own graph-theoretic interest.

Definition 4.13 (Walk reduction threshold)**.** Let H be a sub-digraph of G, k a node of H, and d a positive integer. The *walk reduction threshold* $T^{d,k}_{\text{red}}(G, H)$ is the smallest nonnegative integer T for which the following holds: For all walks $W \in \mathcal{W}(i \xrightarrow{k} j)$ there is a walk $\hat{W} \in \mathcal{W}(i \xrightarrow{k} j)$ obtained from W by removing cycles and possibly inserting cycles of H such that $\ell\left(\hat{W}\right) \equiv \ell(W) \pmod d$, where d is the cyclicity of H_k, and $\ell\left(\hat{W}\right) \leq T$.

We can choose $B^{H,k}_{\text{red}} = T_{\text{red}}(G(A), H)$ in the proof strategy.

We will present one walk reduction method, the "Arithmetic Method", that makes use of the fact that the fixed node k in step 2 is the first one of H on the considered walk. To fully utilize this method, we also give a modified definition of the walk reduction threshold that restricts k to be the first node of H on walk W:

Definition 4.14 (Modified walk reduction threshold)**.** Let H be a sub-digraph of G, k a node of H, and d a positive integer. The *modified walk reduction threshold* $\tilde{T}^{d,k}_{\text{red}}(G, H)$ is the smallest nonnegative integer T for which the following holds: For all walks $W \in \mathcal{W}(i \xrightarrow{k} j)$ on which k is the first node of H there is a walk $\hat{W} \in \mathcal{W}(i \xrightarrow{k} j)$ obtained from W by removing cycles and possibly inserting cycles of H such that $\ell\left(\hat{W}\right) \equiv \ell(W) \pmod d$ and $\ell\left(\hat{W}\right) \leq T$.

Here we can choose $B^{H,k}_{\text{red}} = \tilde{T}^{d_k,k}_{\text{red}}(G(A), H)$ in the proof strategy.

The modified walk reduction threshold is never larger than the ordinary walk reduction threshold because it constrains the considered walks more, the rest of the definition being equal.

In the rest of this section, we give three ways to bound the walk reduction thresholds; two for the first threshold and one for the modified one. Each of the three walk reduction methods has its strengths and weaknesses. The first and second method only remove cycles; the first one even leaves the order of nodes on the walk intact, while the second one does not.

We dub the first method "Repeated Cycle Removal" and we present it in Section 4.5.1. The advantage of leaving the order of nodes intact while reasoning about the structure of the reduced walk is that we will have a description of the walk in terms of a bounded number of alternating paths and cycles. This allows the use of the cab driver's diameter and the circumference as a parameter.

However, if all one has are the two trivial bounds $\text{cdd}(G) \leq n-1$ and $\text{cf}(G) \leq n$, then the second method is strictly better. We call it the "Arithmetic Method" and present it in Section 4.5.2. It proceeds by observing that a walk necessarily includes a closed subwalk whose length is a positive multiple of d as soon as there is a node that appears at least $d+1$ times on the walk. By splitting the walk into two parts, one up to k and one from k on, and using the observation on each of the two parts separately, one can show a bound on the walk reduction threshold. By a closer inspection of the argument, we also find a sharper bound on the modified walk reduction threshold. Because of the way the proof is constructed, we do not obtain a description in terms of alternating paths and cycles, and can therefore not introduce the cab driver's diameter or the circumference as a parameter in the bound.

The third method, the "Cycle Decomposition" method presented in Section 4.5.3, is similar to the technique used in Section 3.2.2 to prove the generalization of Wielandt's bound in the case of the existence of a critical Hamiltonian cycle. In a similar vein, we prove an upper bound on the walk reduction threshold for the case that the sub-digraph H is induced by a Hamiltonian cycle. The resulting bound is strictly better than the other two bounds in this case.

4.5.1 Walk Reduction by Repeated Cycle Removal

Given a walk W, a positive integer d, and a node k of W, we define a reduced walk, denoted $\text{Red}_{d,k}(W)$, such that (a) it contains node k and has the same start and end nodes as W, (b) its length is in the same residue class modulo d as W's length, and (c) its length is bounded by $(d+1)\text{cdd}(G) + (d-1)\text{cf}(G)$. Properties (a) and (b) can be achieved by removing a collection of cycles from W whose combined length is divisible by d and whose removal retains connectivity to k. The key point of the reduction is that we can iterate this removal until the resulting length is bounded as demanded by (c).

We call a finite, possibly empty, sequence of nonempty subcycles $\mathcal{S} = (C_1, C_2, \ldots, C_n)$ a *cycle pattern of a walk* W if there exist walks $U_0, U_1, \ldots, U_n$ such that

$$W = U_0 \cdot C_1 \cdot U_1 \cdot C_2 \cdots U_{n-1} \cdot C_n \cdot U_n \ . \tag{4.22}$$

The choice of the U_m's in (4.22) may be not unique, and we fix some global choice function to make it deterministic. Then we define *the removal of* $\mathcal{S}$ *from* W as

$$\text{Rem}(W, \mathcal{S}) = U_0 \cdot U_1 \cdots U_n \ . \tag{4.23}$$

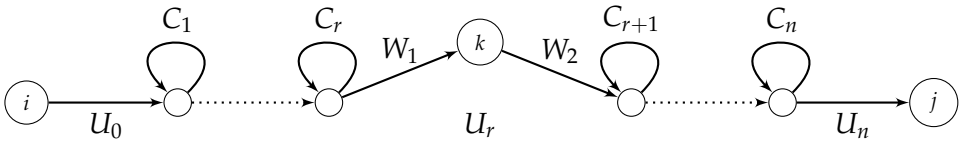


Figure 4.6: Structure of the reduced walk $\hat{W} = \mathrm{Red}_{d,k}(W)$

The walks W and $\mathrm{Rem}(W, \mathcal{S})$ have the same start and end nodes. Furthermore the length satisfies $\ell\big(\mathrm{Rem}(W, \mathcal{S})\big) = \ell(W) - \ell(\mathcal{S})$ where $\ell(\mathcal{S}) = \sum_{C \in \mathcal{S}} \ell(C)$. In particular, $\mathrm{Rem}(W, \mathcal{S}) = W$ if and only if $\ell(\mathcal{S}) = 0$, i.e., $\mathcal{S}$ is the empty cycle pattern.

Given any node k of a walk W, let $\mathbf{S}_k(W)$ denote the set of cycle pattern $\mathcal{S}$ of W whose removal does not impair connectivity to k, i.e., k is a node of $\mathrm{Rem}(W, \mathcal{S})$. Further for any positive integer d, define $\mathbf{S}_{d,k}(W)$ as the subset of cycle pattern $\mathcal{S} \in \mathbf{S}_k(W)$ that, in addition, leave the length's residue class modulo d intact, i.e., $\ell(\mathcal{S}) \equiv 0 \pmod{d}$. The set $\mathbf{S}_{d,k}(W)$ is not empty, because k is a node of W and we can hence choose $\mathcal{S}$ to be the empty cycle pattern.

Choose $\mathcal{S} \in \mathbf{S}_{d,k}(W)$ such that $\ell(\mathcal{S})$ is maximal. There may be several possible choices for $\mathcal{S}$, and we again fix some global choice function to make the choice deterministic; then set

$$\mathrm{Step}_{d,k}(W) = \mathrm{Rem}(W, \mathcal{S}) \ . \tag{4.24}$$

The limit

$$\mathrm{Red}_{d,k}(W) = \lim_{t \to \infty} \mathrm{Step}^t_{d,k}(W) \tag{4.25}$$

exists because the sequence of walks $(\mathrm{Step}^t_{d,k}(W))_{t \geq 0}$ is stationary after at most $\ell(W)$ steps, and we call it the (d,k)*-reduction of* W. More specifically, $\mathrm{Red}_{d,k}(W) = W$ if and only if $\mathbf{S}_{d,k}(W)$ is reduced to the sole empty cycle pattern. The walks W and $\mathrm{Red}_{d,k}(W)$ have the same start and end nodes. Also, k is a node of $\mathrm{Red}_{d,k}(W)$ and $\ell\big(\mathrm{Red}_{d,k}(W)\big) \equiv \ell(W) \pmod{d}$.

Theorem 4.15. *Let G be a digraph. For each positive integer d and each node k, the length of the (d,k)-reduction of any walk W containing node k satisfies*

$$\ell\big(\mathrm{Red}_{d,k}(W)\big) \leq (d+1)\mathrm{cdd}(G) + (d-1)\mathrm{cf}(G) \ . \tag{4.26}$$

Proof. We denote $\hat{W} = \mathrm{Red}_{d,k}(W)$. By definition of the (d,k)-reduction, $\mathrm{Red}_{d,k}(\hat{W}) = \hat{W}$. Let $\mathcal{S}$ be any cycle pattern of $\hat{W}$ in $\mathbf{S}_k(\hat{W})$, and let n be the number of cycles of $\mathcal{S}$. We first show that $n \leq d-1$. Indeed, suppose for contradiction that $n \geq d$. Then Lemma 3.15 implies that there exists a nonempty subsequence of $\mathcal{S}$ that is in $\mathbf{S}_{d,k}(\hat{W})$, which contradicts $\mathrm{Red}_{d,k}(\hat{W}) = \hat{W}$.

Now let us choose $\mathcal{S}$ in $\mathbf{S}_k(\hat{W})$ with maximal $\ell(\mathcal{S})$. If $\mathcal{S} = (C_1, C_2, \ldots, C_n)$, then there exist walks $U_0, U_1, \ldots, U_n$ such that

$$\hat{W} = U_0 \cdot C_1 \cdot U_1 \cdot C_2 \cdots U_{n-1} \cdot C_n \cdot U_n \ . \tag{4.27}$$

By definition of $\mathbf{S}_k(\hat{W})$, k is a node of $\mathrm{Rem}(\hat{W}, \mathcal{S})$. Hence there exists some index r such that k is a node of U_r. Each U_m with $m \neq r$ is a (possibly empty) path, because otherwise we could add a nonempty subcycle of U_m to $\mathcal{S}$, a contradiction to the maximality of $\ell(\mathcal{S})$. Similarly, if $U_r = W_1 \cdot W_2$ such that k is the end node of W_1, then both W_1 and W_2 are (possibly empty) paths. Hence, apart from the at most $(d-1)$ cycles in $\mathcal{S}$, the reduced walk $\hat{W}$ consists of at most $(d+1)$ subpaths. Its structure is shown in Figure 4.6. □

This bound on the length of the reduced walk gives the following bound on the walk reduction threshold:

Corollary 4.16. *We always have* $T_{\text{red}}^{d,k}(G,H) \leq (d+1)\text{cdd}(G) + (d-1)\text{cf}(G)$. *If* n *is the number of nodes of* G, *then* $T_{\text{red}}^{d,k}(G,H) \leq (d-1) + 2d(n-1)$.

4.5.2 Walk Reduction by Arithmetic Method

We begin with the observation that there is a closed subwalk whose length is a positive multiple of d as soon as there is one node that appears at least $d+1$ times in the walk. It is the key lemma in the theorem that follows.

Lemma 4.17. *Let* $d \in \mathbb{N}$ *and let* $W \in \mathcal{W}(i \to j)$. *Then there exists a walk* $W' \in \mathcal{W}(i \to j)$ *obtained from* W *by removing cycles such that* $\ell(W') \equiv \ell(W) \pmod{d}$ *and each node appears at most* d *times in* W'.

Proof. Let $i_0, i_1, \ldots, i_L$ be the sequence of nodes of W.

If a given node appears twice, first as i_a and then as i_b and if $a \equiv b \pmod{d}$, then the subwalk defined by $i_0, \cdots, i_a, i_{b+1}, \cdots, i_L$ is strictly shorter than W and has the same length modulo d.

Iterating this process, we get a sequence of subwalks of W. Since the sequence of length is strictly decreasing, the sequence is finite and we denote the last walk by W'.

Obviously, $\ell(W') \equiv \ell(W) \pmod{d}$ and a node does appear twice as i_a and i_b only if $a \not\equiv b \pmod{d}$, so the pigeonhole principle implies that it appears at most d times; otherwise there would exist i_a and i_b with $a \equiv b \pmod{d}$. □

We can now prove the following upper bound on the modified walk reduction threshold. Note that the bound is decreasing in the number of nodes of the sub-digraph H.

Theorem 4.18. *If* G *has* n *nodes and* k *is a node of* H, *then* $\tilde{T}_{\text{red}}^{d,k}(G,H) \leq (d+1)n - |V(H)| - 1$.

Proof. Let $W \in \mathcal{W}(i \xrightarrow{H} j)$ and let k be the first node of H appearing on W.

We proceed with the following steps:

1. Let W_1 be the shortest prefix of W from i to k, and let W_2 be the remaining subwalk. So we have
$$W_1 \in \mathcal{W}(i \to k),\ W_2 \in \mathcal{W}(k \to j),\ \ell(W_1) + \ell(W_2) = \ell(W) \tag{4.28}$$

2. As long as there is a node l that appears twice in W_1 and at least once in W_2, we can write $W_1 = U_1 \cdot U_2 \cdot U_3$ and $W_2 = V_1 \cdot V_2$, where U_1, U_2, V_1 end with l and U_2, U_3, V_2 start with l. Thus, we can replace W_1 by $U_1 \cdot U_3$ and W_2 by $V_1 \cdot U_2 \cdot V_2$. Equation (4.28) still holds, but now l appears only once in W_1. Step 2 is over when all nodes that appear more than once in W_1 do not appear in W_2. Let us denote the resulting walks by W_3 and W_4 respectively.

3. Apply Lemma 4.17 to W_3 and W_4, obtaining W_1' and W_2' respectively.

4. Set $\hat{W} = W_1' \cdot W_2'$.

Obviously, $\ell(\hat{W}) \equiv \ell(W_1) + \ell(W_2) \equiv \ell(W) \pmod{d}$. Now we take a node l of V and bound the number of its appearances.

- If l is a node of W_2', then it is also a node of W_4, it appears at most once in W_3, thus also in W_1'. Therefore, l appears at most $d+1$ times in $\hat{W}$.
- If l is not a node of W_2', then it appears only in W_1', thus at most d times.
- If $l \in V(H) \setminus \{k\}$, then we have a sharper bound since it only appears in W_2': It appears at most d times.
- Node $l = k$ appears only once in W_1'; at the end. Thus k also appears at most d times because it appears at most d times in W_2'.

In summary, nodes of H appear at most d times, all other nodes at most $d+1$ times. The total number of appearances of all nodes in $\hat{W}$ is hence at most $d \cdot |V(H)| + (d+1)(n - |V(H)|) = (d+1)n - |V(H)|$, so $\ell\left(\hat{W}\right)$ is bounded by $(d+1)n - |V(H)| - 1$ as claimed. □

4.5.3 Walk Reduction by Cycle Decomposition

In this subsection, we use the same technique used in Section 3.2.2 to prove upper bounds on the walk reduction threshold.

Theorem 4.19. *Let Z be a cycle of G and k a node of Z. Then, if n denotes the number of nodes of G, we have:*

$$T_{\text{red}}^{\ell(Z),k}(G,Z) \leq (n-1)\,\text{cf}(G) + \ell(Z) \tag{4.29}$$

This same proof method also leads to:

Theorem 4.20. *If Z is a Hamiltonian cycle of G. Then, if n denotes the number of nodes of G, we have: $T_{\text{red}}^{n,k}(G,Z) \leq n^2 - n + 1$.*

We prove the two theorems in the rest of this subsection.

To any walk $W \in \mathcal{W}(i \xrightarrow{k} j)$, we apply the following procedure.

1. We choose a decomposition of the walk W into a path P and a collection of cycles Z_α for $\alpha \in S$.

   We denote by n_W the number of nodes in the walk's multigraph $M(W)$ and by cdd_W the maximum length of a path in $M(W)$.

2. We take a subset R_1 of S of smallest cardinality such that $M(P) \cup M(Z) \cup \bigcup_{\alpha \in R_1} M(Z_\alpha)$ is connected and contains all nodes appearing in W. We have $|R_1| \leq n_W - \ell(Z) - \ell(P)$ because the connection of $M(P) \cup M(Z)$ with all the nodes of W can be ensured by adding at most $n_W - \ell(Z) - \ell(P)$ edges of W to $M(P) \cup M(Z)$, and hence by adding to it at most $n_W - \ell(Z) - \ell(P)$ of the cycles Z_α.

3. Let R_2 be a result of recursively removing from $S \setminus R_1$ sets of indices whose corresponding cycles have a combined length that is a multiple of $\ell(Z)$. By Lemma 3.15, $|R_2| \leq \ell(Z) - 1$. Let $R = R_1 \cup R_2$ and set $\text{cf}_W = \max_{\alpha \in R} \ell(Z_\alpha)$. It is $|R| \leq n_W - \ell(P) - 1$.

4. If $M_0 = M(P) \cup \bigcup_{\alpha \in R} M(Z_\alpha)$ is connected, then we choose walk $\hat{W} \in \mathcal{W}(i \xrightarrow{k} j)$ such that $M(\hat{W}) = M_0$.

5. Otherwise, we choose $\hat{W} \in \mathcal{W}(i \xrightarrow{k} j)$ such that $M(\hat{W}) = M_0 \cup M(Z)$.

By construction, $\ell\left(\hat{W}\right) \equiv \ell(W) \pmod{\ell(Z)}$ in both cases. We now bound the length of $\hat{W}$: If $M(W)$ does not contain any cycles, then $R = \emptyset$ and thus $\ell\left(\hat{W}\right) \leq \ell(P) \leq n_W - 1$. We hence assume that $M(W)$ contains cycles in the rest of the proof. In particular, this assumption means that $\mathrm{cf}_W \geq 1$.

If M_0 is connected, then

$$\begin{aligned} \ell\left(\hat{W}\right) &= \ell(P) + \sum_{\alpha \in R} \ell(Z_\alpha) \leq \ell(P) + \mathrm{cf}_W \cdot (n_W - \ell(P) - 1) \\ &\leq (n_W - 1) \cdot \mathrm{cf}_W + \ell(P) \cdot (1 - \mathrm{cf}_W) \leq (n_W - 1) \cdot \mathrm{cf}_W \ , \end{aligned} \tag{4.30}$$

which gives

$$\ell\left(\hat{W}\right) \leq (n-1)\,\mathrm{cf}(G) \ . \tag{4.31}$$

If M_0 is not connected, then we have, in the same vein,

$$\ell\left(\hat{W}\right) \leq (n_W - 1)\,\mathrm{cf}_W + \ell(Z) \leq (n-1)\,\mathrm{cf}(G) + \ell(Z) \ , \tag{4.32}$$

which completes the proof of Theorem 4.19.

On the other hand, there is some $\hat{\alpha} \in R$ such that $\ell(P) + \ell(Z_{\hat{\alpha}}) \leq n_W - 1$, because otherwise every Z_α with $\alpha \in R$ would share a node with P. Because $|R \setminus \{\hat{\alpha}\}| \leq n_W - \ell(P) - 2$, we have

$$\begin{aligned} \ell(\hat{W}) &= \ell(Z) + \ell(P) + \ell(Z_{\hat{\alpha}}) + \sum_{\substack{\alpha \in R \\ \alpha \neq \hat{\alpha}}} \ell(Z_\alpha) \\ &\leq \ell(Z) + n_W - 1 + (n_W - \ell(P) - 2) \cdot n_W \\ &\leq n_W \cdot (n_W - 1) + \ell(Z) - 1 \ . \end{aligned} \tag{4.33}$$

If $\ell(Z) = n$, i.e., Z is Hamiltonian, then R_1 is empty and the cycles in R_2 have length at most $n-1$. So we obtain $\ell\left(\hat{W}\right) \leq (n-1)(n-1) + \ell(Z) = n^2 - n + 1$ in the same way as (4.33). This proves Theorem 4.20.

4.6 Critical Bound

In this section, we seek to give possible choices for the critical bound B^H_{crit}. We hence want to show the existence of realizers that visit H for sufficiently lengths. Formally, define the *H-criticality threshold* of A, written as T^H_{crit}, to be the smallest T such that, whenever there exists an $\mathbf{N}^{(t,\pi)}_{\geq T}$-realizer of (i,j), then there also exists $\mathbf{N}^{(t,\pi)}_{\geq T}$-realizer of (i,j) that contains a node of H.

We prove bounds on the H-criticality threshold by reducing thresholds for smaller H to that of larger H. Clearly, a priori, the criticality thresholds of larger H are lower than that of smaller H. In our proof strategy, we decided to choose H as a completely reducible subdigraph of the critical digraph that contain at least one cycle of every critical component. We call each such H a *representing sub-digraph* of the critical digraph.

By repeating a critical closed walk connecting an arbitrary critical node with a node in H, we can see that the criticality threshold of every representing sub-digraph is upper bounded by the threshold of the critical digraph:

Lemma 4.21. *If H contains at least one node of every critical component, then $T^H_{\mathrm{crit}} \leq T^{G_c}_{\mathrm{crit}}$.*

4.6.1 Avoiding Super-Digraphs of the Critical Digraph

In this subsection we will consider a super-digraph G_0 of the critical digraph and study the G_0-criticality threshold by comparing weights of walks that do not visit G_0 to ones that do. In subsequent subsections, we show how to derive G_0-criticality thresholds for smaller G_0 from the results of this section.

Theorem 4.22. *Let A be an irreducible $n \times n$ max-plus matrix, let G_0 be a super-digraph of the critical digraph $G_c(A)$, and let B be defined as in* (4.1)*. If $\lambda(B) = -\infty$, then $T_{\text{crit}}^{G_0}(A) \leq \text{cdd}(G(B)) + 1 \leq n - 1$. Otherwise,*

$$T_{\text{crit}}^{G_0}(A) \leq \frac{T_{\text{red}}(G(A), H) \cdot (\lambda(A) - A_{\min}) + \text{cdd}(G(B)) \cdot (B_{\max} - \lambda(B))}{\lambda(A) - \lambda(B)} \tag{4.34}$$

for all representing sub-digraphs H of $G_c(A)$, where $A_{\min}$ is the minimal finite entry of A and $B_{\max}$ is the maximum (necessarily finite) entry of B.

Proof. Because neither the critical threshold nor the bound changes, we assume A to be normalized and $\lambda(B)$ to be finite because the other case is easy. This implies that $B_{\max} \geq \lambda(B)$ and $\lambda(A) = 0 > \lambda(B) \geq A_{\min}$. Set $G_A = G(A)$, $G_B = G(B)$, and $G_c = G_c(A)$.

Denote by T the right-hand side of (4.34). Assume by contradiction that there exist i, j and a $t \geq T$ such that all $\mathbf{N}_{\geq T}^{(\pi,t)}$-realizers for (i, j) are walks in $G(B)$ and that there is one.

Let V be such a realizer. By reducing V to a path we see that it has weight at most

$$A(V) \leq \text{cdd}(G_B) \cdot (B_{\max} - \lambda(B)) - t \cdot \lambda(B) \ . \tag{4.35}$$

On the other hand, there is some walk $W_0 \in \mathcal{W}(i \xrightarrow{H} j)$ with $\ell(W_0) \equiv t \pmod{\pi}$. By definition of the walk reduction threshold, there is a strongly connected component H_k of G_c and a walk $\hat{W}_0 \in \mathcal{W}(i \xrightarrow{H_k} j)$ obtained from W_0 by removing cycles and adding cycles in H_k such that $\ell(\hat{W}_0) \equiv \ell(W_0) \equiv t \pmod{d_k}$ and $\ell(\hat{W}_0) \leq T_{\text{red}}(G_A, G_c)$. Because d_k is the cyclicity of k's component H_k, there is a closed walk in H that can be added to $\hat{W}_0$ to obtain a walk W with $\ell(W) \equiv t \pmod{\pi}$, $\ell(W) \geq T$, and

$$A(W) = A(\hat{W}_0) \geq A_{\min} \cdot \ell(\hat{W}_0) \geq A_{\min} \cdot T_{\text{red}}(G_A, G_c) \ . \tag{4.36}$$

By the contradictory assumption, because W contains nodes of G_c and hence of G_0, W is not a $\mathbf{N}_{\geq T}^{(\pi,t)}$-realizer, i.e., $A(W) < A(V)$. But this means

$$\begin{aligned} 0 &< A(V) - A(W) \\ &\leq \text{cdd}(G_B) \cdot (B_{\max} - \lambda(B)) - t \cdot \lambda(B) + T_{\text{red}}(G_A, G_c) \cdot (-A_{\max}) \ , \end{aligned} \tag{4.37}$$

which implies

$$t < \frac{T_{\text{red}}(G_A, G_c) \cdot (-A_{\max}) + \text{cdd}(G_B) \cdot (B_{\max} - \lambda(B))}{-\lambda(B)} = T \ , \tag{4.38}$$

a contradiction to $t \geq T$. □

We can also follow a different, more constructive way to show existence of a realizer that visits G_0. In the proof of the previous theorem, we also constructed a realizer, but we did not pay all too much attention to its length during construction because we then reduced its length using the walk reduction threshold. In the next theorem, we do a "hands-on" construction of the realizer. The resulting bound is often worse than that of Theorem 4.22, but one can construct examples where it is better.

Theorem 4.23. *Let A be an irreducible $n \times n$ max-plus matrix, let G_0 be a super-digraph of the critical digraph $G_c(A)$, and let B be defined as in* (4.1)*. If $\lambda(B) = -\infty$, then $T_{\mathrm{crit}}^{G_0}(A) \leq \mathrm{cdd}(G(B)) + 1 \leq n - 1$. Otherwise,*

$$T_{\mathrm{crit}}^{G_0}(A) \leq \|A\| \cdot \frac{\mathrm{cdd}(G(B)) + \mathrm{diam}(G(A)) + g(G_c(A)) + \mathrm{ind}(G(A))}{\lambda(A) - \lambda(B)} \ . \tag{4.39}$$

In particular, $T_{\mathrm{crit}}^{G_0}(A) \leq n^2 \cdot \|A\| / (\lambda(A) - \lambda(B))$.

Proof. Note that the terms in the theorem statement remain unchanged when passing to the normalization of A. We hence assume that A is normalized.

Denote by $A_{\max}$ and $A_{\min}$ the maximum and minimum finite entries of A, respectively. By definition, $\|A\| = A_{\max} - A_{\min}$. Set $G = G(A)$, $G_B = G(B)$, and $G_c = G_c(A)$.

If $\lambda(B) = -\infty$, then $G(B)$ does not contain an cycles and hence the maximum length of walks in $G(B)$ is $\mathrm{cdd}(G(B))$, which is at most $n - |V(G_0)| - 1 \leq n - 2$. In summary, every walk in $G(A)$ of length at least $\mathrm{cdd}(G(B)) + 1 \leq n - 1$ is not in $G(B)$ and thus contains a node of G.

In the rest of the proof, we hence consider the case $\lambda(B) \neq -\infty$. The normalization assumption $\lambda(A) = 0$ implies $A_{\min} \leq \lambda(B) < 0 \leq A_{\max}$.

Denote by T the right-hand side of (4.39). Assume by contradiction that there exist i, j and a t such that all $\mathbf{N}_{\geq T}^{(\pi,t)}$-realizers for (i, j) are walks in $G(B)$ and that there is one.

Let V be such a realizer. By reducing V to a path we see that it has weight at most

$$A(V) \leq T \cdot \lambda(B) + \|A\| \cdot \mathrm{cdd}(G_B) \ . \tag{4.40}$$

We now construct a walk W in $G(A)$ from i and j of length $\ell(V)$ and show $A(W) > A(V)$. This then shows the theorem in the case $\lambda(B) \neq -\infty$. Let Z be a shortest critical cycle. By the normalization assumption, $A(Z) = 0$. Further let k be the start node of Z and let W_1 be a shortest path from i to k. Because W_1 is a shortest path, $\ell(W_1) \leq \mathrm{diam}(G)$. Set $r = \left\lfloor \left(\ell(V) - \ell(W_1) - \mathrm{ind}(G)\right) / \ell(Z) \right\rfloor$. We have

$$\begin{aligned} \ell(V) - \ell(W_1 \cdot Z^r) &= \ell(V) - \ell(W_1) - r \cdot \ell(c) \\ &\leq \mathrm{ind}(G(A)) + \ell(c) - 1 \\ &\leq \mathrm{ind}(G) + g(G_c) - 1 \end{aligned} \tag{4.41}$$

since Z a shortest cycle in G_c.

Because also $\ell(V) - \ell(W_1 \cdot Z^r) \geq \mathrm{ind}(G)$, there exists a walk W_2 whose length is equal to $\ell(V) - \ell(W_1 \cdot Z^r)$ from k to j: Denote by γ the cyclicity of G. By the definition of $\mathrm{ind}(G)$, there exists a walk W_2 from k to j of such that $\ell(V) - \ell(W_1 \cdot Z^r \cdot W_2)$ is between 0 and $\gamma - 1$. But then, because both V and $W_1 \cdot Z^r \cdot W_2$ share their start and end nodes, $\ell(V) \equiv \ell(W_1 \cdot Z^r \cdot W_2) \pmod{\gamma}$, hence $\ell(V) - \ell(W_1 \cdot Z^r \cdot W_2)$ is zero.

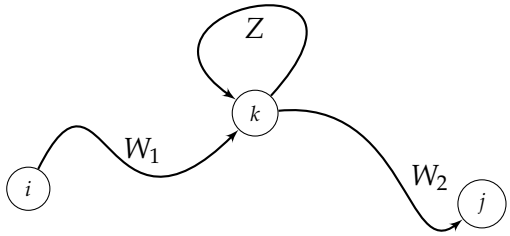


Figure 4.7: Walk W in proof of Theorem 4.23

Define the walk $W = W_1 \cdot Z^r \cdot W_2$, which is depicted in Figure 4.7. Its weight satisfies

$$\begin{aligned} A(W) &= A(W_1) + A(W_2) \geq A_{\min} \cdot \big(\ell(W_1) + \ell(W_2)\big) \\ &\geq -\|A\| \cdot \big(\operatorname{diam}(G) + g(G_c) + \operatorname{ind}(G)\big) \ . \end{aligned} \tag{4.42}$$

Together with (4.40), the contradictory assumption $A(W) < A(V)$ yields

$$T < \|A\| \cdot \frac{\operatorname{cdd}(G_B) + \operatorname{diam}(G) + g(G_c) + \operatorname{ind}(G)}{-\lambda(B)} \ , \tag{4.43}$$

a contradiction to the definition of T. This concludes the proof of (4.39).

It remains to show that $\operatorname{cdd}(G_B) + \operatorname{diam}(G) + g(G_c) + \operatorname{ind}(G) \leq n^2$. We have already seen that $\operatorname{cdd}(G_B) \leq n - |V(G_0)| - 1$. Trivially, $\operatorname{diam}(G) \leq n - 1$. Moreover, because G_c is a sub-digraph of G_0, we have $g(G_c) \leq |V(G_0)|$. If we upper-bound $\operatorname{ind}(G)$ with Theorem 2.9, we deduce $\operatorname{ind}(G) \leq n^2 - 2n + 2$. Putting all these estimates together, we have

$$\begin{aligned} &\operatorname{cdd}(G_B) + \operatorname{diam}(G) + g(G_c) + \operatorname{ind}(G) \\ \leq &n - |V(G_0)| - 1 + n - 1 + |V(G_0)| + n^2 - 2n + 2 = n^2 \ . \end{aligned} \tag{4.44}$$

This concludes the proof. □

4.6.2 Hartmann-Arguelles Scheme

Lemma 4.24. *We have $T^H_{\text{crit}} \leq \max\left\{T^{G_{\text{ha}}}_{\text{crit}}\ ,\ T_{\text{rp}}(G,H)\right\}$ for every representing sub-digraph H of G_c where $T_{\text{rp}}(G,H)$ is the maximum of values*

$$\tilde{T}^{d_k,k}_{\text{red}}(G,H) + d_k \cdot \left\lceil \frac{\operatorname{ind}(H_k)}{d_k} \right\rceil - d_k + 1 \tag{4.45}$$

where k is a node of H, H_k is k's strongly connected component in H, and d_k is the cyclicity of H_k.

Proof. We can assume without loss of generality that A is normalized and visualized because the set realizers does not change when passing to the visualization. Denote by T the right-hand side of the claimed inequality.

Let i and j be two nodes of $G(A)$, $t \in \mathbb{N}_0$, and let V be a $\mathbf{N}^{(\pi,t)}_{\geq T}$-realizer for (i,j). Then, because $T \geq T^{G_{\text{ha}}}_{\text{crit}}$, there exists a $\mathbf{N}^{(\pi,t)}_{\geq T}$-realizer W_0 for (i,j) that contains a node of G_{ha}.

Denote the maximum weight of edges in W_0 by $\mu(W_0)$ and define the digraph

$$\tilde{G} = \begin{cases} G_{\mathrm{ha}} & \text{if } \mu(W_0) \le \mu^{ha} \ , \\ T^{\mathrm{ha}}(\mu(W_0)) & \text{otherwise} \ . \end{cases} \tag{4.46}$$

By the definition of Hartmann-Arguelles threshold digraphs, $G_c \subseteq \tilde{G} \subseteq G_{\mathrm{ha}}$. In both cases of (4.46), walk W_0 contains a node l of digraph $\tilde{G}$, which is completely reducible due to the assumed max-balancing of A.

Let $W_0 = W_1 \cdot W_2$ with W_1 ending at node l. By definition of $\tilde{G}$, there exists a critical node k of H in the same strongly connected component of $\tilde{G}$ as k. Let V_1 be a walk in $\tilde{G}$ from l to k and V_2 be a walk in $\tilde{G}$ from k to l. Set $V = V_1V_2$ and $W_3 = W_1 \cdot V^\pi \cdot W_2$. We have $\ell(W_3) \equiv t \pmod{\pi}$.

By definition of the walk reduction threshold, there exists a walk $\hat{W}_3 \in \mathcal{W}(i \xrightarrow{k} j)$ obtained from W_3 by removing cycles and possibly inserting cycles in H such that $\ell\left(\hat{W}_3\right) \le T_{\mathrm{red}}^{d_k,k}(G,H)$ and $\ell\left(\hat{W}_3\right) \equiv \ell(W_3) \equiv t \pmod{d_k}$. By Lemma 4.12, we have $t - \ell\left(\hat{W}_3\right) \ge \mathrm{ind}(H_k)$, and there hence exists a critical closed walk in H at node k whose addition to $\hat{W}_3$ yields a walk W with $\ell(W) = t$.

Since A is max-balanced and $\lambda(A) = 0$, all edges have nonpositive weights, and the weight of each edge of $\tilde{G}$ is not smaller than that of any edge of W. Each edge of W is either removed, kept, or replaced by an edge of $\tilde{G}$ in $\tilde{W}$, thus we conclude that $A(W) \ge A(W_0)$. Hence W is also a $\mathbf{N}_{\ge T}^{(\pi,t)}$-realizer, but one that includes a node of H. □

4.6.3 Cycle Threshold Scheme

A finite sequence of cycles $Z_1, \ldots, Z_m$ in G_0 is called a *staircase* in G_0 if, for all $1 \le r \le m-1$, Z_r and Z_{r+1} share a node, $A(Z_r)/\ell(Z_r) \le A(Z_{r+1})/\ell(Z_{r+1})$ and, moreover, the cycle mean of Z_{r+1} is the greatest among all the cycles sharing a node with Z_r.

Lemma 4.25. *Let $\mu > \mu^{ct}$ and Z be a cycle in $T^{\mathrm{ct}}(\mu)$ or $\mu = \mu^{ct}$ and Z be a cycle in $G_{\mathrm{ct}}(\mu)$ with $A(Z)/\ell(Z) = \mu$. Then there exists a staircase $Z_1, \ldots, Z_m$ in $T^{\mathrm{ct}}(\mu)$ such that $Z_1 = Z$ and Z_m is critical.*

Proof. Suppose by contradiction that no such staircase exists. Let $Z_1, \ldots, Z_m$ be a staircase in $T^{\mathrm{ct}}(\mu)$ such that $Z_1 = Z$ and $A(Z_m)/\ell(Z_m)$ is maximal.

Denote $\mu' = A(Z_m)/\ell(Z_m)$, so $\mu' < \lambda(A)$. If the strongly connected component of $T^{\mathrm{ct}}(\mu')$, in which Z_m lies, contains a cycle of mean weight strictly greater than μ', then we can build a staircase with a greater cycle mean of the final cycle, a contradiction. So that component of $T^{\mathrm{ct}}(\mu')$ does not contain a cycle of mean weight strictly greater than μ', which is a contradiction to the definition of μ^{ct} and the fact that $\mu' \ge \mu^{ct}$. Thus we must have $\mu' = \lambda(A)$. □

Lemma 4.26. *We have $T_{\mathrm{crit}}^{G_c} \le \max\left\{T_{\mathrm{crit}}^{G_{\mathrm{ct}}},\ T'_{\mathrm{rp}}\right\}$ where T'_{rp} is the maximum of the values*

$$\tilde{T}_{\mathrm{red}}^{\ell(Z),k} + 1 \tag{4.47}$$

where Z is a cycle of $G(A)$ and k is a node of Z.

Proof. We assume without loss of generality that A is normalized. Denote by T the right-hand side of the claimed inequality.

Let i and j be two nodes of $G(A)$, $t \geq T$, and let V be a $\mathbf{N}_{\geq T}^{(\pi,t)}$-realizer for (i,j). Then, because $T \geq T_{\text{crit}}^{G_{\text{ct}}}$, there exists a $\mathbf{N}_{\geq T}^{(\pi,t)}$-realizer W_0 for (i,j) that contains a node of G_{ct}.

Denote by $\nu(W)$ the largest cycle mean of cycles in the walk's multigraph $M(W)$. We assume in the following that $\nu(W)$ is maximal among all $W \in \mathcal{W}^t(i \to j)$ with $A(W) = A_{i,j}^{\otimes t}$. We prove the lemma by showing $\nu(W) = 0$. Assume by contradiction that $\nu(W) < 0$, and define

$$\tilde{G} = \begin{cases} G_{\text{ct}} & \text{if } \nu(W) \leq \mu^{ct} \ , \\ T^{\text{ct}}(\nu(W)) & \text{otherwise} \ . \end{cases} \tag{4.48}$$

By the definition of cycle threshold digraphs, $G_c \subseteq \tilde{G} \subseteq G_{\text{ct}}$.

By Lemma 4.25, there exists a staircase $Z_1, \ldots, Z_m$ in $\tilde{G}$ such that Z_1 has $A(Z_1) = \nu(W)$ and shares a node k_1 with W, and Z_m is critical. We inductively define walks $W_1, \ldots, W_m$ as follows: Set $W_1 = W$. For every $2 \leq r \leq m$, let k_r be a node in both Z_{r-1} and Z_r. These exist by definition of a staircase.

Let $1 \leq r < m$ and assume that W_r is already defined. There is a walk $\hat{W}_r \in \mathcal{W}(i \xrightarrow{k_r} j)$ with $\ell\left(\hat{W}_r\right) \equiv t \pmod{\ell(Z_r)}$, obtained from W_r by removing cycles and inserting copies of Z_r such that

$$\ell\left(\hat{W}_r\right) \leq \tilde{T}_{\text{red}}^{\ell(Z_r),k_r}(G, Z_r) \leq t - 1 \ . \tag{4.49}$$

We hence have $t - \ell\left(\hat{W}_r\right) > 0$. Thus, the number $\tau = \left(t - \ell\left(\hat{W}_r\right)\right)/\ell(Z_r)$ is a positive integer. Now define W_{r+1} as walk $\hat{W}_r$ after inserting τ copies of Z_r, to have $\ell(W_{r+1}) = t$. Thus Z_r is a subwalk of W_{r+1} and hence contains node k_{r+1}. The walk W_m contains a critical node.

We now show that $A(W_{r+1}) \geq A(W_r)$. For this we will prove by induction that, for all $1 \leq r \leq m-1$, the mean weight of Z_r is not less than that of any cycle of $M(W_r)$. This is true for $r = 1$ by definition of $\tilde{G}$. Observe that the cycles of $M(W_{r+1})$ are (1) Z_r and cycles using the edges of Z_r, (2) cycles that were already in $M(W_r)$. For the latter cycles we use the inductive assumption, while the cycles using edges of Z_r share a common node with it and hence their mean weight does not exceed that of Z_{r+1} by the definition of a staircase.

Setting $\tilde{W} = W_m$ we obtain $\tilde{W} \in \mathcal{W}^t(i \xrightarrow{G_c} j)$ and $A(\tilde{W}) \geq A(W)$, i.e., $\tilde{W}$ is an $\mathbf{N}_{\geq T}^{(\pi,t)}$-realizer for (i,j). □

Lemma 4.27. *We have* $T_{\text{crit}}^{G_c} \leq \max\left\{T_{\text{crit}}^{G_{\text{ct}}}, \text{Wi}(n)\right\}$ *where* n *is the number of nodes of* $G(A)$.

Proof. We use the same proof as for Lemma 4.26, except that we specifically use the cycle decomposition walk reduction in the definition of $\hat{W}_r$ and that we demand that Z_m be the first critical cycle in the staircase.

The latter assumption guarantees the number n_{W_r} of distinct nodes of W_r is not n, that is, at most $n-1$, for all $1 \leq r \leq m-1$. Plugging this estimate into inequality (4.33) and using $\ell(Z_r) \leq n$, we get that the length of the reduced walks $\hat{W}_r$ is at most

$$\ell\left(\hat{W}_r\right) \leq (n-1)(n-2) + n - 1 = n^2 - 2n + 1 \ , \tag{4.50}$$

which is at most $t - 1$ because $t \geq \text{Wi}(n) = n^2 - 2n + 2$.

The rest of the proof is the same as that of Lemma 4.26. □

4.7 Putting it Together

Proof of Theorem 4.1. We use the proof strategy in Section 4.3. For all four claimed bounds, we list our parameter choices and the walk reductions used. We use Theorem 4.22 with the Hartmann-Arguelles scheme $G_0 = G_{\text{ha}}$ for all four bounds. Note that $\operatorname{cdd}(G(B_{\text{ha}}))$ is lower than the first term in the maximum of all four claimed bounds.

For bounding the walk reduction threshold for the criticality threshold in Theorem 4.22, the bound, B_T, is larger or equal to $\operatorname{cdd}(B_G)$ in all four claimed bounds. We use this fact to bound the second term, R, in the maximum of the claimed bounds by

$$\begin{aligned} R &= \frac{B_T(\lambda(A) - A_{\min}) + \operatorname{cdd}(G_B)(B_{\max} - \lambda(B))}{\lambda(A) - \lambda(B)} \\ &= \frac{(B_T - \operatorname{cdd}(G_B))(\lambda(A) - A_{\min}) + \operatorname{cdd}(G_B)(\lambda(A) - A_{\min} + B_{\max} - \lambda(B))}{\lambda(A) - \lambda(B)} \\ &= \frac{(B_T - \operatorname{cdd}(G_B))(\lambda(A) - A_{\min}) + \operatorname{cdd}(G_B)(B_{\max} - A_{\min})}{\lambda(A) - \lambda(B)} + \operatorname{cdd}(G_B) \\ &\leq \frac{B_T \cdot \|A\|}{\lambda(A) - \lambda(B)} + \operatorname{cdd}(G_B) \end{aligned} \tag{4.51}$$

because $\lambda(A), B_{\max} \leq A_{\max}$ and $B_T \geq \operatorname{cdd}(G_B)$.

For (4.2), we choose $H = G_c(A)$ for bounding the walk reduction threshold in the critical bound Theorem 4.22, and choose H repetitively and use Theorem 4.18 for bounding the walk reduction threshold for the reduction bound $B_{\text{red}}^{H,k}$ and in Lemma 4.24.

For (4.3), we choose $H = G_c(A)$ for bounding the walk reduction threshold in the critical bound Theorem 4.22, and choose H repetitively and use Theorem 4.15 for bounding the walk reduction threshold for the reduction bound $B_{\text{red}}^{H,k}$ and in Lemma 4.24.

For (4.4), we choose $H = G_c(A)$ for bounding the walk reduction threshold in the critical bound Theorem 4.22, and choose $H = G_c(A)$ exploratively and use Theorem 4.18 for bounding the walk reduction threshold for the reduction bound $B_{\text{red}}^{H,k}$ and in Lemma 4.24.

For (4.5), we choose $H = G_c(A)$ for bounding the walk reduction threshold in the critical bound Theorem 4.22, and choose $H = G_c(A)$ exploratively and use Theorem 4.15 for bounding the walk reduction threshold for the reduction bound $B_{\text{red}}^{H,k}$ and in Lemma 4.24. □

Proof of Theorem 4.4. Using the proof strategy laid out in Section 4.3, we choose for H one critical cycle in every critical component and for B_{crit}^{H} the maximum of $\operatorname{Wi}(n)$ and the second term in the maximum in theorem statement. If H does not contain a Hamiltonian cycle, i.e., all its cycles have length at most $n-1$, then we choose $B_{\text{red}}^{H,k} = (d_k + 1)(n-1)$ and $B_{\text{pump}}^{H,k} = -d_k + 1$ where d_k is the length of the critical cycle in H on which k is included. This choice is in accordance with Theorem 4.18 and we have $B_{\text{red}}^{H,k} + B_{\text{pump}}^{H,k} = (d_k+1)(n-2)+2$, which is at most $\operatorname{Wi}(n)$ because $d_k \leq n-1$. If H does contain a Hamiltonian cycle, then it is induced by one and we choose $B_{\text{red}}^{H,k} = n^2 - n + 1$ and $B_{\text{pump}}^{H,k} = -n + 1$ in accordance with Theorem 4.20. In this case, we have $B_{\text{red}}^{H,k} + B_{\text{pump}}^{H,k} = \operatorname{Wi}(n)$.

By noting that, in any case, we have $\tilde{T}_{\text{red}}(G, H) \leq n^2 - n + 1$ concludes the proof with Theorem 4.22 and Lemma 4.27. □

Lemma 4.28. *Let A be a max-plus matrix with $\lambda(A) = 0$. Then, for every $t \in \mathbb{N}_0$, there exists a walk W in $G(A)$ with $\ell(W) = t$ and $A(W) \geq 0$.*

Proof. Let Z be a critical cycle in $G(A)$. By forming blocks of length t, we decompose $Z^t = W_1 \cdot W_2 \cdots W_{\ell(Z)}$ with $\ell(W_k) = t$ for all k. Because $0 = A(Z^t) = \sum_{k=1}^{\ell(Z)} A(W_k)$, there is at least one W_k of nonnegative A-weight. □

Theorem 4.29. *Let A be an $n \times n$ max-plus matrix with all finite entries, i.e., $A \in \mathbb{R}^{n\times n}$. Further, let G_0 be a super-digraph of the critical digraph $G_c(A)$ and let B be defined as in* (4.1)*. Then,*

$$T_{\text{crit}}^{G_0}(A) \leq \frac{2\|A\|}{\lambda(A) - \lambda(B)} + \text{cdd}(G(B)) \ . \tag{4.52}$$

Proof. Because neither the critical threshold nor the bound changes, we assume A to be normalized. This implies that $\lambda(A) = 0 > \lambda(B) \geq A_{\min}$.

Denote by T the right-hand side of (4.52). Assume by contradiction that there exist i, j and a t such that all $\mathbf{N}_{\geq T}^{(\pi,t)}$-realizers for (i, j) are walks in $G(B)$ and that there is one.

Let V be such a realizer. Denoting by $\hat{V}$ any reduction of V to a path by removing cycles, we have

$$\begin{aligned} A(V) &\leq \ell(V) \cdot \lambda(B) - \ell\left(\hat{V}\right) \cdot \lambda(B) + A(\hat{V}) \\ &\leq T \cdot \lambda(B) - \text{cdd}(G(B)) \cdot \lambda(B) + A(\hat{V}) \ . \end{aligned} \tag{4.53}$$

On the other hand, we construct a walk W from i to j of length in $\mathbf{N}_{\geq T}^{(\pi,t)}$ as follows: Since $\ell(V) \geq T \geq 2 + \text{cdd}(G(B))$, the integer $s = \ell(V) - \ell\left(\hat{V}\right) - 2$ is nonnegative. By Lemma 4.28, there is a walk W_1 with $\ell(W_1) = s$ and $A(W_1) \geq 0$. Denote by k the start node of W_1 and by l its end node. Since the digraph $G(A)$ is complete, the weight of walk $W = \hat{V} \cdot (j, k) \cdot W_1 \cdot (l, j)$ satisfies

$$A(W) \geq A(\hat{V}) + 2A_{\min} \tag{4.54}$$

and its length is equal to $\ell(V)$.

By assumption, $A(W) < A(V)$, which, in combination with (4.53) and (4.54), means

$$T < \frac{-2A_{\min}}{-\lambda(B)} + \text{cdd}(G(B)) \leq \frac{2\|A\|}{-\lambda(B)} + \text{cdd}(G(B)) \ , \tag{4.55}$$

a contradiction to the choice of T. □

4.8 Linear Systems

In this section, we study transients of max-plus linear systems, i.e., the sequence of vectors $A^{\otimes t} \otimes v$ for an irreducible system matrix A and an initial vector v. Because the sequence $A^{\otimes t}$ is eventually periodic if A is irreducible, so is the linear system and the system's transient is upper bounded by that of the matrix. And the system's transient can really be lower, depending on the initial vector v. In this section, we adapt the proof technique of Section 4.3 to linear systems whose initial vector's entries are all finite. The reason for this is that the matrix's transient is always attained with an initial vector with $-\infty$ entries, so the analysis would lead to the same bounds as those for matrices. However, with all finite initial vectors, one can prove a sharper critical bound B_{crit}^{H} that includes $\|v\|$, the maximum difference of entries of v. We discuss the relationship between matrix and system transients in more detail in Section 4.9.

4.8.1 Modified Proof Strategy

The proof strategy to prove transience bounds for systems is almost the same as that for matrices. The only difference is an adapted definition of a realizer and a different critical bound. A different definition of realizers is need because the graph correspondence for entries of the system's vectors is different from that of matrices: Setting $x(t) = A^{\otimes t} \otimes v$, we have

$$x_i(t) = \max_{j\in[n]} \left(A^{\otimes}_{i,j} + v_j\right) = \max\left\{A(W) + v_j \mid W \in \mathcal{W}^t(i \to) \text{ and } j = \mathrm{End}(W)\right\} \ , \tag{4.56}$$

where $\mathcal{W}^t(i \to)$ denotes the set of walks of length t starting at node i and $\mathrm{End}(W)$ denotes the end node of walk W. This equality leads us to not use the A-weight when dealing with linear systems, but to adapt the definition of the weight, taking into account the initial vector. We hence define $A_v(W) = A(W) + v_j$ where $j = \mathrm{End}(W)$ is the end node of W and call it the *A_v-weight* of W. We also use the notation $\mathcal{W}(i \to)$ for all walks starting at node i. With this adapted walk weight definition, (4.56) transforms into

$$x_i(t) = (A^{\otimes} \otimes v)_i = \max\left\{A_v(W) \mid W \in \mathcal{W}^t(i \to)\right\} \ . \tag{4.57}$$

We write $T_v(A)$ for the transient of the linear system $A^{\otimes t} \otimes v$.

To repeat our analysis of linear systems analogously to matrices, we need to define an adapted version of a realizer: An $\mathbf{N}$-*realizer* for i is a maximum A_v-weight walk in $\mathcal{W}(i \to)$ with length in $\mathbf{N}$.

However, we do not need to re-prove all the lemmas and theorems we used in the matrix case. This is due to the fact that we can reuse many results because of the following correspondence between realizers for one node i and realizers for a pair of nodes (i, j):

Lemma 4.30. *Let A be an $n \times n$ max-plus matrix and v a vector in $\mathbb{R}^n$. If $W \in \mathcal{W}(i \to j)$ is an $\mathbf{N}$-realizer for i, then it is also an $\mathbf{N}$-realizer for (i, j).*

Moreover, if there is an $\mathbf{N}$-realizer for i that ends in j, then every $\mathbf{N}$-realizer for (i, j) is an $\mathbf{N}$-realizer for i.

Proof. Being an $\mathbf{N}$-realizer for i means that $A_v(W) = A(W) + v_j \geq A_v(W')$ for all walks $W' \in \mathcal{W}(i \to)$ with $\ell(W') \in \mathbf{N}$. Because $\mathcal{W}(i \to j) \subseteq \mathcal{W}(i \to)$, the inequality is in particular true for all $W' \in \mathcal{W}(i \to j)$ with $\ell(W') \in \mathbf{N}$, for which $A_v(W') = A(W') + v_j$. Subtracting v_j from both sides gives $A(W) \geq A(W')$, which shows that W is an $\mathbf{N}$-realizer for (i, j).

On the other hand, if W_0 is a $\mathbf{N}$-realizer for (i, j), then $A(W_0) \geq A(W')$ for all $W' \in \mathcal{W}(i \to j)$ with $\ell(W') \in \mathbf{N}$. In particular, for $W' = W$, we have $A(W_0) \geq A(W)$. Adding v_j to both sides gives $A_v(W_0) \geq A_v(W)$ and hence W_0 is a $\mathbf{N}$-realizer for i because W is. □

An analog of Theorem 4.11 for proving bounds on the transient via realizers exists and is proved very similarly:

Theorem 4.31. *Let A be an $n \times n$ max-plus matrix, let v be a vector in $\mathbb{R}^n_{\max}$, and let $i \in [n]$. If B and π are such that for every integer $t \geq B$ either*

- *there exists an $\mathbf{N}^{(t,\pi)}_{\geq B}$-realizer for i of length t, or*
- *there exists no $\mathbf{N}^{(t,\pi)}_{\geq B}$-realizer for i at all,*

then the sequence $(A^{\otimes t} \otimes v)_i$ is eventually periodic with period π and transient at most B.

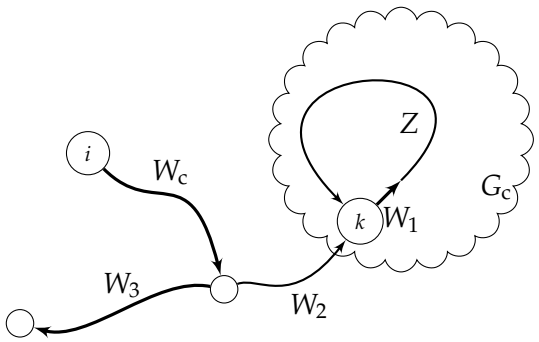


Figure 4.8: Walk W in proof of Theorem 4.32

4.8.2 Critical Bounds for Systems

The following theorem gives a way to choose B^H_{crit} for step 2 of the proof strategy for linear systems. The rest of the proof stays the same, noting Lemma 4.30.

Theorem 4.32. *Let A be an irreducible $n \times n$ max-plus matrix, let G_0 be a super-digraph of the critical digraph $G_c(A)$, and let B be defined as in* (4.1)*. Further, let*

$$T \geq \max\left\{ n-1\,,\ \frac{\|v\| + \|A\| \cdot (n-1)}{\lambda(A) - \lambda(B)} \right\}\ , \tag{4.58}$$

$t \in \mathbb{N}_0$, and $i \in [n]$. there is a $\mathbf{N}^{(t,\pi)}_{\geq T}$-realizer for i that includes a node of G_0.

Proof. Because neither the critical threshold nor the bound changes, we assume A to be normalized and $\lambda(B)$ to be finite because the other case is easy. This implies that $A_{\max} \geq \lambda(A) = 0 > \lambda(B) \geq A_{\min}$. Set $G_A = G(A)$, $G_B = G(B)$, and $G_c = G_c(A)$.

We proceed by contradiction: Suppose that no $\mathbf{N}^{(t,\pi)}_{\geq T}$-realizer for i contains a node of G_0. Let W_0 be such a walk. Let $\hat{W}_0$ be W_0 after removing all subcycles in any order. This is a path.

Next choose a critical node k, and then a prefix W_c of $\hat{W}_0$, such that the distance between k and the end node of W_c is minimal. Let W_2 be a path of minimal length from the end node of W_c to k. Let W_3 be the walk such that $\hat{W}_0 = W_c \cdot W_3$. Further let Z be a critical cycle starting at k.

We distinguish two cases for $\ell(W_0)$, namely (a) $\ell(W_0) \geq \ell(W_c) + \ell(W_2)$, and (b) $\ell(W_0) < \ell(W_c) + \ell(W_2)$.

Case a: Let $m \in \mathbb{N}_0$ be the quotient in the Euclidean division of $\ell(W_0) - \ell(W_c) - \ell(W_2)$ by $\ell(Z)$, and choose W_1 to be a prefix of Z of length $\ell(W_0) - \big(\ell(W_c) + \ell(W_2) + m \cdot \ell(Z)\big)$ (see Figure 4.8). Clearly W_1 starts at k. If we set $W = W_c \cdot W_2 \cdot Z^m \cdot W_1$, we get $\ell(W) = \ell(W_0)$ and

$$A_v(W) \geq \min_{1 \leq j \leq n}(v_j) + A(W_c) + A(W_2) + A(W_1) \tag{4.59}$$

since we assume $\lambda = 0$.

For the A_v-weight of W_0, we have

$$\begin{aligned} A_v(W_0) &\leq A_v(\hat{W}_0) + \lambda(B) \cdot \big(\ell(W_0) - \ell(\hat{W}_0)\big) \\ &\leq \max_{1 \leq j \leq n}(v_j) + A(\hat{W}_0) + \lambda(B) \cdot \big(\ell(W_0) - \ell(\hat{W}_0)\big) \end{aligned} \tag{4.60}$$

By assumption $A_v(\hat{W}) > A_v(W)$, and from (4.59), (4.60), and $\lambda(B) < 0$ we therefore obtain

$$\begin{aligned}\ell(W_0) &< \frac{\|v\| + A(W_3) - A(W_1) - A(W_2)}{-\lambda(B)} + \ell(\hat{W}_0) \\ &\leq \frac{\|v\| + A_{\max}\,\ell(W_3) - A_{\min}\,(\ell(W_1) + \ell(W_2))}{-\lambda(B)} + \ell(\hat{W}_0) \\ &\leq \frac{\|v\| + A_{\max}\,\ell(W_3) - A_{\min}\,(\ell(W_1) + \ell(W_2) + \ell(\hat{W}_0))}{-\lambda(B)}\end{aligned} \tag{4.61}$$

Denote by n_0 the number of nodes in G_0. The following three inequalities hold: $\ell(W_3) \leq n - n_0 - 1$, $\lambda(B) \geq A_{\min}$, and $\ell(W_1) < \ell\,(Z) \leq n_c$. Moreover from the minimality constraint for the length of W_2 follows that $\ell(W_2) + \ell(\hat{W}_0) \leq n - n_c$. Thereby

$$\ell(W_0) < \frac{\|v\| + \|A\| \cdot (n-1)}{-\lambda(B)}\ , \tag{4.62}$$

a contradiction to $\ell\,(W_0) \geq T$ and the lemma follows in case (a).

Case b: In this case $\ell(W_c) \leq n < \ell(W_c) + \ell(W_2)$, and we set $W = W_c \cdot W_2'$, where W_2' is a prefix of W_2, such that $\ell(W) = \ell(W_0)$. Hence,

$$A_v(W) \geq \min_{1 \leq j \leq n}(v_j) + A(W_c) + A(W_2')\ . \tag{4.63}$$

We again obtain (4.60). By assumption $A_v(\hat{W}) > A_v(W)$, and by similar arguments as in case a we derive

$$\ell(\hat{W}) \leq \frac{\|v\| + A(W_3) - A(W_2')}{-\lambda(B)} + \ell(W_0) \tag{4.64}$$

and since W_2' is a prefix of W_2 with $\ell(W_2') < \ell(W_2)$,

$$\ell(\hat{W}) < \frac{\|v\| + A_{\max}\,\ell(W_3) - A_{\min}\,\ell(W_2)}{-\lambda(B)} + \ell(W_0)\ , \tag{4.65}$$

which is less or equal to the bound obtained in (4.61) of case (a). By similar arguments as in case (a), the lemma follows also in case (b). □

4.9 Matrix vs. System Transients

By letting one component of the initial vector v tend to infinity, one can see that the transient $T(A)$ of a max-plus matrix A is equal to the transient $T_v(A)$ of some linear system with system matrix A. However, in this argument, the value $\|v\|$ also tends to infinity, which categorically prohibits the use of our bounds for linear systems to bound the transients of matrices. In this section, we give an alternative argument that theoretically permits this use. While it may not provide better bounds for matrices than that of Section 4.7, it provides further insights on the relationship between the transients of matrices and systems. We show that the transient of matrix A is actually equal to the transient of a specific linear system with matrix A and initial vector v with $\|v\|$ is in $O\big(\|A\| \cdot n^2\big)$, provided the system transient is sufficiently large, namely at most equal to some term quadratic in n.

Obviously, $T(A)$ is an upper bound on the $T_v(A)$'s. Conversely, the equalities $A^{\otimes t}_{i,j} = \left(A^{\otimes t} \otimes e^j\right)_i$, where the e^j's are the unit vectors defined by $e^j_i = 0$ if $i = j$ and $e^j_i = -\infty$ otherwise, show that $\max\left\{T_{e^j}(A) \mid j \in [n]\right\} \geq T(A)$. Hence,

$$\sup\left\{T_v(A) \mid v \in \mathbb{R}^n_{\max}\right\} = T(A) \ . \tag{4.66}$$

We now seek a similar expression of $T(A)$, but with *finite* initial vectors v, i.e., with $v \in \mathbb{R}^n$. We define:

$$\tilde{B} = 2(n-1) + \hat{\text{ind}} + (\text{ind}(G) + \hat{\gamma} - 1) \tag{4.67}$$

$$\mu = \sup\left\{A^{\otimes t}_{i,h} - A^{\otimes t}_{i,j} \mid h, i, j \text{ nodes of } G \ , \ t \geq \tilde{B} \ , \ A^{\otimes t}_{i,j} \neq -\infty\right\} \tag{4.68}$$

Clearly μ is finite, i.e., $\mu \in \mathbb{R}$. Then we consider the *μ-truncated* unit vectors obtained by replacing the infinite entries of the e^j's by $-\mu$.

In Theorem 4.35 below, we show that if $B \geq \tilde{B}$ and B is a bound on the system transients for all μ-truncated unit vectors, then B is also a bound on the matrix transient. A technical difficulty in the proof lies in the fact that, contrary to the sets $\mathcal{W}^t(i \to)$ which occur in the expression of the i^{th} component of linear systems, the sets $\mathcal{W}^t(i \to j)$ that we consider for matrix powers may be empty. The next two lemmas deal with this technicality.

Lemma 4.33. *For any pair of nodes i, j of G and any integer $t \geq \text{ind}(G) + \gamma(G) + n - 2$, there exists a walk W from i to j such that $t - \ell(W) \in \{0, \ldots, \gamma(G) - 1\}$.*

Proof. Let i, j be any two nodes, and let W_0 be a path from i to j. For any integer t, consider the residue r of $t - \ell(W_0)$ modulo $\gamma(G)$. By definition of $\text{ind}(G)$, if $t - \ell(W_0) - r \geq \text{ind}(G)$, then there exists a closed walk C starting at node j with length equal to $t - \ell(W_0) - r$. Then, $W_0 \cdot C$ is a walk from i to j with length $t - r$, where $r \in \{0, \ldots, \gamma(G) - 1\}$. The lemma follows since $t - \ell(W_0) - r \geq \text{ind}(G)$ as soon as $t \geq \text{ind}(G) + (n-1) + \gamma(G) - 1$. □

Lemma 4.34. *Let $t \geq \text{ind}(G) + \gamma(G) + n - 2$. Then $A^{\otimes t+\gamma(G)}_{i,j} = -\infty$ if and only if $A^{\otimes t}_{i,j} = -\infty$.*

Proof. It is equivalent to claim that $\mathcal{W}^{t+\gamma(G)}(i \to j) = \emptyset$ if and only if $\mathcal{W}^t(i \to j) = \emptyset$ for every integer $t \geq \text{ind}(G) + \gamma(G) + n - 2$.

Suppose $\mathcal{W}^{t+\gamma(G)}(i, j) \neq \emptyset$, and let $W_0 \in \mathcal{W}^{t+\gamma(G)}(i, j)$. By Lemma 4.33, there exists a walk $W \in \mathcal{W}(i, j)$ such that $t = \ell(W) + r$ with $r \in \{0, 1, \ldots, \gamma(G) - 1\}$. Lemma 2.4 implies that $\gamma(G)$ divides $\ell(W_0) - \ell(W) = (t + \gamma(G)) - (t - r) = \gamma(G) + r$; hence $\gamma(G)$ divides r. Therefore, $r = 0$, i.e., $\ell(W) = t$ and thus $\mathcal{W}^t(i \to j) \neq \emptyset$.

The converse implication is proved similarly. □

Theorem 4.35. *If $t \geq \tilde{B}$ and $A^{\otimes(t+\gamma)} \otimes v = A^{\otimes t} \otimes v$ for all μ-truncated unit vectors v, then $A^{\otimes(t+\gamma)} = A^{\otimes t}$.*

Proof. Let i and j be nodes in G, and let $t \geq \tilde{B}$. Further let v be the μ-truncated unit vector with $v_j = 0$ and $v_h = -\mu$ for $h \neq j$. Since $\tilde{B} \geq \text{ind}(G) + \gamma(G) + n - 2$ and $\gamma = \gamma(G_c)$ is a multiple of $\gamma(G)$, we derive from Lemma 4.34 that $A^{\otimes t+\gamma}_{i,j} = -\infty$ if and only if $A^{\otimes t}_{i,j} = -\infty$. There are two cases to consider:

1. $A^{\otimes t}_{i,j} - \infty$ and $A^{\otimes t+\gamma}_{i,j} = -\infty$. In this case, $A^{\otimes t+\gamma}_{i,j} = A^{\otimes t}_{i,j}$ trivially holds.

2. $A_{i,j}^{\otimes t} \neq -\infty$ and $A_{i,j}^{\otimes t+\gamma} \neq -\infty$. Recall that

$$\left(A^{\otimes t} \otimes v\right)_i = \max\left\{A_{i,j}^{\otimes t} + v_h \mid h \in [n]\right\} . \tag{4.69}$$

By definition of μ and v, for any node $h \neq j$,

$$A_{i,h}^{\otimes t} - A_{i,j}^{\otimes t} \leq \mu = v_j - v_h . \tag{4.70}$$

It follows that

$$\left(A^{\otimes t} \otimes v\right)_i = A_{i,j}^{\otimes t} + v_j . \tag{4.71}$$

As $t + \gamma \geq t$, we similarly have

$$A_{i,j}^{\otimes t+\gamma} = \left(A^{\otimes t+\gamma} \otimes v\right)_i - v_j = \left(A^{\otimes t} \otimes v\right)_i - v_j = A_{i,j}^{\otimes t} . \tag{4.72}$$

Thus $A_{i,j}^{\otimes t+\gamma} = A_{i,j}^{\otimes t}$ holds also in this case. □

The key point for establishing our bound on matrix transients is the following upper bound on μ, which is quadratic in n. The proof uses the pumping technique developed for the explorative bound twice.

Theorem 4.36. $\mu \leq \|A\| \cdot \tilde{B}$

Proof. First, we observe that each term in the inequality to show is invariant under substituting A by $\overline{A}$. Hence we assume that $\lambda = 0$. It follows that

$$A_{i,h}^{\otimes t} \leq A_{\max} \cdot (n-1) \leq A_{\max} \cdot \tilde{B} . \tag{4.73}$$

We now give a lower bound on $A_{i,j}^{\otimes t}$ in the case that it is finite, i.e., if $\mathcal{W}^t(i \to j) \neq \emptyset$. Let k be a critical node in the strongly connected component H of G_c with minimal distance from i and let W_1 be a shortest path from i to k. Further, let W_2 be a shortest path from k to j. Let r denote the residue of $t - \ell(W_1 \cdot W_2) - \mathrm{ind}(G)$ modulo $\gamma(H)$, and let $s = t - \ell(W_1 \cdot W_2) - \mathrm{ind}(G) - r$. Since $s \equiv 0 \pmod{\gamma(H)}$, and

$$s \geq \tilde{B} - 2(n-1) - \mathrm{ind}(G) - \left(\gamma(H) - 1\right) \geq \hat{\mathrm{ind}} \geq \mathrm{ind}(H) , \tag{4.74}$$

there exists a closed walk Z_c of length s in component H starting at node k. Let $\tau = \mathrm{ind}(G) + r$; then, $\tau \geq \mathrm{ind}(G)$. Moreover, $s = \tau - \ell(W_1 \cdot Z_c \cdot W_2)$, and $W_1 \cdot Z_c \cdot W_2 \in \mathcal{W}(i \to j)$. By Lemma 2.4, it follows that $\gamma(G)$ divides τ, because $\mathcal{W}^t(i \to j) \neq \emptyset$. Hence there exists a closed walk Z_{nc} of length τ starting at node j.

Now define $W = W_1 \cdot Z_c \cdot W_2 \cdot Z_{nc}$. Clearly, $\ell(W) = t$ and

$$A(W) \geq A_{\min} \cdot (t - s) \geq A_{\min} \cdot \left(2(n-1) + \mathrm{ind}(G) + \gamma(H) - 1\right) , \tag{4.75}$$

and so

$$A_{i,j}^{\otimes t} \geq A_{\min} \cdot \left(2(n-1) + \mathrm{ind}(G) + \hat{\gamma} - 1\right) \geq A_{\min} \cdot \tilde{B} . \tag{4.76}$$

From (4.73) and (4.76) follows $\mu \leq (A_{\max} - A_{\min}) \cdot \tilde{B} = \|A\| \cdot \tilde{B}$. □

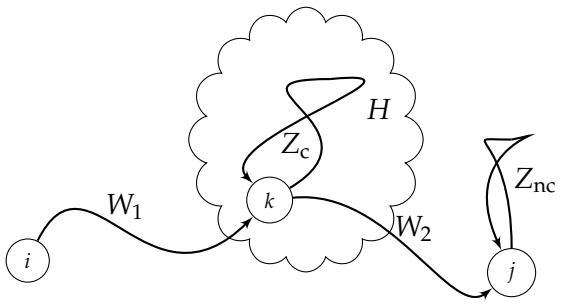


Figure 4.9: Walk W in proof of Theorem 4.36

4.10 A Closer Look at the Nachtigall Decomposition

In this section, we present a different technique and proof strategy to prove transience bounds for max-plus matrices that are almost as tight as those developed in the previous sections of this chapter. For this reason, we do not try to optimize the bounds, but rather focus on the novelty of the proof technique. We will use and improve the max-plus matrix decomposition defined by Nachtigall.

The next two lemmas play an important role in our proofs: They state transience bounds for a sequence obtained by simple composition of two eventually periodic sequences f and g with common ratio in terms of the transients of f and g. Their proofs are not hard.

Lemma 4.37. *Let $f, g : \mathbb{N}_0 \to \mathbb{R}_{\max}$ be eventually periodic with common ratio ϱ and respective transients T_f and T_g. Then the sequence $\max\{f, g\}$ is eventually periodic with ratio ϱ and transient at most $\max\{T_f, T_g\}$.*

In analogy to classical convolution, the *max-plus convolution* $f \otimes g$ of two sequences f and g is given by

$$(f \otimes g)(t) = \max_{t_1+t_2=t} \big(f(t_1) + g(t_2)\big) \ . \tag{4.77}$$

Lemma 4.38 ([72, Lemma 6.1])**.** *Let $f, g : \mathbb{N}_0 \to \mathbb{R}_{\max}$ be eventually periodic with common ratio ϱ, common period p, and respective transients T_f and T_g. Then the convolution $f \otimes g$ is eventually periodic with ratio ϱ, period p, and transient at most equal to $T_f + T_g + p - 1$.*

The next lemma characterizes the convolution in terms of walks in the matrix's digraph.

Lemma 4.39. *Let $A \in \mathbb{R}_{\max}^{n\times n}$ and $i, j, k \in [n]$. Then, for all $t \in \mathbb{N}_0$:*

$$\big(A_{i,k} \otimes A_{k,j}\big)(t) = \max\big\{A(W) \mid W \in \mathcal{W}^t(i \xrightarrow{k} j)\big\} \tag{4.78}$$

Proof. Denote by L the left-hand side and by R the right-hand side, respectively, of the claimed equality.

We first prove $L \leq R$ by showing $A_{i,k}^{\otimes t_1} + A_{k,j}^{\otimes t_2} \leq R$ whenever $t = t_1 + t_2$ with nonnegative integers t_1 and t_2: This inequality is trivial if $A_{i,k}^{\otimes t_1} = -\infty$ or $A_{k,j}^{\otimes t_2} = -\infty$. Otherwise, let $W_1 \in \mathcal{W}^{t_1}(i \to k)$ such that $A(W_1) = A_{i,k}^{\otimes t_1}$ and $W_2 \in \mathcal{W}^{t_2}(k \to j)$ such that $A(W_2) = A_{k,j}^{\otimes t_2}$. Setting $W = W_1 \cdot W_2$ yields $W \in \mathcal{W}^t(i \xrightarrow{k} j)$. Hence $A_{i,k}^{\otimes t_1} + A_{k,j}^{\otimes t_2} = A(W) \leq R$ by definition of R.

We now prove $L \geq R$: The inequality is trivial if $R = -\infty$ Otherwise, let $W \in \mathcal{W}^t(i \xrightarrow{k} j)$. Then W can be written as $W = W_1 \cdot W_2$ with $\mathrm{End}(W_1) = \mathrm{Start}(W_2) = k$. Set $t_1 = \ell(W_1)$ and $t_2 = \ell(W_2)$. Trivially, $t = t_1 + t_2$. Since $W_1 \in \mathcal{W}^{t_1}(i \to k)$ and $W_2 \in \mathcal{W}^{t_2}(k \to j)$, we have $A(W_1) \leq A_{i,k}^{\otimes t_1}$ and $A(W_2) \leq A_{k,j}^{\otimes t_2}$, i.e., $A(W) = A(W_1) + A(W_2) \leq A_{i,k}^{\otimes t_1} + A_{k,j}^{\otimes t_2} \leq L$. This concludes the proof. □

4.10.1 Transience Bounds via Critical Bound

In this subsection, we show how to quickly arrive at a transience bound when combining the convolution approach with a critical bound. It is possible to get a transience bound without explicitly using a critical bound; we discuss this in the following two subsections. The following theorem is a critical bound that is not too hard to prove, even without the apparatus of the preceding sections. It is a particular consequence of Theorem 4.22 when choosing G_0 to be the critical digraph.

Lemma 4.40 (Critical Bound)**.** *Let $A \in \mathbb{R}_{\max}^{n\times n}$ be irreducible. For all $i, j \in [n]$ and $t \in \mathbb{N}_0$, each walk with maximum A-weight in $\mathcal{W}^t(i \to j)$ contains a critical node if*

$$t = B_c \geq \max\left\{ n\,,\ \frac{\|A\| \cdot n^2}{\lambda(A) - \lambda_{nc}(A)} \right\}\,. \tag{4.79}$$

Together with Nachtigall's lemma (Lemma 3.7), this critical bound easily gives a transience bound when using the convolution representation. It is worse than the bounds we presented in the previous sections, but it is nonetheless asymptotically tight, as is shown by Theorem 2.19.

Theorem 4.41. *Let $A \in \mathbb{R}_{\max}^{n\times n}$ be irreducible. Let* cf *be the circumference of the critical digraph $G_c(A)$. Then the transient of A is at most*

$$\max\left\{ \frac{\|A\| \cdot 2n^2}{\lambda(A) - \lambda_{nc}(A)}\,,\ \mathrm{cf} \cdot (2n-1) - 1 \right\}\,. \tag{4.80}$$

Proof. From Lemmas 4.39 and 4.40, we know that

$$A_{i,j}^{\otimes t} = \max_{k \text{ crit.}} \left((A_{i,k} \otimes A_{k,j})(t) \right) \tag{4.81}$$

when $t \geq B_c$. For each critical node k, let ℓ_k denote the length of a critical cycle containing k. By Lemmas 2.7 and 3.7, we obtain that all sequences $A_{i,k}$ and $A_{k,j}$ are eventually periodic, with period ℓ_k, ratio $\lambda(A)$, and transient less or equal to $\mathrm{cf} \cdot (n-1)$ because $\ell_k \leq \mathrm{cf}$. Lemma 4.38 shows that the sequence $\left((A_{i,k} \otimes A_{k,j})(t)\right)_{t\geq 0}$ is eventually periodic, with ratio $\lambda(A)$, and transient less or equal to $2\mathrm{cf} \cdot (n-1) + \mathrm{cf} - 1$. By Lemma 4.37, the same property holds for the sequence of values $\max_{k \text{ crit.}}((A_{i,k} \otimes A_{k,j})(t))$. This proves that each sequence $\left(A_{i,j}^{\otimes t}\right)_t$ is eventually periodic, with ratio $\lambda(A)$, and transient at most equal to

$$\max\left\{ B_c\,,\ \mathrm{cf} \cdot (2n-1) - 1 \right\}\,, \tag{4.82}$$

which concludes the proof. □

4.10.2 Nachtigall Decomposition

Nachtigall [72] introduced a representation of the sequence of matrix powers of an $n \times n$ square matrix as the maximum of at most n "simple" matrix sequences, i.e., eventually periodic sequences with small period and small transient (Theorem 2.15). He showed that this representation can be computed efficiently. However, no results on the transient of the original matrix were obtained. In this section, we present and prove a more precise version of this representation using arguments similar to the original proof. This refined representation will allow us to derive a transience bound of the sequence of powers of the original matrix in the next section.

The proof of Nachtigall's representation picks a maximum mean A-weight cycle in the corresponding digraph of the matrix and decomposes the sequence of matrix powers into the part that involves this strongly connected component and the part that does not. It turns out that the first part can be written as convolutions of two "simple" sequences. The simple structure of the sequences in the convolution relies on the fact that a *maximum* mean A-weight cycle was chosen. The second part is the sequence of powers of the matrix in which the cycle is deleted. The procedure is then recursively applied to the second part.

Given a matrix $A \in \mathbb{R}_{\max}^{n\times n}$ and a set $I \subseteq [n]$ of indices, we define the *deletion* of I in A as the matrix $B \in \mathbb{R}_{\max}^{n\times n}$ whose entries satisfy $B_{i,j} = -\infty$ if $i \in I$ or $j \in I$, and $B_{i,j} = A_{i,j}$ otherwise.

The results of this section are the following: Lemma 4.42 shows the decomposition of the sequence of matrix powers into the part that involves a given set of indices and the part that does not. Finally, in Theorem 4.43, we show an improved version of Nachtigall's decomposition. The improvements lie in a bound of $2n^2$ on the involved transients, whereas Nachtigall states a bound of $3n^2$, and a more precise statement on the form of the involved sequences.

Lemma 4.42. *Let $A \in \mathbb{R}_{\max}^{n\times n}$, $I \subseteq [n]$, and B the deletion of I in A. Then for all $i, j \in [n]$ and all $t \in \mathbb{N}_0$:*

$$A_{i,j}^{\otimes t} = \max\left\{ \max_{k\in I} \left(A_{i,k} \otimes A_{k,j}\right)(t)\,,\, B_{i,j}^{\otimes t} \right\} \tag{4.83}$$

Proof. By Lemma 4.39, we have

$$\max_{k\in I} \left(A_{i,k} \otimes A_{k,j}\right)(t) = \max\left\{ A(W) \mid W \in \mathcal{W}^t(i \xrightarrow{I} j) \right\} \ . \tag{4.84}$$

By definition of deletion, $\mathcal{W}^t_{G(B)}(i \to j)$ is equal to the set of walks in $\mathcal{W}^t_{G(A)}(i \to j)$ that do not contain a node in I. For all these walks W, we have $B(W) = A(W)$. We can hence see that

$$(B^t)_{i,j} = \max\left\{ B(W) \mid W \in \mathcal{W}^t_{G(B)}(i \to j) \right\} \tag{4.85}$$

$$= \max\left\{ A(W) \mid W \in \mathcal{W}^t_{G(A)}(i \to j) \text{ and } W \text{ does not contain a node in } I \right\} \ . \tag{4.86}$$

Forming the maximum over both sides of (4.84) and (4.86) concludes the proof. □

Theorem 4.43 (Improved Nachtigall decomposition)**.** *Let $A \in \mathbb{R}_{\max}^{n\times n}$. Then there exist eventually periodic matrix sequences $S_1(n), S_2(n), \ldots, S_N(n)$ with periods at most n and transients at most $2n^2$ such that for all $t \in \mathbb{N}_0$:*

$$A^t = \max\left\{ S_1(t), S_2(t), \ldots, S_n(t) \right\} \tag{4.87}$$

Moreover, there exist pairwise disjoint subsets $I_1, I_2, \ldots, I_n$ of $[N]$ such that

1. *for all $i, j, m \in [N]$ and all $t \in \mathbb{N}_0$ we have*

$$\big(S_m(t)\big)_{i,j} = \max_{k \in I_t} \Big(\big(B_m\big)_{i,k} \otimes \big(B_m\big)_{k,j} \Big)\,(n) \tag{4.88}$$

   where matrix B_m is defined as the deletion of $\bigcup_{r=1}^{m-1} I_r$ in A.

2. *if $G(A)$ contains a cycle, then $I_1 \neq \emptyset$, $S_1(t)$ has the ratio $\lambda(A)$, and for all $m \in [n]$ the sequence $S_m(t)$ has a ratio equal to the mean A-weight of some cycle in $G(A)$.*

Proof. We define the sets I_m, and hence the matrices B_m and the sequences $S_m(t)$, inductively. By definition, $B_1 = A$. If $G(B_m)$ contains a cycle, denote by C_m a cycle in $G(B_m)$ of maximal mean A-weight and let I_m be the set of nodes of C_m. Otherwise, set $I_m = [n] \setminus \bigcup_{r=1}^{m-1} I_r$. It is clear, in both cases, that I_m is disjoint to every I_r with $r < m$. Also, $|I_m| > 0$ if and only if $|\bigcup_{r=1}^{m-1} I_r| < n$. Because B_{m+1} is the deletion of I_m in B_m, Lemma 4.42 implies

$$(B_m)^t = \max\left\{ S_m(t)\,,\, (B_{m+1})^t \right\}\ . \tag{4.89}$$

Let h denote the greatest positive integer such that $|I_h| > 0$. We derive that $h \leq n$. For all $m > h$, we have $(B_m)_{i,j} = -\infty$ for all i and j, because then B_m is the deletion of $[n]$ in A. Repeated application of (4.89) hence shows (4.87).

Lemmas 4.37, 4.38, and 3.7 show that, whenever $G(B_m)$ contains a cycle, then the transient of $S_m(t)$ is at most

$$2 \cdot \ell(C_m) \cdot (n-1) + \ell(C_m) - 1 \leq 2n^2 - n - 1\ , \tag{4.90}$$

its period is at most $\ell(C_m) \leq n$, and has a ratio equal to $A(C_m)/\ell(C_m)$. If $G(B_m)$ does not contain a cycle, then necessarily $S_m(t) = (B_m)^t$, which is infinite for all $t \geq n$. Hence in this case $S_m(t)$ has transient at most n, period equal to 1, and arbitrary ratio; in particular the smallest mean A-weight of cycles in $G(A)$, if it contains a cycle. □

Note that Theorem 4.43 does *not* imply that the transient of any sequence of matrix powers is at most $2n^2$. The reason for this is that Lemma 4.37 is not necessarily applicable to the maximum in the Nachtigall decomposition because the involved sequences can have different ratios.

4.10.3 Transience Bounds via Nachtigall Decomposition

In this section we deduce from Theorem 4.43 an upper bound on the transient of the sequence of matrix powers A^t of an irreducible square matrix A.

Theorem 4.43 expresses A^t as the maximum of eventually periodic sequences with small transients, i.e., at most $2n^2$. Some of them will share a common ratio, some of them will not. For sequences with a common ratio, Lemma 4.37 is applicable and shows that their maximum has also a transient of at most $2n^2$. For a pair of sequences with different ratios, however, Lemma 4.37 gives no information.

It is possible that the maximum is not even eventually periodic: If f and g are two eventually periodic scalar sequences such that f's ratio is strictly larger than that of g, then the maximum $\max\{f, g\}$ is eventually periodic if and only if, for all t large enough, $f(t) = -\infty$ implies $g(t) = -\infty$. This condition is not necessary for eventual periodicity if the two ratios are equal. For example, define f by setting $f(t) = 2t$ if t is even and $f(t) = -\infty$ if t is odd,

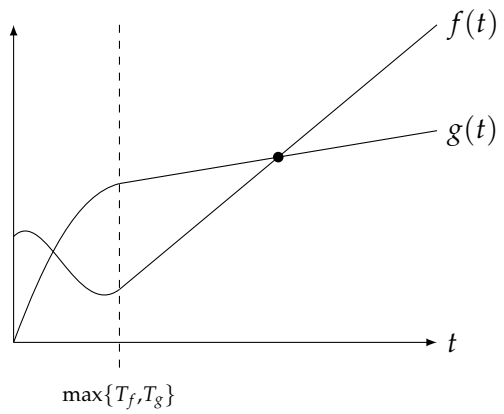


Figure 4.10: Eventually periodic sequences with differing ratios

and g by setting $g(t) = -\infty$ if t is even and $g(t) = t$ if t is odd. Both f and g are eventually periodic and their ratios satisfy $\varrho_f = 2 > \varrho_g = 1$. However, $f(t) = -\infty$ if and only if $g(t) \neq -\infty$ for all $t \in \mathbb{N}_0$. The sequence $\max\{f, g\}$ has the form $\max\{f(t), g(t)\} = 2t$ whenever t is even and $\max\{f(t), g(t)\} = t$ whenever t is odd, which is not eventually periodic.

The next lemma closes the gap by bounding the transient of a maximum of eventually periodic sequences with different ratios. A similar result for the case that both sequences are eventually finite can be constructed from the proof of [15, Proposition 4(2)].

Lemma 4.44. *Let $f, g : \mathbb{N}_0 \to \mathbb{R}_{\max}$ be eventually periodic with respective periods p_f and p_g, respective ratios ϱ_f and ϱ_g, and respective transients T_f and T_g. Assume that $\varrho_f > \varrho_g$ and that there exists an $R \in \mathbb{N}_0$ such that for all $t \geq R$, $f(t) = -\infty$ implies $g(t) = -\infty$.*

Then the sequence $\max\{f, g\}$ *is eventually periodic with period p_f, ratio ϱ_f, and transient at most*

$$T + p - 1 + \frac{\Gamma}{\varrho_f - \varrho_g} \tag{4.91}$$

where $\Gamma = \max\left\{g(s) - f(s') + (s' - s)\varrho_f \mid T \leq s, s' \leq T + p - 1 \text{ and } f(s') \neq -\infty\right\} \cup \{0\}$, $T = \max\{T_f, T_g\}$, *and* $p = \max\{p_f, p_g\}$.

Proof. By eventual periodicity of both f and g, we can assume $R \leq \max\{T_f, T_g\}$, which we do. By replacing $f(t)$ and $g(t)$ by $f(t) - t\varrho_f$ and $g(t) - t\varrho_f$, respectively, we can assume $\varrho_f = 0$.

Let $t \geq T + p - 1 + \Gamma/(\varrho_f - \varrho_g)$. We show that $\max\{f(t), g(t)\} = f(t)$, which then concludes the proof because $t \geq T_f$.

The statement is trivial if $g(t) = -\infty$. So let $g(t) \neq -\infty$. Setting

$$s = t - p_g \cdot \left\lfloor \frac{t - T - p_g + 1}{p_g} \right\rfloor \quad \text{and} \quad s' = t - p_f \cdot \left\lfloor \frac{t - T - p_f + 1}{p_f} \right\rfloor , \tag{4.92}$$

we get $T \leq s, s' \leq T + p - 1$. Because $t \geq R$, also $f(t) \neq -\infty$, which implies $f(s') \neq -\infty$ because $s' \geq T_f$. Hence $g(s) - f(s') \leq \Gamma$.

Because $s, s' \geq T_f, T_g$, we have:

$$g(t) - f(t) = g(s) - f(s') + \varrho_g \cdot p_g \cdot \left\lfloor \frac{t - T - p_g + 1}{p_g} \right\rfloor \tag{4.93}$$

$$\leq \Gamma + \varrho_g \cdot \frac{\Gamma}{-\varrho_g} = 0 \tag{4.94}$$

Hence $f(t) \geq g(t)$, which concludes the proof. □

We now provide a general transience bound for sequences of matrix powers using the Nachtigall decomposition. We do this by applying Lemma 4.37 and Lemma 4.44 entry-wise to the maximum in (4.87). We do this by pairwise comparing the entries of $S_m(t)$ to the entries of some other $S_m(t)$. If they have equal ratios, Lemma 4.37 is applicable. If not, we show by using Theorem 2.9 that Lemma 4.44 is applicable. By part 2 of Theorem 4.43, we know that all S_m have a ratio equal to some cycle mean and that S_1 has ratio $\lambda(A)$. Hence, if S_1 and S_m do not have a common ratio, S_m has a ratio at most equal to the second largest cycle mean, i.e., $\lambda_2(A)$.

We arrive at the following theorem bounding the transient of the sequence of matrix powers. Numerically, it is even worse than the bound of Theorem 4.41, but its proof avoid an explicit critical bound, and is solely based on the Nachtigall decomposition.

Theorem 4.45. *Let $A \in \mathbb{R}_{\max}^{n \times n}$ be irreducible. Then its transient is at most*

$$2n^2 + \frac{3n^2 \|A\|}{\lambda(A) - \lambda_2(A)} \ . \tag{4.95}$$

Proof. Denote by $A_{\min}$ the smallest finite entry of A and by $A_{\max}$ the largest. It is $\|A\| = A_{\max} - A_{\min}$. Let

$$A^t = \max\left\{S_1(t), S_2(t), \ldots, S_n(t)\right\} = \max_{1 \leq m \leq n} \max\left\{S_1(t), S_m(t)\right\} \tag{4.96}$$

as in Theorem 4.43. Let $i, j, r \in [n]$. We will show that the sequence $\max\left\{S_1, S_m\right\}$ has ratio $\lambda(A)$ and transient at most $2n^2 + 3n^2\|A\|/(\lambda(A) - \lambda_2(A))$. Component-wise application of Lemma 4.37 then concludes the proof.

If the ratios of S_1 and S_m are equal, we simply apply Lemma 4.37 to see that the sequence $\max\left\{S_1, S_m\right\}$ is eventually periodic with ratio $\lambda(A)$ and transient at most $2n^2$.

We apply Lemma 4.44 to the two sequences $f(t) = \big(S_1(t)\big)_{i,j}$ and $g(t) = \big(S_m(t)\big)_{i,j}$. Both T_f and T_g are at most $2n^2$, ϱ_f is equal to $\lambda(A)$, and ϱ_g is at most $\lambda_2(A)$ by part 2 of Theorem 4.43. Because both sequences have periods at most equal to n, they have a common period p less than n^2.

We now show that we can choose $R = n^2$: Let $n \geq n^2$ and $\big(S_m(t)\big)_{i,j} \neq -\infty$. In particular, by noting (4.88) and Lemma 4.39, there exist a walk $\hat{W}$ in $G(B_m)$ from i to j of length t containing a node of I_m. Walk $\hat{W}$ is also a walk in $G(A)$ because B_m is a sub-digraph of $G(A)$. Because $G(A)$ is strongly connected, there exists a $k \in I_1$ by part 2 of Theorem 4.43. Theorem 2.9 shows that there exists a walk in $\mathcal{W}^n_{G(A)}(i \to j)$ that contains node k, which implies $\big(S_1(t)\big)_{i,j} \neq \infty$ by Lemma 4.39.

Hence $\max\{T_f, T_g, R\} \leq 2n^2$ and thus

$$p + \max\{T_f, T_g, R\} \leq 3n^2 \ . \tag{4.97}$$

By Lemma 4.39, whenever $\big(S_1(t)\big)_{i,j}$ is finite, it is the A-weight of some walk in $G(A)$ of length t. In particular,

$$\big(S_1(t)\big)_{i,j} \leq A_{\max} \cdot t \ . \tag{4.98}$$

Similarly,

$$\big(S_m(t)\big)_{i,j} \geq A_{\min} \cdot n \quad \text{if } \big(S_m(t)\big)_{i,j} \neq -\infty \ . \tag{4.99}$$

Combining (4.97), (4.98), and (4.99) hence yields $\Gamma \leq 3n^2(A_{\max} - A_{\min}) = 3n^2\|A\|$.

In summary, Lemma 4.44 shows that $\max\big\{S_1, S_m\big\}$ is eventually periodic with ratio $\lambda(A)$ and transient at most $2n^2 + 3n^2\|A\|/(\lambda(A) - \lambda_2(A))$. This concludes the proof. □

Chapter 5

Asymptotic Consensus

5.1 Introduction

Asymptotic consensus is a phenomenon observed in certain biological, physical, and sociological systems. It is also utilized in some engineered man-made computer systems. The phenomenon consists in agents communicating in a very simple fashion to asymptotically reach agreement on a common real value. In nature, it can be observed (e.g.,[76, 58, 86, 7, 54, 55]) in bird flocking, firefly synchronization, synchronization of coupled oscillators, or opinion spreading. In engineering, it is used for sensor fusion, dynamic load balancing protocols, robot formation protocols, replication techniques, or rendezvous in space.

The distributed computing model in which we study asymptotic consensus is the following: There are n distinguishable agents, each agent $i \in [n] = \{1, 2, \ldots, n\}$ possessing a real state variable x_i and communicating by exchanging messages. There is a global discrete time base, referred to by nonnegative integers in $\mathbb{N}_0 = \{0, 1, 2, \ldots\}$. At every time $t \in \mathbb{N}_0$, we denote the content of the agents' state variables by $x_i(t)$. The initial value of state variable x_i is $x_i(0)$. At every time $t \in \mathbb{N}_0$, every agent sends the content of its state variable to all other agents. Messages may be delayed and/or lost. All agents simultaneously update their state variable at all positive times $t = 1, 2, 3, \ldots$ to some weighted average value of the received values, at most one of each other agent, and its current content of its own state variable.

Since the new content of the state variable is a mean value, for each agent i and each time $t \geq 1$, there exist coefficients $A_{i,j,\tau}(t)$ with

$$x_i(t) = \sum_{j=1}^{n} \sum_{\tau=0}^{t-1} A_{i,j,\tau}(t) \cdot x_j(\tau) \tag{5.1}$$

and

$$\sum_{j=1}^{n} \sum_{\tau=0}^{t-1} A_{i,j,\tau}(t) = 1 \ . \tag{5.2}$$

Since at most one value of every agent appears in the mean value, there exists a $\delta_{i,j}(t) > 0$ for every $j \in [n]$ such that $A_{i,j,\tau}(t)$ is zero for all τ except possibly for $\tau = t - \delta_{i,j}(t)$. Hence (5.1) can be rewritten as

$$x_i(t) = \sum_{j=1}^{n} A_{i,j}(t) \cdot x_j\big(t - \delta_{i,j}(t)\big) \tag{5.3}$$

with

$$\sum_{j=1}^{n} A_{i,j}(t) = 1 \ . \tag{5.4}$$

A *configuration* of asymptotic consensus is a collection of real values, one for each agent's state variable, i.e., a vector in $\mathbb{R}^n$. An *execution* of asymptotic consensus is an infinite sequence of configurations $x(t) \in \mathbb{R}^n$ following the evolution (5.3) for some choice of the $A_{i,j}(t)$ and the $\delta_{i,j}(t)$. An execution *reaches asymptotic consensus* if $x(t)$ converges and all component-wise limits $\lim_{t\to\infty} x_i(t)$ are equal.

An *averaging matrix* is a matrix whose entries are all nonnegative and whose row sums are all 1. In other words, it is a row stochastic matrix. Equation (5.4) assures that the collection of the $A_{i,j}(t)$ is an averaging matrix for all t. A *delay matrix* for time t is a matrix of integers between 1 and t. For every t, the collection of the $\delta_{i,j}(t)$ is a delay matrix for t. Hence an execution is determined by the initial configuration $x(0)$, the sequence of the averaging matrices $A(t)$, and the sequence of the delay matrices $\delta(t)$. A pair consisting of a sequence of averaging matrices $A(t)$ and a sequence of vectors $\delta(t)$ such that every $\delta(t)$ is a delay matrix for t is referred to as a *setting*. An *environment* is a nonempty set of settings. We say that a setting or an environment reaches asymptotic consensus if all of its executions do.

An important parameter of a setting is the maximum entry of the delay matrices, if it exists. We call a setting *B-bounded* if all entries of its delay matrices are at most B. A 1-bounded setting is called *synchronous* and is determined by only the sequence of averaging matrices. If the nonzero entries of the averaging matrices are lower bounded by some positive α, then we say that the setting has *minimal confidence* α. It has *self-confidence* if all diagonal entries are positive. The *communication digraph* of a stochastic matrix A in $\mathbb{R}^{n\times n}$ has node set $[n]$ and contains an edge (i, j) if and only if $A_{i,j} > 0$.

In a synchronous setting, the evolution of configurations $x(t)$ is governed by the linear recursive law

$$x(t) = A(t) \cdot x(t-1) \tag{5.5}$$

where $A(t)$ is a row stochastic matrix. Defining the product matrices

$$P(t) = A(t) \cdot A(t-1) \cdots A(1) \ , \tag{5.6}$$

we have

$$x(t) = P(t) \cdot x(0) \ . \tag{5.7}$$

In particular, the sequence of state vectors is determined by the initial vector $x(0)$ and the sequence of row stochastic matrices $A(t)$.

In the following sections, we will also use the notation

$$P(t,s) = A(t) \cdot A(t-1) \cdots A(s+1) \tag{5.8}$$

for partial products. It is $P(t) = P(t, 0)$ for all t and $P(t, s) = I$, the identity matrix, if $t \leq s$. We will also refer to row stochastic matrices simply as *stochastic matrices*. If all $A(t)$ are equal to a constant matrix A, then $P(t) = A^t$. We will use the notation ${}^tA$ for the transpose of A to distinguish it from its t^{th} power.

We now exhibit asymptotic consensus through the example of bird flocking with n birds: At every time t of some discrete time base, every bird i has a position in Euclidean space

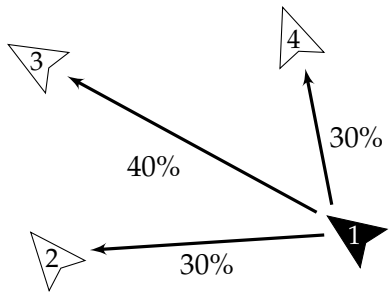


Figure 5.1: Birds' observations while flocking

$s_i(t) = \left(s_i^{(1)}(t)\,,\, s_i^{(2)}(t)\,,\, s_i^{(3)}(t)\right) \in \mathbb{R}^3$ and a velocity $v_i(t) = \left(v_i^{(1)}(t)\,,\, v_i^{(2)}(t)\,,\, v_i^{(3)}(t)\right) \in \mathbb{R}^3$. We assume that the time base is $\mathbb{N}_0$ and that the birds' velocities influence their positions in the physically obvious manner, i.e., that

$$s_i(t) = s_i(t-1) + v_i(t) \tag{5.9}$$

for all $t \geq 1$ and all birds $i \in [n]$.

As described, among others, by Ballerini et al. [7], birds seem to follow three rule sets while flocking: (1) attraction towards each other, (2) short range repulsion to avoid collisions, and (3) alignment of the velocities. The application of these rule sets is determined by the position and velocity of surrounding birds, each being taken into account with a certain weight depending, among other factors, on metric distance and the angle between the vision center and the line of sight to the neighbor. Figure 5.1 shows an example of the relative weights with which a bird perceives its neighbors.

Based on these observations, the birds change their velocity following the three rule sets. We ignore the first two and focus only on the alignment of velocities, which is an instance of asymptotic consensus. To align the velocities, a bird observes a set of neighbors and changes its own velocity to a weighted average of their velocities and its own current velocity. The new adapted velocity is influenced by its own current velocity, if only because of its inertia. There hence exists a stochastic matrix $A(t)$ such that

$$v_i(t) = \sum_{j=1}^{n} A_{i,j}(t) \cdot v_j(t-1) \tag{5.10}$$

for all $i \in [n]$. Figure 5.2 depicts the influence on the velocity for our example; the first row of $A(t)$ is equal to $\left(0.5\,,\, 0.15\,,\, 0.2\,,\, 0.15\right)$.

Writing Equation (5.10) for all entries of the vectors on both sides, we have

$$v_i^{(d)}(t) = \sum_{j=1}^{n} A_{i,j}(t) \cdot v_j^{(d)}(t-1) \tag{5.11}$$

for all $d \in \{1,2,3\}$ and all $i \in [n]$. Writing $v^{(d)}(t)$ for the vector $\left(v_1^{(d)}(t)\,,\, v_2^{(d)}(t)\,,\, \ldots\,,\, v_n^{(d)}(t)\right)$ in $\mathbb{R}^n$, this translates into

$$v^{(d)}(t) = A(t) \cdot v^{(d)}(t-1) \tag{5.12}$$

for all $d \in \{1,2,3\}$. Hence the sequences $v^{(1)}(t)$, $v^{(2)}(t)$, and $v^{(3)}(t)$ are executions of asymptotic consensus with identical synchronous settings.

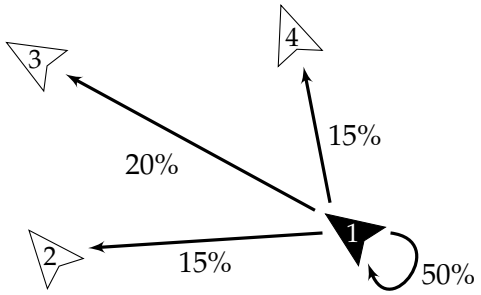


Figure 5.2: Birds' influence on new adapted velocity

Note that the matrices $A(t)$ are dependent on the birds' positions $s_i(t)$, which in turn depend on the previous velocities $v_i(t-1)$. Hence the setting is dependent on the initial positions and velocities. In the following sections, we will consider the setting and the initial vector to be fixed by an external force. We will give necessary conditions for settings for which asymptotic consensus is reached and bound the speed of convergence. Hence, to use these results for an application like bird flocking, it is necessary to show that the settings that occur in the applications satisfy one of the necessary conditions.

5.2 State of the Art

There is a well-known sufficient graph-theoretic condition for constant synchronous settings with averaging matrix A to reach asymptotic consensus. It uses the notion of the *digraph of a stochastic matrix* A in $\mathbb{R}^{n\times n}$, which is defined as the digraph with node set $V = [n]$ containing an edge $(i,j) \in E$ if and only if $A_{i,j} > 0$. We denote it by $G(A)$. We have the equality $G(A^t) = G(A)^t$ for all $t \in \mathbb{N}_0$.

Definition 5.1 (Ergodic matrices)**.** A stochastic matrix A is *ergodic* if $G(A)$ is primitive.

The vector $\mathbf{1} = {}^t(1,1,\ldots,1)$ is a right-eigenvector to the eigenvalue 1 of every stochastic matrix. By the Perron-Frobenius theorem [43], if A is an ergodic matrix, there exists a unique left-eigenvector π to the eigenvalue 1 that is a probability vector, i.e., ${}^t\pi \cdot A = {}^t\pi$. We call this eigenvector π the *Perron vector* of A.

The following theorem states the sufficiency of ergodicity and also the fact that the speed of convergence is exponential. It is a classical result in Markov theory.

Theorem 5.2. *Let A be an ergodic matrix in $\mathbb{R}^{n\times n}$ with Perron vector π. Then there exists some $\varrho < 1$ such that $|A^{(t)}_{i,j} - \pi_j| = O(\varrho^t)$ for all $i,j \in [n]$ as $t \to \infty$.*

Translated into the language of asymptotic consensus, this theorem states that every constant synchronous execution with ergodic averaging matrix A reaches asymptotic consensus, and that $|x_i(t) - c_\infty| = O(\varrho^t)$ where c_∞ is the common limit of the $x_i(t)$. It also motivates the definition of the *rate of convergence* of an execution as the value $\lim_{t\to\infty}\|x(t) - c_\infty \cdot \mathbf{1}\|^{1/t}$ where $\|\cdot\|$ is some norm. The rate of convergence of a setting is the maximum rate of convergence of its executions.

Numerous techniques have been developed for bounding the rate of convergence more concretely. Two of the main techniques are coupling [75, 2] and stopping and stationary times [3, 4], which both are probabilistic techniques and are often adapted to special classes of stochastic matrices. A third large class of techniques are spectral methods, which study the eigenvalues of the stochastic matrix. We will detail this class of techniques a bit more in the remainder of this subsection.

The fundamental result for the spectral method is equality of the rate of convergence and the second largest eigenvalue. Denote by $\rho_2(A)$ the second largest absolute value of eigenvalues of A, counted with multiplicity. All eigenvalues of a stochastic matrix have absolute value at most 1, and $\rho_2(A)$ is strictly less than 1 if A is ergodic. The term $1 - \rho_2(A)$ is called the matrix's *spectral gap*. A lower bound on the spectral gap directly translates to an upper bound on $\rho_2(A)$.

Theorem 5.3. *The rate of convergence of a constant synchronous setting with averaging matrix A is equal to $\rho_2(A)$.*

Many authors have proposed bounds on the value $\varrho_2(A)$, including Mihail [70], Diaconis and Stroock [35], Sinclair and Jerrum [83], Fill [42], Chung [29], and Landau and Odlyzko [61].

Tsitsiklis introduced the *bounded intercommunication* assumption. It states that if an edge (i, j) appears in infinitely many communication digraphs, then is appears in one of the digraphs $G(A(t)), G(A(t+1)), \ldots, G(A(t+B-1))$ for a fixed B and all t.

Theorem 5.4 (Tsitsiklis [88])**.** *A synchronous setting with averaging matrices $A(1), A(2), \ldots$ with self-confidence and minimal confidence α reaches asymptotic consensus if the digraph G_∞ formed by the edges appearing in infinitely many communication digraphs is strongly connected and the bounded intercommunication assumption holds.*

Moreau and Hendrickx and Blondel independently showed that the bounded intercommunication assumption can be replaced by the assumption that every communication digraph is bi-directional:

Theorem 5.5 (Moreau [71], Hendrickx and Blondel [57])**.** *A synchronous setting with averaging matrices $A(1), A(2), \ldots$ with self-confidence and minimal confidence α reaches asymptotic consensus if the digraph G_∞ formed by the edges appearing in infinitely many communication digraphs is strongly connected and every communication digraph is bi-directional.*

Blondel et al. generalized this result to B-bounded settings:

Theorem 5.6 (Blondel et al. [13])**.** *A B-bounded setting with averaging matrices $A(1), A(2), \ldots$ with self-confidence and minimal confidence α reaches asymptotic consensus if the digraph G_∞ formed by the edges appearing in infinitely many communication digraphs is strongly connected and every communication digraph is bi-directional.*

Touri and Nedić generalized the assumption of bi-directional digraphs to digraphs that are completely reducible. Charron-Bost [24] very recently showed its extension to B-bounded settings.

Theorem 5.7 (Touri and Nedić [87])**.** *A synchronous setting with averaging matrices $A(t)$ with self-confidence and minimal confidence α reaches asymptotic consensus if the digraph G_∞ formed by the edges appearing in infinitely many communication digraphs is strongly connected and every communication digraph is completely reducible.*

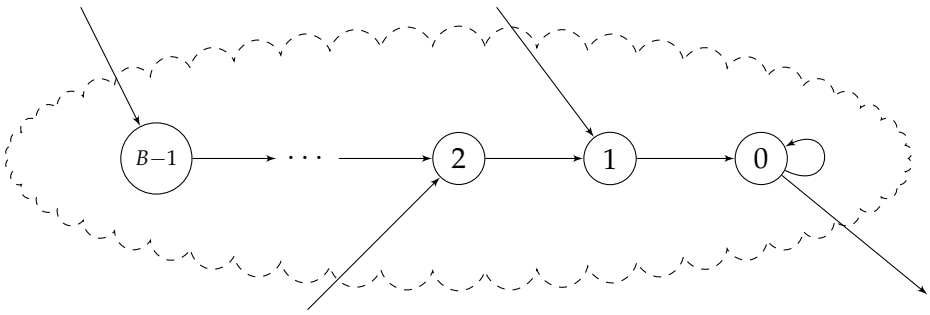


Figure 5.3: The B copies of an agent in Cao, Morse, and Anderson's reduction

If an execution $x(t)$ reaches asymptotic consensus, one can ask the question of the speed at which this convergence occurs. Olshevsky and Tsitsiklis noted that this speed is sometimes exponential and have hence defined the *rate of convergence* as

$$\lim_{t\to\infty} \|x(t) - x^*\|_2^{1/t} \ . \tag{5.13}$$

Theorem 5.8 (Olshevsky and Tsitsiklis [73])**.** *A synchronous setting with constant equal-neighbor averaging matrices $A(t) = A$ of a connected bi-directional digraph reaches consensus. Moreover, the rate of convergence is at most $1 - \Omega(n^3)$. There exist connected bi-directional digraphs such that the rate of convergence is $1 - O(n^3)$.*

Cao, Morse, and Anderson studied *coordinated* communication digraphs, i.e., digraphs that have a node j such that every other node has a path to j. They obtained the following result:

Theorem 5.9 (Cao, Morse, and Anderson [21, 22])**.** *A B-bounded setting with averaging matrices $A(1), A(2), \ldots$ with self-confidence and minimal confidence α reaches asymptotic if every communication digraph is coordinated. Moreover, the rate of convergence is less than 1.*

To prove their result, they described a reduction of B-bounded settings to synchronous settings, albeit with B times as many agents as the original setting [22, Section 4.1]. The idea is to replicate every agent B times, but to shift the copies in time, i.e., at time t there is one copy holding the value $x_i(t)$, one $x_i(t-1)$, and so on until $x_i(t-B+1)$. This results in synchronous setting for asymptotic consensus. The replication of agents is illustrated in Figure 5.3. Only the copy for the current value $x_i(t)$ has links to other agents' copies. Nonetheless, no such restriction exists for incoming edges. In the new resulting communication digraphs, even if all agents have self-loops in the original communication digraphs, not all nodes have them.

Chazelle introduced and studied the *s-energy* of executions, which gives a means to talk about the efficiency of convergence even in the case when the convergence rate is 1 and the convergence speed is arbitrarily slow. He also proved bounds on the rate of convergence for *fixed confidence equal-neighbor* settings. Their averaging matrices have the form

$$A_{i,j} = \begin{cases} c_i & \text{if } (i,j) \in E \text{ and } i \neq j \\ 1 - (d(i) - 1)c_i & \text{if } i = j \\ 0 & \text{else} \end{cases} \tag{5.14}$$

with $c_i d(i) \leq 1$ and fixed c_i, where $d(i)$ is the degree of agent i.

Theorem 5.10 (Chazelle [28]). *A synchronous setting with averaging matrices $A(1), A(2), \ldots$ that are fixed self-confidence equal neighbor matrices of connected bi-directional digraphs reaches asymptotic consensus. Moreover, the rate of convergence is less than $1 - c/n^2$ where c is the minimal confidence.*

5.3 Rate of Convergence in Constant Synchronous Settings

In this section, we study synchronous settings with a constant averaging matrix $A(t) = A$. We further restrict our attention to the case that A is ergodic, i.e., the communication graph $G(A)$ is primitive. Ergodicity of the matrix is a sufficient condition for reaching asymptotic consensus in a constant synchronous setting (Theorem 5.2). We make this restriction to be able to use linear algebraic spectral methods to assess the speed of convergence. The condition of ergodicity, however, is not necessary for reaching asymptotic consensus in a constant setting.

For every probability vector $\pi \in \mathbb{R}^n$, one defines the inner product

$$\langle x, y \rangle_\pi = \sum_{i=1} \pi_i \overline{x_i} y_i \tag{5.15}$$

on $\mathbb{C}^n$. It is positive definite if and only if π is positive. If it is, then

$$\|x\|_\pi = \sqrt{\langle x, x \rangle_\pi} \tag{5.16}$$

is a norm on $\mathbb{C}^n$. An ergodic matrix $A \in \mathbb{R}^{n \times n}$ with Perron vector π is *reversible* if $\pi_i A_{i,j} = \pi_j A_{j,i}$ for all $i, j \in [n]$. An ergodic matrix A with Perron vector π is self-adjoint with respect to $\langle \cdot, \cdot \rangle_\pi$ if and only if it is reversible.

To give an upper bound on $\varrho_2(A)$, and hence the rate of convergence, we will need the following lemma about the real quadratic form defined by a strongly connected digraph.

Lemma 5.11 (Poincaré inequality). *Let $G = ([n], E)$ be a digraph with diameter at most D. Then the real quadratic form defined by $Q(z) = \sum_{(i,j) \in E} (z_i - z_j)^2$ satisfies the inequality $Q(z) \geq 1/D \cdot (z_a - z_b)^2$ for all $a, b \in [n]$.*

Proof. Let P be a path from a to b in G. Trivially,

$$Q(z) \geq \sum_{(i,j) \text{ in } P} (z_i - z_j)^2 \ . \tag{5.17}$$

By the Cauchy-Schwarz inequality, because there are at most D edges in P, we now deduce

$$Q(z) \geq \frac{1}{D} \left(\sum_{(i,j) \text{ in } P} (z_i - z_j) \right)^2 = \frac{1}{D} (z_a - z_b)^2 \ , \tag{5.18}$$

which concludes the proof. □

The next lemma is an expression of an eigenvalue of A as a sum involving the squared differences of components of a corresponding eigenvector. We denote the real part of a complex number x by $\Re x$.

Lemma 5.12. *Let A be an $n \times n$ ergodic matrix with Perron vector π and let z be an eigenvector of A corresponding to some eigenvalue λ, i.e., $Az = \lambda z$. Further assume that $\|z\|_\pi = 1$. Then*

$$1 - \Re\lambda = \frac{1}{2} \sum_{1 \le i,j \le n} \pi_i A_{i,j} |z_i - z_j|^2 \ . \tag{5.19}$$

Proof. We denote by S the sum on the right-hand side of the claimed equality. We first calculate:

$$\begin{aligned} S &= \frac{1}{2} \sum_{i,j} \pi_i A_{i,j} (z_i - z_j)(\bar{z}_i - \bar{z}_j) \\ &= \frac{1}{2} \sum_{i,j} \pi_i A_{i,j} |z_i|^2 - \frac{1}{2} \sum_{i,j} \pi_i A_{i,j} z_i \bar{z}_j - \frac{1}{2} \sum_{i,j} \pi_i A_{i,j} \bar{z}_i z_j + \frac{1}{2} \sum_{i,j} \pi_i A_{i,j} |z_j|^2 \end{aligned} \tag{5.20}$$

Because A is stochastic, we have $\sum_j A_{i,j} = 1$ and hence

$$\frac{1}{2} \sum_{i,j} \pi_i A_{i,j} |z_i|^2 = \frac{1}{2} \sum_i \pi_i |z_i|^2 = \frac{1}{2} \|z\|_\pi^2 = \frac{1}{2} \ . \tag{5.21}$$

Because π is a left-eigenvector to the eigenvalue 1, we have $\sum_i \pi_i A_{i,j} = \pi_j$ and hence

$$\frac{1}{2} \sum_{i,j} \pi_i A_{i,j} |z_j|^2 = \frac{1}{2} \sum_j \pi_j |z_j|^2 = \frac{1}{2} \|z\|_\pi^2 = \frac{1}{2} \ . \tag{5.22}$$

Because all π_i and $A_{i,j}$ are real, we also have

$$\begin{aligned} \frac{1}{2} \sum_{i,j} \pi_i A_{i,j} z_i \bar{z}_j + \frac{1}{2} \sum_{i,j} \pi_i A_{i,j} \bar{z}_i z_j &= \Re \sum_{i,j} \pi_i A_{i,j} \bar{z}_i z_j \\ &= \Re \langle z, Az \rangle_\pi = \Re \langle z, \lambda z \rangle_\pi \\ &= \Re\lambda \, \|z\|_\pi^2 = \Re\lambda \end{aligned} \tag{5.23}$$

Plugging (5.21), (5.22), and (5.23) into (5.20) now yields $S = 1 - \Re\lambda$. □

5.3.1 The Reversible Case

We now state our main theorem about reversible ergodic matrices. The proof essentially follows a proof given by Chazelle [28, Section 3.3] for fixed confidence equal-neighbor settings.

Theorem 5.13. *Let A be the averaging matrix of a constant synchronous setting Σ with self-confidence at least $\hat{\alpha} > 0$. Assume further that A is ergodic and reversible with Perron vector π and that the communication digraph contains a spanning strongly connected sub-digraph H of diameter D such that $\pi_i A_{i,j} \ge \beta > 0$ for all edges (i,j) in H. Then Σ's spectral gap is at least $\min\{\beta/2D, 2\hat{\alpha}\}$.*

In particular, it is at least $\pi_{\min}\alpha/2D$ where α is Σ's minimal confidence and $\pi_{\min}$ is the minimal entry of the Perron vector π.

Proof. The matrix A is self-adjoint with respect to $\langle \cdot, \cdot \rangle_\pi$ and hence all of its eigenvalues are real. Let $1 = \lambda_1 > \lambda_2 \ge \lambda_3 \ge \cdots \ge \lambda_n > -1$ be the eigenvalues of A and let (v_k) be an orthonormal eigenbasis (with respect to $\langle \cdot, \cdot \rangle_\pi$) of A with $v_1 = {}^t(1, \ldots, 1)$ and $Av_k = \lambda_k v_k$. It

is $\varrho_2(A) = \max\{|\lambda_2|, |\lambda_n|\}$. We will show that $\lambda_n \geq -1 + 2\alpha$ and $\lambda_2 \leq 1 - \beta/2n$. This then concludes the proof.

To show the first inequality, define the stochastic matrix $B = (A - \hat{\alpha}I)/(1 - \hat{\alpha})$. Because of the relation $Bv_k = (\lambda_k - \hat{\alpha})/(1 - \hat{\alpha}) \cdot v_k$, all of its eigenvalues are real. In fact, matrix B is self-adjoint. Denote by $1 = \mu_1 \geq \mu_2 \geq \cdots \geq \mu_n \geq -1$ the eigenvalues of B. It is $\mu_k = (\lambda_k - \hat{\alpha})/(1 - \hat{\alpha})$. In particular $(\lambda_n - \hat{\alpha})/(1 - \hat{\alpha}) \geq -1$, from which $\lambda_n \geq -1 + 2\hat{\alpha}$ follows.

Set $z = v_2$. There exists at least one index a with $z_a^2 \geq 1$ since otherwise we have $\|z\|_\pi^2 = \sum_{i=1}^n \pi_i z_i^2 < \sum_{i=1}^n \pi_i = 1$. Also, there exists at least one index b with $\operatorname{sgn}(z_a) \neq \operatorname{sgn}(z_b)$ because otherwise $\langle z, v_1 \rangle_\pi = \sum_{i=1}^n \pi_i z_i \neq 0$. In particular, $|z_a - z_b| \geq 1$. We have, by definition of $\pi_{\min}$ and α, and by Lemmas 5.11 and 5.12,

$$\begin{aligned} 1 - \lambda_2 = \langle z, (1 - \lambda_2)z \rangle_\pi = \langle z, (I - A)z \rangle_\pi = \frac{1}{2} \sum_{1 \leq i,j \leq n} \pi_i A_{i,j} |z_i - z_j|^2 \\ \geq \frac{1}{2} \sum_{(i,j) \text{ in } H} \pi_i A_{i,j} |z_i - z_j|^2 \geq \frac{\beta}{2} \sum_{(i,j) \text{ in } H} |z_i - z_j|^2 \geq \frac{\beta}{2D} |z_a - z_b|^2 \geq \frac{\beta}{2D} \ . \end{aligned} \tag{5.24}$$

This concludes the proof. □

Example 5.14 (Equal neighbor settings with bi-directional communication digraphs)**.** If A is the equal-neighbor matrix of a connected bi-directional digraph $G = (V, E)$ with self-loops, then it is of the form $A_{i,j} = 1/d(i)$ if $(i, j) \in E$ where $d(i)$ is the out-degree of node i in G. The fact that G includes self-loops ensures its primitivity. Hence A is ergodic. A short computation verifies that its Perron vector is given by $\pi_i = d(i)/|E|$ and that A is reversible.

In the constant synchronous setting with averaging matrix A, all confidences are lower-bounded by $\alpha = \hat{\alpha} \geq 1/n$ where n is the number of nodes in G. For all edges $(i, j) \in E$, we have $\pi_i \cdot A_{i,j} = 1/|E| \geq \beta = 1/n^2$. Theorem 5.13 hence gives the bound $1/2Dn^2 \geq 1/2n^3$ on the setting's spectral gap where D is the diameter of G.

This cubic lower bound is well-known and usually derived by different methods such as the cover time of the equivalent lazy random walk (e.g., [9]). It is asymptotically tight in the number of nodes. The estimation $\beta \geq \pi_{\min}\alpha$ does *not* yield the same bound, as $\pi_{\min}$ can be in the order of $1/n^2$, which would yield the lower bound $\geq 1/2Dn^3 \geq 1/2n^4$ on the spectral gap. This is one power of n worse than utilizing the precise definition of β.

Similarly, if A is a fixed confidence equal neighbor matrix of a bi-directional graph with self-loops: Every diagonal entry is at least $\hat{\alpha} \geq c$ where c is the smallest confidence. Furthermore, $\pi_i = 1/\left(c_i \sum_k 1/c_k\right)$ and thus $\pi_i A_{i,j} \geq \beta = c/n$ since $\sum_k 1/c_k \leq n/c$. This yields a lower bound of $c/2Dn$ on the spectral gap with the theorem.

We hence recover a particular case of Theorem 5.10. Its general form will be a special case of Corollary 5.25.

Corollary 5.15. *A constant synchronous fixed confidence equal-neighbor setting with communication digraph G reaches asymptotic consensus if G is bi-directional and connected. Furthermore, the rate of convergence is at most $1 - 1/2n^2c$ where n is the number of agents and c is the smallest confidence.*

We have seen that utilizing the bound $1 - \varrho_2(A) \geq \pi_{\min}\alpha/2D$ might not always be sufficient to prove tight bounds on the spectral gap. Nevertheless, we can use it to prove worst-case bounds independent of the particular structure and entries of A. For this, we prove a lower bound on $\pi_{\min}$ solely in terms of the minimal positive entry α and the diameter D:

Lemma 5.16. *Let A be an $n \times n$ ergodic matrix with Perron vector π, minimal positive entry α, and let D be the diameter of $G(A)$. Then the minimal entry $\pi_{\min}$ of π fulfills $\pi_{\min} \geq \alpha^D/n$.*

Proof. Let $j \in [n]$ be any index. We first show that $\pi_j \leq \alpha^{-D}\pi_{\min}$:

There exists a path $j_0, j_1, \ldots, j_\ell$ in $G(A)$ from $j_0 = j$ to j_ℓ with $\pi_{j_\ell} = \pi_{\min}$ and length at most D, i.e., $\ell \leq D$. Because π is a left-eigenvector to 1 and the edge (j_{k-1}, j_k) is in $G(A)$, we have

$$\pi_{j_k} = \sum_{i=1}^{n} \pi_i A_{i,j_k} \geq \pi_{j_{k-1}} A_{j_{k-1},j_k} \geq \alpha \pi_{j_{k-1}} \tag{5.25}$$

for all $1 \leq k \leq \ell$, and we have thus shown $\pi_{\min} \geq \alpha^\ell \pi_j \geq \alpha^D \pi_j$.

But then

$$1 = \sum_{j=1}^{n} \pi_j \leq n\alpha^{-D}\pi_{\min} \ , \tag{5.26}$$

which implies $\pi_{\min} \geq \alpha^D/n$. □

Combining the lower bound on $\pi_{\min}$ with the second part of Theorem 5.13, we hence get a bound on the spectral gap solely in terms of the minimal confidence, the diameter, and the number of agents:

Corollary 5.17. *The spectral gap of a constant synchronous setting with reversible averaging matrix and self-confidence is at least $\alpha^{D+1}/2Dn$ where α is the minimal confidence, D is the communication digraph's diameter, and n is the number of agents.*

5.3.2 Extension to the Non-Reversible Case

If A is not reversible, we cannot directly use the proof of Theorem 5.13 because it relies on the fact that the second largest eigenvalues are real, as it uses Lemma 5.12, which only talks about the real part of the eigenvalues. It also relies on the fact that the relevant eigenvector is real. Both are not true in general if the matrix is not reversible. In this subsection, we will use a reversibilization of A, i.e., a reversible matrix whose eigenvalues have a certain relation to the eigenvalues of the original matrix A.

Two natural candidates for reversibilizations are the "multiplicative" one $A^* \cdot A$ and the "additive" one $(A + A^*)/2$ (cf. [42]) where A^* denotes the adjoint matrix with respect to the inner product $\langle \cdot, \cdot \rangle_\pi$ of A's Perron vector π. We will use the multiplicative one. Clearly, both are self-adjoint and hence reversible with respect to π. The additive reversibilization is always ergodic if A is because the digraph $G(A)$ is contained in that of $(A + A^*)/2$. The multiplicative reversibilization is not necessarily ergodic if A is, but it is ergodic if the diagonal of A is positive because then $G(A)$ is also a sub-digraph of $G(A^*A)$.

The fundamental use of the multiplicative reversibilization is the fact that it occurs when looking at the norm $\|Az\|_\pi^2 = \langle A^*Az, z\rangle_\pi$. The next lemma shows that the latter value is upper bounded by $\varrho_2(A^*A) \cdot \|z\|_\pi^2$ whenever z is π-orthogonal to $\mathbf{1}$. It is the variational characterization of the second largest eigenvalue. As π-orthogonality is preserved when applying A, this then shows that $A^t z \to 0$ for all z with $\langle z, \mathbf{1}\rangle_\pi = 0$ if $\varrho_2(A^*A) < 1$, i.e., if A^*A is ergodic.

Lemma 5.18. *Let B be ergodic and reversible with Perron vector π. Then*

$$\varrho_2(B) = \max\left\{ |\langle Bz, z\rangle_\pi| \mid \langle z, \mathbf{1}\rangle_\pi = 0 \text{ and } \|z\|_\pi = 1 \right\} \ . \tag{5.27}$$

Proof. Matrix B is self-adjoint and hence all of its eigenvalues are real and it has an orthonormal eigenbasis. Let $\{v_1, \ldots, v_n\}$ be such a basis with $Bv_k = \lambda_k v_k$, $v_1 = \mathbf{1}$, and $|\lambda_2| = \varrho_2(B)$.

It is $\|z\|_\pi = 1$ if and only if there exists an $\alpha \in \mathbb{R}^n$ with $\|\alpha\|_2 = 1$ such that $z = \sum_{k=1}^n \alpha_k v_k$. It is $\langle z, \mathbf{1}\rangle_\pi = 1$ if and only if $\alpha_1 = 0$. We have

$$|\langle Bz, z\rangle_\pi| = \sum_{k=1}^n |\lambda_k| \cdot |\alpha_k|^2 = |\alpha_1|^2 + \varrho_2(B) \cdot |\alpha_2|^2 + \sum_{k=3}^n |\lambda_k| \cdot |\alpha_k|^2 \ . \tag{5.28}$$

This implies the lemma. □

Applying the lemma to $B = A^*A$ hence shows that A is contracting on the π-orthogonal complement of the subspace generated by $\mathbf{1}$. On one hand, this is interesting as we will use it to bound the remaining eigenvalues of A, whose eigenvectors necessarily lie in the complement. On the other hand, it will allow us to prove convergence and rate bounds for dynamic matrices, supposing that all have the same Perron vector π and that the second eigenvalues of their multiplicative reversibilizations are uniformly bounded away from 1.

Lemma 5.19. *Let A be an ergodic matrix with Perron vector π. Then $\|Az\|_\pi \leq \sqrt{\varrho_2(A^*A)}\, \|z\|_\pi$ for all z with $\langle z, \mathbf{1}\rangle_\pi = 0$.*

The fact that all eigenvectors of nonzero eigenvalues are, in fact, π-orthogonal to $\mathbf{1}$ thus allows us to upper bound the second eigenvalue of A in terms of the second eigenvalue of its multiplicative reversibilization:

Lemma 5.20. *Let A be an ergodic matrix. Then $\varrho_2(A) \leq \sqrt{\varrho_2(A^*A)}$.*

If A is reversible with positive diagonal, then equality holds.

Proof. Every eigenvector z of A not corresponding to eigenvalue 1 is π-orthogonal to $\mathbf{1}$: If $Az = \lambda z$ with $\lambda \neq 1$, then

$$\langle z, \mathbf{1}\rangle_\pi = \langle z, A^*\mathbf{1}\rangle_\pi = \langle Az, \mathbf{1}\rangle_\pi = \lambda \langle z, \mathbf{1}\rangle_\pi \tag{5.29}$$

shows that $\langle z, \mathbf{1}\rangle_\pi = 0$.

In particular, if nonzero, this applies to an eigenvector z corresponding to an eigenvalue of absolute value $\varrho_2(A)$. Hence, $\varrho_2(A) = \|Az\|_\pi / \|z\|_\pi \leq \sqrt{\varrho_2(A^*A)}$ by Lemma 5.19. If $\varrho_2(A)$ is zero, then the inequality is trivial.

Now assume that A is reversible with positive diagonal. Then $A^*A = A^2$ is ergodic. By Theorem 5.3, we have

$$\varrho_2(A) = \lim_{t\to\infty} \|A^t - \mathbf{1} \cdot {}^t\pi\|^{1/t} = \lim_{t\to\infty} \|A^{2t} - \mathbf{1} \cdot {}^t\pi\|^{1/2t} = \varrho_2(A^2)^{1/2} \ , \tag{5.30}$$

which concludes the proof. □

We can now assemble the lemmas with Theorem 5.13 for the reversible case to give an extension to non-reversible ergodic matrices. Compared to the reversible case, we lose a factor of $\hat{\alpha}/2$ in the bound on the convergence rate.

Theorem 5.21. *Let A be the averaging matrix of a constant synchronous setting with communication digraph diameter D and self-confidence at least $\hat{\alpha} > 0$. If π is the Perron vector of A and β is the minimal positive value of $\pi_i A_{i,j}$, then the setting's spectral gap is at least $\min\{\hat{\alpha}\beta/4D, \hat{\alpha}^2\}$.*

In particular, it is at least $\pi_{\min}\alpha^2/4D$ where α is the minimal confidence and $\pi_{\min}$ is the minimal entry of the Perron vector π.

Proof. Set $A' = A^*A$, $\hat{\alpha}' = \hat{\alpha}^2$, $H' = G(A)$, and $\beta' = \hat{\alpha}\beta$. We want to apply Theorem 5.13 to A'. We have already established that A' is ergodic and reversible, as well as the fact that $H = G(A)$ is a sub-digraph of $G(A')$. Because A is ergodic, the graph H is strongly connected; it has diameter D. The diagonal entries of A' are at least $\hat{\alpha}'$ because

$$A'_{i,i} = (A^*A)_{i,i} = \sum_k \frac{A_{k,i}\pi_k}{\pi_i} A_{k,i} \geq A_{i,i}^2 \geq \hat{\alpha}^2 = \hat{\alpha}' \ . \tag{5.31}$$

Now let (i,j) be an edge of H', i.e., $\pi_i A_{i,j} \geq \beta$. Then

$$\pi_i A'_{i,j} = \pi_i (A^*A)_{i,j} = \sum_k \pi_k A_{k,i} A_{k,j} \geq A_{i,i}\pi_i A_{i,j} \geq \hat{\alpha}\beta \ . \tag{5.32}$$

Theorem 5.13 is hence applicable and shows

$$1 - \varrho_2(A) \geq 1 - \sqrt{1 - \min\{\beta'/2D, 2\hat{\alpha}'\}} \geq \min\{\beta'/4D, \hat{\alpha}'\} \tag{5.33}$$

with Lemma 5.20 and the elementary inequality $\sqrt{1-x} \leq 1 - x/2$ for $x \in [0,1]$. □

As in the reversible case, we can give a bound only in terms of smallest confidence, the diameter, and the number of agents, using Lemma 5.16:

Corollary 5.22. *The spectral gap of a constant synchronous setting with self-confidence is at least $\alpha^{D+2}/4Dn$. where α is the minimal confidence, D is the communication digraph's diameter, and n is the number of agents.*

Example 5.23 (Equal neighbor settings with Eulerian digraphs)**.** A Eulerian digraph is one in which every node's in-degree is equal to its out-degree. Let $G = (V, E)$ be a Eulerian digraph with self-loops. Let A be the equal neighbor matrix for G. Like in the bi-directional case, A is ergodic. Furthermore, its Perron vector is again given by $\pi_i = d(i)/|E|$. But, since $G(A) = G$ is not bi-directional, A is not reversible.

The diagonal entries of A are at lower-bounded $\hat{\alpha} = 1/n$ where n is the number of nodes in G. For all edges $(i,j) \in E$, we have $\pi_i \cdot A_{i,j} = 1/|E| \geq \beta = 1/n^2$. Theorem 5.13 hence gives the bound $1 - \varrho_2(A) \geq 1/4Dn^3 \geq 1/4n^4$ where D is the diameter of G.

However, this bound is not tight. Cover time methods are able to show a cubic bound of $1 - \varrho_2(A) = \Omega(1/n^3)$ like in the bi-directional case. Theorem 5.21 is nonetheless useful, also for equal-neighbor matrices in Eulerian digraphs, if we look at dynamic settings, which we do in the following subsection.

5.3.3 Dynamic Settings with Constant Perron Vector

In this subsection, we show how to can generalize the proofs for a constant setting to dynamic settings if their matrices' Perron vector is constant. The main tool is the fact that the multiplicative reversibilization gives a bound on the contraction constant in the orthogonal complement of the Perron vector. Constant Perron vectors occur under the popular assumption that the averaging matrices be doubly stochastic. A non-coordinator based technique to achieve doubly is stochastic matrices are fixed confidence equal-neighbor algorithms.

Theorem 5.24. *Let $A(1), A(2), \dots$ be the averaging matrices of a synchronous setting with the same Perron vector π. Further suppose that all communication digraphs are strongly connected and that there exists some $\varrho < 1$ such that $\varrho_2\big(A(t)^* A(t)\big) \leq \varrho$ for all $t \in \mathbb{N}_0$.*

Then the setting reaches asymptotic consensus and its rate of convergence is at most $\sqrt{\varrho}$.

Proof. Let y be any vector. We will show that $\lim_{t\to\infty} \left\|P(t)y - \mathbf{1}^t \pi y\right\|_\pi^{1/t} \leq \sqrt{\varrho}$, which then concludes the proof.

We can write $y = \alpha\mathbf{1} + \beta z$ where $\langle z, \mathbf{1}\rangle_\pi = 0$. Then,

$$\left\|P(t)y - \mathbf{1}^t \pi y\right\|_\pi = |\beta| \cdot \|P(t)z\|_\pi \leq |\beta| \cdot \varrho^{t/2} \cdot \|z\|_\pi \tag{5.34}$$

by repeated application of Lemma 5.19 because $\langle P(t)z, \mathbf{1}\rangle_\pi = 0$ for all $t \in \mathbb{N}_0$. □

This theorem, together with Theorems 5.13 and 5.21, provides a means to show asymptotic consensus and to bound the rate of convergence of synchronous dynamic settings whose averaging matrices are all ergodic and have a common Perron vector. A particular application for equal neighbor matrices is the following corollary. It is a generalization of Theorem 5.10 from bi-directional to Eulerian digraphs.

Corollary 5.25. *Let $A(1), A(2), \dots$ be the averaging matrices of a synchronous fixed confidence equal-neighbor setting whose communication digraphs are connected and Eulerian. Then the setting reaches asymptotic consensus and its rate of convergence is at most $1 - c/4Dn^2$ where n is the number of agents, c is the smallest confidence, and D is the maximum diameter.*

If all digraphs are bi-directional, then the rate of convergence is at most $1 - c/4Dn$.

Proof. For the Eulerian case, we use Theorem 5.21 with $\hat{\alpha} = c$ and $\beta \geq c/n$ and then Theorem 5.24. For the bi-directional case, we use Theorem 5.13, then Lemma 5.20, and then Theorem 5.24. □

Combining the theorem with our general lower bound in Corollary 5.22, we get:

Corollary 5.26. *Let $A(1), A(2), \dots$ be the averaging matrices of a synchronous setting having the same Perron vector and whose communication digraphs are strongly connected, have self-loops, and the minimal confidence is at least $\alpha > 0$. Then the setting reaches asymptotic consensus and its rate of convergence is at most $1 - \alpha^{D+2}/4nD$ where n is the number of agents and D is the maximum diameter.*

If the matrices are *doubly stochastic*, i.e., also their columns sum to 1, then their Perron vector is uniform $\pi_i = 1/n$. In fact, this is a characterization of doubly stochastic matrices. Doubly stochastic matrices are widely used in engineering applications. We can give a bound on their rate of convergence:

Corollary 5.27. *Let $A(1), A(2), \dots$ be the averaging matrices of a synchronous setting that are doubly stochastic and whose communication digraphs are strongly connected, have self-loops, and the minimal confidence is at least $\alpha > 0$. Then the setting reaches asymptotic consensus and its rate of convergence is at most $1 - \alpha^2/4nD$ where n is the number of agents and D is the maximum diameter.*

5.3.4 Worst-Case Lower Bound

We have seen worst-case bounds on the convergence rate bounding it away from 1 with a term exponential in n, e.g., Corollary 5.26. In the next section, we will see a wider class of dynamic settings in which the convergence rate is bounded away by a term of the order α^n. In this subsection, we present an example showing that there exists even a *constant* setting that indeed has a convergence rate exponentially close to 1.

To prove the lower bound of the example, we need a result that lower bounds $\varrho_2(A)$ in terms of the bottleneck ratio of A.

The bottleneck ratio is also referred to as Cheeger constant or conductance. It measures the minimal normalized weight of outgoing edges from a set of nodes in $G(A)$.

Definition 5.28 (Bottleneck ratio [62, Section 7.2])**.** Let A be an ergodic matrix in $\mathbb{R}^{n\times n}$ with Perron vector π and let $S \subseteq [n]$. The *bottleneck ratio* of S in A is defined as

$$\Phi_S(A) = \pi(S)^{-1} \sum_{i\in S} \sum_{j\in[n]\setminus S} \pi_i A_{i,j} \ . \tag{5.35}$$

The *bottleneck ratio* of the matrix A is defined as the minimal bottleneck ratios of all S with $\pi(S) \leq 1/2$, i.e.,

$$\Phi(A) = \min_{\substack{S\subseteq[n] \\ \pi(S)\leq 1/2}} \Phi_S(A) \ . \tag{5.36}$$

A proof of the following theorem can be found in textbooks (e.g., [62, Theorem 7.2 and Theorem 12.3]).

Theorem 5.29. *Let A be an ergodic matrix with Perron vector π. If $\pi_{\min}$ denotes the minimal entry of π, then*

$$1 - \varrho_2(A) \leq 4 \cdot \Phi(A) \cdot \log \frac{4}{\pi_{\min}} \ . \tag{5.37}$$

Example 5.30. The following example was described by Chung [29, Proof of Lemma 6.3] to show that the bottleneck ratio can be exponentially small in the size of A. More specifically, it is an ergodic matrix in $\mathbb{R}^{n\times n}$ with minimal positive entry $\alpha = 1/2$ whose spectral gap is at most $\alpha^{\Omega(n)}$, i.e., exponential in n. Chung used the bottleneck ratio to bound the eigenvalues of the Laplacian. We, on the other hand, use it in conjunction with Theorem 5.29 to bound the eigenvalues of the matrix itself, and not of its Laplacian. This allows us to directly infer a lower bound on its convergence rate.

The example is an equal-neighbor matrix. Its digraph is heavily non-bidirectional in the sense that there is a node whose difference between in-degree and out-degree is in the order of n. The digraph has to be non-bidirectional because otherwise its spectral gap would be at least n^{-3}, which is far away from being exponential in n. Nonetheless, all nodes have out-degree 2, which means that all matrix entries are either 0 or $1/2$.

The digraph is depicted in Figure 5.4. It has $n = 2m$ nodes and consists of two isomorphic parts that are connected via two anti-parallel edges. We will list the edges between the nodes $1, 2, \ldots, m$, which also determine the edges between the nodes $m+1, m+2, \ldots, 2m$ via the isomorphism $\bar{i} = 2m - i + 1$. The digraph also contains the two anti-parallel edges $(m, \bar{m})$ and $(\bar{m}, m)$. The edges between the nodes $1, 2, \ldots, m$ are: (a) the edges $(i, i+1)$ for all $i < m$ and (b) the edges $(i, 1)$ for all $i \in \{1, 2, \ldots, m\}$.

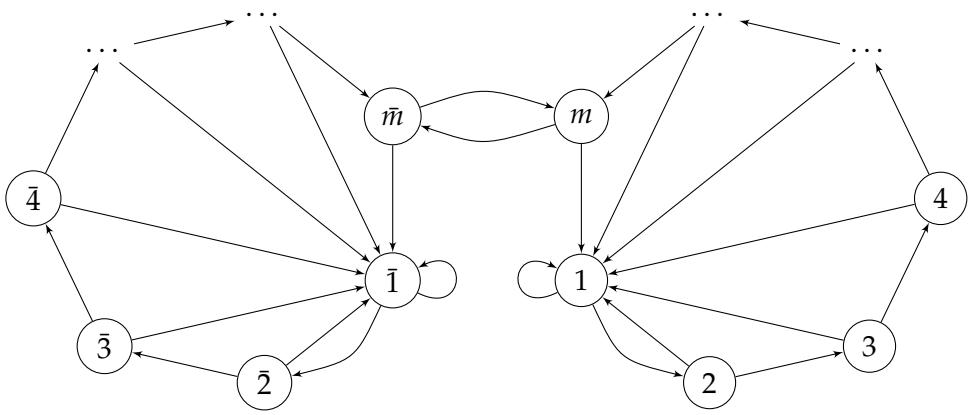


Figure 5.4: Chung's example of a digraph with small bottleneck ratio

We can verify that the Perron vector is given by

$$\pi_i = \frac{1}{2^{i+1}} \text{ for } i \in \{1, 2, \ldots, m-1\} \quad \text{and} \quad \pi_m = \frac{1}{2^m} \ . \tag{5.38}$$

By symmetry, this also defines the Perron vector for the remaining indices between $m+1$ and $2m$ since $\pi_i = \pi_{2m-i+1}$. We note the particular value $\pi_m = 1/2^m$, which is exponentially small in m, and hence in n. Choosing $S = \{1, 2, \ldots, m\}$, we have $\pi(S) = 1/2$ by symmetry, and its bottleneck ratio is equal to $\Phi_S(A) = 2\pi_m \cdot A_{m,\bar{m}} = 1/2^m$. Hence the bottleneck ratio of A is at most $\Phi(A) \leq 1/2^m$. The smallest entry of its Perron vector is $\pi_{\min} = \pi_m = 1/2^m$. Hence Theorem 5.29 gives the bound

$$1 - \varrho_2(A) \leq \frac{1}{2^{m-2}} \cdot \log 2^{m+2} = O\left(\frac{m}{2^m}\right) \tag{5.39}$$

on the spectral gap.

5.4 Asymptotic Consensus in Dynamic Settings with Aperiodic Core

In this section, we study asymptotic consensus within dynamic settings, i.e., settings in which the averaging matrices $A(t)$ and delay vectors $\delta(t)$ depend on t. We give both convergence conditions and upper-bound the rate of convergence. We start the chapter by studying synchronous settings, which translate to products $P(t)$ of stochastic matrices $A(t)$. We treat B-bounded settings as special cases of our results on synchronous settings via the reduction of B-bounded to synchronous settings.

5.4.1 Coefficient of Ergodicity

Definition 5.31 (Coefficients of ergodicity [82, Definition 4.6])**.** Let $\mathcal{S}_n$ be the set of stochastic matrices in $\mathbb{R}^{n\times n}$. A *coefficient of ergodicity* on $\mathbb{R}^{n\times n}$ is a continuous mapping $\mu : \mathcal{S}_n \to [0, 1]$. It is *proper* if $\mu(A) = 0$ if and only if A has rank 1, i.e., is of the form $A = \mathbf{1} \cdot {}^t v$ for some stochastic vector v.

The use of a coefficient of ergodicity μ shows that $\mu\big(P(t)\big)$ converges to zero as $t \to \infty$. Much of the complication in the proofs comes from the need to show some form of sub-multiplicativity for proving convergence to zero.

Theorem 5.32 (Hajnal [51, Theorem 2])**.** *The two mappings*

$$\delta(A) = \max_j \max_{i_1,i_2} |A_{i_1,j} - A_{i_2,j}| \tag{5.40}$$

and

$$\lambda(A) = 1 - \min_{i_1,i_2} \sum_j \min\{A_{i_1,j}, A_{i_2,j}\} \ . \tag{5.41}$$

are proper coefficients of ergodicity. Furthermore, we have $\delta(AB) \leq \lambda(A)\delta(B)$.

It is natural to ask what this technique yields when applied to a constant sequence $A(t) = A$, i.e., powers of a stochastic matrix A. The next theorem is a result by Hajnal that answers this question. It characterizes the matrices A for which Theorem 5.32 is able to show ergodicity of $P(t) = A^t$.

Definition 5.33 (Scrambling matrices [51])**.** A stochastic matrix A in $\mathbb{R}^{n\times n}$ is *scrambling* if for all $i_1, i_2 \in [n]$ there exists some $j \in [n]$ such that both $A_{i_1,j} > 0$ and $A_{i_2,j} > 0$.

Theorem 5.34 (Hajnal [51])**.** *A stochastic matrix A is scrambling if and only if $\lambda(A) < 1$.*

We will further provide sufficient conditions on the averaging matrices of synchronous and B-bounded settings for reaching asymptotic consensus in *all* executions; that is, independent of the agents' initial values. Since this also includes the standard basis vectors, in synchronous settings, this is equivalent for the sequence of matrix products $P(t)$ converging and the limit being a rank 1 matrix. For this, we will use a coefficient of ergodicity that measures how far a stochastic matrix is from having rank 1. It will turn out to be equal to the coefficient λ defined by Hajnal [51], but we introduce it in a different manner, namely in form of a semi-norm, that establishes its sub-multiplicativity; a fact that is not obvious from Hajnal's definition. In fact, authors like Hajnal, Wolfowitz, and Chatterjee and Seneta used a substitute for sub-multiplicativity involving a second coefficient, see Theorem 5.32.

An execution $x(t)$ reaches asymptotic consensus if and only if the limit $x^* = \lim_{t\to\infty} x(t)$ exists and all of its entries are equal, i.e., $x^* = c^* \cdot \mathbf{1}$ for some scalar $c^* \in \mathbb{R}$. Thus, a necessary condition for reaching asymptotic consensus is that the distance of $x(t)$ to the vector space $\langle \mathbf{1} \rangle$ generated by $\mathbf{1}$ tends to zero. We will show, at least for executions in synchronous and B-bounded settings, that this condition is actually also sufficient for reaching asymptotic consensus.

We choose the infinity norm to define the distance to the subspace $\langle \mathbf{1} \rangle$. This distance is actually a vector semi-norm, which we will use to show that the resulting matrix semi-norm is sub-multiplicative. We then show that the matrix semi-norm is equal to the coefficient of erogdicity λ, which shows sub-multiplicativity of λ. We choose a normalization factor of 2 to have the additional property that the distance is equal to the maximum distance between vector entries. We hence define

$$\|x\|_\perp = 2 \inf_{c\in\mathbb{R}} \|x - c \cdot \mathbf{1}\|_\infty \ . \tag{5.42}$$

Lemma 5.35. *The following propositions are true.*

1. *The mapping* $x \mapsto \|x\|_\perp$ *is a semi-norm on* $\mathbb{R}^n$.
2. $\|x\|_\perp = x_{\max} - x_{\min}$.

We now prove that $\|x(t)\|_\perp \to 0$ is necessary and sufficient for $x(t)$ to converge to a multiple of $\mathbf{1}$. For a vector $x \in \mathbb{R}^n$, we denote by $\operatorname{hull}(x)$ the convex hull of the set $\{x_i \mid i \in [n]\}$ of its entries. More concretely, $\operatorname{hull}(x)$ is the interval $[x_{\min}, x_{\max}]$ where $x_{\min}$ is the minimal and $x_{\max}$ the maximum entry of x. For an interval $I \subseteq \mathbb{R}$, we write $\operatorname{len}(I)$ for its length. In our case, $\operatorname{len}(\operatorname{hull}(x)) = x_{\max} - x_{\min}$.

Because every entry of Ax is a convex combination of the x_i, we get:

Lemma 5.36. *If A is a stochastic matrix in $\mathbb{R}^{n\times n}$ and $x \in \mathbb{R}^n$, then* $\operatorname{hull}(Ax) \subseteq \operatorname{hull}(x)$.

We can now prove our claimed characterization.

Theorem 5.37. *Let $A(1), A(2), \dots$ be a sequence of stochastic matrices in $\mathbb{R}^{n\times n}$ and let $x \in \mathbb{R}^n$. If the sequence of vectors defined by $x(0) = x$ and $x(t) = A(t) \cdot x(t-1)$ converges to some vector $x^* = c^* \cdot \mathbf{1} \in \langle \mathbf{1} \rangle$, then*

$$\frac{1}{2}\|x(t)\|_\perp \leq \|x(t) - x^*\|_\infty \leq \|x(t)\|_\perp \tag{5.43}$$

for all $t \in \mathbb{N}_0$.

Proof. Since $\operatorname{hull}(x^*) = \{c^*\}$ and $\operatorname{hull}(x^*) \subseteq \operatorname{hull}(x(t))$ for all $t \in \mathbb{N}_0$ by the second part of Lemma 5.36, it is $c^* \in \operatorname{hull}(x(t))$ for all $t \in \mathbb{N}_0$. In particular, $|x_i(t) - c^*| \leq \operatorname{len}(\operatorname{hull}(x(t)))$ for all $i \in [n]$ since $x_i(t) \in \operatorname{hull}(x(t))$ by definition. Hence

$$\|x(t) - x^*\|_\infty = \max_{i\in[n]} |x_i(t) - c^*| \leq \operatorname{len}(\operatorname{hull}(x(t))) = \|x(t)\|_\perp \tag{5.44}$$

by the first part of Lemma 5.36. The inequality $\|x(t)\|_\perp / 2 \leq \|x(t) - c^* \cdot \mathbf{1}\|_\infty = \|x(t) - x^*\|_\infty$ holds by the definition of $\|x(t)\|_\perp$ as an infimum. □

Because the convex hull of the agents' values in the reduced execution $y(t)$ of a B-bounded settings to a synchronous settings is equal to the convex hull of the values in $x(t), x(t-1), \dots, x(t-B+1)$, we have $\|y(t)\|_\perp \to 0$ if and only if $\|x(t-d)\|_\perp \to 0$ for all $0 \leq d \leq B-1$, which in turn is equivalent to $\|x(t)\|_\perp \to 0$ because of monotonicity. We hence have:

Corollary 5.38. *An execution $x(t)$ of a B-bounded setting reaches asymptotic consensus if and only if $\|x(t)\|_\perp \to 0$.*

Starting from the vector semi-norm, we define a matrix semi-norm by mimicking the definition of the operator norm. More explicitly, for a matrix $A \in \mathbb{R}^{n\times n}$, we set

$$\|A\|_\perp = \sup_{\substack{x\in\mathbb{R}^n \\ \|x\|_\perp \neq 0}} \frac{\|Ax\|_\perp}{\|x\|_\perp} \ . \tag{5.45}$$

Lemma 5.39. *The following propositions are true.*

1. *The mapping $A \mapsto \|A\|_\perp$ is a sub-multiplicative semi-norm on $\mathbb{R}^{n\times n}$.*
2. *If A is stochastic, then $\|A\|_\perp = 0$ if and only if* $\operatorname{img} A = \langle \mathbf{1} \rangle$.

The following theorem characterizes convergence of the semi-norm of an infinite matrix product of stochastic matrices to zero. It shows that it is equivalent that the product converges to a rank 1 matrix.

Theorem 5.40. *Let $\big(A(t)\big)$ be a sequence of stochastic matrices. The sequence of products $P(t) = A(t) \cdot A(t-1) \cdots A(1)$ converges to a stochastic matrix of rank 1 if and only if $\|P(t)\|_\perp \to 0$.*

Proof. If $P(t)$ converges to a rank 1 stochastic matrix P, then $\|P\|_\perp = 0$. The mapping $A \mapsto \|A\|_\perp$ is continuous because

$$\big|\, \|A\|_\perp - \|B\|_\perp \,\big| \le \|A - B\|_\perp \le \|A - B\|_\infty \tag{5.46}$$

by the reverse triangle inequality and because $\|C\|_\perp \le \|C\|_\infty$. Hence $\|P(t)\|_\perp \to \|P\|_\perp = 0$.

To prove the converse implication, we show that $P(t) \cdot x$ is a Cauchy sequence for all $x \in \mathbb{R}^n$ if $\|P(t)\|_\perp \to 0$. This then concludes the proof because then the limit $P = \lim P(t)$ exists and is of rank at most 1 by the second part of Lemma 5.39 because $\|P\|_\perp = 0$. Every $P(t)$ is stochastic because the product of two stochastic matrices is stochastic. The condition $\forall i\colon \sum_{j=1}^n P_{i,j}(t) = 1$ is preserved when taking the limit, hence P is stochastic, and of rank 1 because stochastic matrices are nonzero.

So let $x \in \mathbb{R}^n$ and $\varepsilon > 0$. Because $\|P(t) \cdot x\|_\perp \to 0$, there exists a T such that $\|P(T) \cdot x\|_\perp \le \varepsilon$. Let $c \in \mathbb{R}$ such that $\|P(T) \cdot x\|_\perp = 2\|P(T) \cdot x - c \cdot \mathbf{1}\|_\infty$. Then, denoting $P(t,T) = P(t) \cdot P(t-1) \cdots P(T+1)$, for every $t \ge T$, we have

$$\begin{aligned} \|P(t) \cdot x - P(T) \cdot x\|_\infty &\le \|P(t,T) \cdot P(T) \cdot x - c \cdot \mathbf{1}\|_\infty + \|P(T) \cdot x - c \cdot \mathbf{1}\|_\infty \\ &\le \|P(t,T) \cdot (P(T) \cdot x - c \cdot \mathbf{1})\|_\infty + \varepsilon/2 \\ &\le \|P(T) \cdot x - c \cdot \mathbf{1}\|_\infty + \varepsilon/2 \le \varepsilon \end{aligned} \tag{5.47}$$

because $\|P(t,T)\|_\infty \le 1$ since it is stochastic. This concludes the proof. □

Corollary 5.41. *A synchronous setting with averaging matrices $A(t)$ reaches asymptotic consensus in all executions if and only if $\|P(t)\|_\perp \to 0$.*

We now give different expressions for it. In particular, we show that it is equal to the coefficient of ergodicity λ used by Hajnal.

Lemma 5.42. *Let A be a stochastic matrix in $\mathbb{R}^{n \times n}$. The following equalities for $\|A\|_\perp$ are true.*

1. $\|A\|_\perp = \max_{x \in \{0,1\}^n} \|Ax\|_\perp$
2. $\|A\|_\perp = \max_{i_1, i_2 \in [n]} \sum_{j=1}^n (A_{i_1,j} - A_{i_2,j})_+$ *where $(z)_+ = \max\{z, 0\}$ is the positive part of $z \in \mathbb{R}$*
3. $\|A\|_\perp = 1 - \min_{i_1, i_2 \in [n]} \sum_{j=1}^n \min\{A_{i_1,j}, A_{i_2,j}\}$

Proof. By homogeneity of the vector semi-norm, we can restrict the supremum in the definition of $\|A\|_\perp$ to all x with $\|x\|_\perp = 1$, i.e., $\mathrm{len}(\mathrm{hull}(x)) = 1$. Further, if we denote the minimal entry of x by $x_{\min}$ and set $x' = x - x_{\min} \cdot \mathbf{1}$, we have $\|x'\|_\perp = \|x\|_\perp = 1$ and $\|Ax'\|_\perp = \|Ax\|_\perp$.

We can hence restrict the supremum to all x with $\operatorname{hull}(x) = [0,1]$. Using the third part of Lemma 5.35, we thus have

$$\begin{aligned} \|A\|_\perp &= \sup_{\operatorname{hull}(x)=[0,1]} \|Ax\|_\perp = \sup_{\operatorname{hull}(x)=[0,1]} \max_{i_1,i_2\in[n]} \big(\sum_{j=1}^{n} A_{i_1,j}x_j - \sum_{j=1}^{n} A_{i_2,j}x_j\big) \\ &= \max_{i_1,i_2\in[n]} \sup_{\operatorname{hull}(x)=[0,1]} \sum_{j=1}^{n} (A_{i_1,j} - A_{i_2,j})x_j \ . \end{aligned} \tag{5.48}$$

Clearly, the supremum in the last expression in (5.48) is attained by the vector $x \in \mathbb{R}^n$ by setting $x_j = 1$ if $A_{i_1,j} - A_{i_2,j} > 0$ and $x_j = 0$ otherwise. This shows the second claimed equality. Because $\|Ax\|_\perp$ is equal to the last expression in (5.48), we have also shown the first claimed equality.

Starting from the second claimed equality, we get

$$\begin{aligned} \|A\|_\perp &= \max_{i_1,i_2\in[n]} \sum_{j=1}^{n} \max\{A_{i_1,j} - A_{i_2,j}\,,\, 0\} = \max_{i_1,i_2\in[n]} \sum_{j=1}^{n} \big(A_{i_1,j} + \min\{-A_{i_2,j}, -A_{i_1,j}\}\big) \\ &= 1 + \max_{i_1,i_2\in[n]} \sum_{j=1}^{n} \max\{-A_{i_1,j}, -A_{i_2,j}\} = 1 - \min_{i_1,i_2\in[n]} \sum_{j=1}^{n} \min\{A_{i_1,j}\,,\, A_{i_2,j}\} \end{aligned} \tag{5.49}$$

because $\sum_{j=1}^{n} A_{i_1,j} = 1$ and $\max(-\mathcal{A}) = -\min \mathcal{A}$. This shows the third claimed equality. □

5.4.2 Proving Convergence with the Semi-norm

The semi-norm of a stochastic matrix is at most one. The following lemma gives a sufficient condition for the case that it is strictly less than one, and gives a means to bound the distance to one.

Theorem 5.43. *Let $A \in \mathbb{R}^{n\times n}$ be a stochastic matrix, $\alpha \geq 0$, and $j_0 \in [n]$. Suppose that all of A's entries in the $j_0{}^{th}$ column are at least α, i.e., $A_{i,j_0} \geq \alpha$ for all $i \in [n]$. Then $\|A\|_\perp \leq 1 - \alpha$.*

Proof. Write $A = E + B$ where $E_{i,j} = \alpha$ if $j = j_0$ and $E_{i,j} = 0$ otherwise. Because the image of E is contained in $\langle \mathbf{1} \rangle$, we have $\|E\|_\perp = 0$ and hence $\|A\|_\perp = \|B\|_\perp$ by the triangle inequality. The entries of B are all nonnegative and its row sums $\sum_j B_{i,j}$ are all equal to $1 - \alpha$.

Let $x \in \mathbb{R}^n$. For every $c \in \mathbb{R}$, we have:

$$\begin{aligned} \frac{1}{2}\|Bx\|_\perp &\leq \|Bx - (1-\alpha)c\cdot\mathbf{1}\|_\infty = \max_{i\in[n]} \big|\big(\sum_{j=1}^{n} B_{i,j}x_j\big) - (1-\alpha)c\big| \\ &= \max_{i\in[n]} \big|\sum_{j=1}^{n} B_{i,j}\cdot(x_j - c)\big| \leq \max_{i\in[n]} \sum_{j=1}^{n} B_{i,j}\cdot|x_j - c| \\ &\leq \max_{i\in[n]} \sum_{j=1}^{n} B_{i,j}\cdot\|x - c\cdot\mathbf{1}\|_\infty = (1-\alpha)\cdot\|x - c\cdot\mathbf{1}\|_\infty \end{aligned} \tag{5.50}$$

Forming the infimum over all $c \in \mathbb{R}$ now shows that $\|Bx\|_\perp \leq (1-\alpha)\cdot\|x\|_\perp$ and hence $\|A\|_\perp = \|B\|_\perp \leq 1 - \alpha$. □

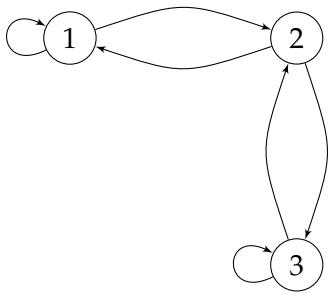


Figure 5.5: Digraph $G(A)$ of matrix A in Example 5.44

Example 5.44. We now give an example of a matrix whose semi-norm is strictly less than 1, but that does not have a strictly positive column, i.e., Theorem 5.43 cannot be used to prove this. The matrix is equal to

$$A = \begin{pmatrix} 1/2 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 \end{pmatrix} \tag{5.51}$$

and its digraph is depicted in Figure 5.5. Using the first expression in Lemma 5.42, one can see that the semi-norm $\|A\|_\perp$ is equal to $1/2$.

However, if the sequence of powers A^t, or the sequence of products $P(t)$, or any sequence of matrices $B(t)$ really, converges to a rank 1 stochastic matrix B, then necessarily some column is eventually positive. This is because every rank 1 stochastic matrix can be written as $B = \mathbf{1} \cdot {}^t v$ where v is a probability vector. A probability vector never being zero, it has some positive entry, say, v_{j_0}. But for all i, the sequence $B_{i,j_0}(t)$ converges to v_{j_0}, i.e., is eventually positive at some point.

Hence, while the hypothesis of Theorem 5.43 is not necessary for A^t converging to a rank 1 matrix, the hypothesis that *some* power of A fulfills it is both necessary and sufficient. We develop this argument more carefully in the following section.

The following lemma characterizes positivity of entries in products of stochastic matrices solely in terms of the matrices' associated digraphs. It should be noted that, because we study backward products, the walks grow at the start node and not at the end node.

Lemma 5.45. *Let $0 \leq s \leq t$ and $i, j \in [n]$. Then $P_{i,j}(t,s)$ is positive if and only if there exist $i_t, i_{t-1}, \ldots, i_s \in [n]$ with $i_t = i$ and $i_s = j$ such that $(i_\tau, i_{\tau-1})$ is an edge of $G\big(A(\tau)\big)$ for all $s+1 \leq \tau \leq t$.*

5.4.3 Aperiodic Cores

Classically, in asymptotic consensus, self-confidence of the agents is assumed. That is, every communication digraph contains self-loops at all nodes. This can model the fact that an agent does not ignore or forget its own previous value. We generalize the existence of self-loops, however: A missing self-loop in a specific communication digraph can model memory loss of an agent. We replace the assumption of self-loops to *aperiodic cores*, which are sub-digraphs of all of the settings' communication digraphs. They can be seen as a "distributed safety

net against memory loss". In this sense, existence of self-loops is the assumption of a non-distributed safety measure against memory loss or temporary self-distrust. Their function in the proofs is similar to that of self-loops, but they are more general. A parameter that we use over and over in our results is that of the index of convergence of the aperiodic core. If one assumes self-loops, then this parameter is equal to 0. So, in our theorem statements, if one assumes self-confidence, then $\operatorname{ind}(H) = 0$.

We call a node j in a digraph G a *leader* of another node i if G contains a path from i to j. A digraph is *j-coordinated* if j is a leader of every node. In this case, node j is called a *leader* of G. A digraph is *coordinated* if it is j-coordinated for some j. If j is a node of a digraph G, we say that G is *j-aperiodic* if j's strongly connected component in G is primitive. A digraph H is a *core* of a sequence $G_1, G_2, \dots$ of digraphs if H is a sub-digraph of every G_t.

5.4.4 Coordinated Aperiodic Cores

We start with assuming that there is a core that is coordinated and leader-aperiodic. The assumption of a core in particular applies if the communication digraph is constant. We hence get a direct generalization of the constant ergodic case:

Theorem 5.46. *A synchronous setting with averaging matrices $A(t)$ with spanning core H and minimal confidence α reaches asymptotic consensus if there exists some agent j_0 such that H is j_0-coordinated and j_0-aperiodic. Moreover, the rate of convergence is at most $1 - \alpha^{\operatorname{ind}(H)} / \operatorname{ind}(H)$.*

We prove this theorem in the rest of the subsection.

In general, given a sequence of stochastic matrices $A(1), A(2), \dots$ in $\mathbb{R}^{N\times N}$ and a node $j \in [N]$, we define $S_j(t,s)$ to be the set of indices $i \in [N]$ such that $P_{i,j}(t,s)$ is positive. Denote by $\mu_j(t,s)$ the smallest (positive) $P_{i,j}(t,s)$ with $i \in S_j(t,s)$. We also define $S_j(t) = S_j(t,0)$ and $\mu_j(t) = \mu_j(t,0)$.

It is easy to see that $\mu_j(t,s) \geq \alpha^{t-s}$ if α is the minimal confidence. This will be our main tool to bound the convergence rate: If $S_j(t,s) = [N]$, then $\|P(t,s)\|_\perp \leq 1 - \alpha^{t-s}$ by Theorem 5.43. And if we can show $S_j(t,s) = [N]$ whenever $t - s \geq T$ where T is some constant, then

$$\lim_{t\to\infty} \|P(t)\|_\perp^{1/t} = \lim_{k\to\infty} \|P(kT)\|_\perp^{1/kT} \leq (1-\alpha^T)^{1/T} \leq 1 - \alpha^T/T \ . \tag{5.52}$$

Because all hypotheses we consider are time-invariant, it is sufficient to show $S_j(T) = [N]$.

For Theorem 5.46, we choose $T = \operatorname{ind}(H)$: We show that $S_j(\operatorname{ind}(H)) = [N]$. This is done by reducing the problem to one with a constant matrix. So let A be any stochastic matrix whose digraph $G(A)$ is equal to H. If A^t has a positive column, then so does $P(t)$ because H is a sub-digraph of every communication digraph. This shows the claim since $\operatorname{ind}(G(A)) = \operatorname{ind}(H)$.

5.4.5 Clusterings

We pair the idea of the distributed safety net in form of an aperiodic core with the notion of *clusters*, which have a leader that is the sole agent of the cluster to regard values of agents other than the cluster's. We will prove that it is not necessary for every agent to be contained in an aperiodic component, but only for the cluster leaders.

A natural example of these clusterings occurs in the reduction of B-bounded settings with self-confidence to synchronous ones (see Figure 5.3), for which $\operatorname{ind}(H) = B - 1$. If we do

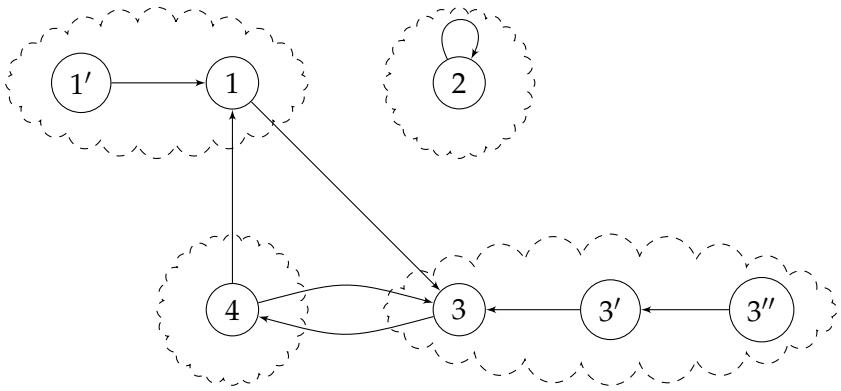


Figure 5.6: $\mathcal{C}$-aperiodic digraph with leaders $1, 2, 3, 4$

not assume self-confidence in B-bounded settings, then asymptotic consensus is not necessarily reached, even if the averaging matrices are constant and ergodic. By proving results on cluster-aperiodic cores in synchronous settings, we are hence also proving results on B-bounded settings with self-confidence.

A digraph is a *cluster* with leader l if it is l-coordinated. A *clustering* $\mathcal{C}$ is a collection of node-disjoint clusters $C_1, C_2, \ldots, C_n$ together with respective leaders $l_1, l_2, \ldots, l_n$. A digraph is *$\mathcal{C}$-aperiodic* if every cluster C_m is a sub-digraph, every node is contained in some cluster, and it is l-aperiodic for every leader l_m of $\mathcal{C}$. Figure 5.6 shows an example of a $\mathcal{C}$-aperiodic digraph.

A digraph *respects* a clustering $\mathcal{C}$ if the only edges leaving a cluster are the leader's. Given a digraph that respects clustering $\mathcal{C}$, the corresponding *cluster digraph* is the digraph when collapsing all clusters of $\mathcal{C}$ to single node.

5.4.6 Dynamic Coordinated Communication Digraphs

We now prove that asymptotic consensus is also reached if there is no coordinated core, but that coordination at every time step suffices.

Theorem 5.47. *A synchronous setting with averaging matrices $A(1), A(2), \ldots$ with a $\mathcal{C}$-aperiodic spanning core H and minimal confidence α reaches asymptotic consensus if every communication digraph respects $\mathcal{C}$ and is coordinated. Moreover, the rate of convergence is at most*

$$1 - \alpha^{(n-1)^2(\mathrm{ind}(H)+1)} / (n-1)^2(\mathrm{ind}(H)+1) \tag{5.53}$$

where n is the number of clusters in $\mathcal{C}$.

Corollary 5.48. *A B-bounded setting with averaging matrices $A(1), A(2), \ldots$ with self-confidence and minimal confidence α reaches asymptotic consensus if every communication digraph is coordinated. Moreover, the rate of convergence is at most $1 - \alpha^{(n-1)^2 B} / (n-1)^2 B$.*

Corollary 5.48, without the explicit bound on the rate of convergence is included in Theorem 5.9.

We prove the theorem in the rest of the subsection. Denote the number of nodes by N.

The sets $S_j(t)$ satisfy a weak form of monotonicity if the sequence of communication graphs have an aperiodic core. If there are self-loops in all communication digraphs, then clearly $S_j(t) \subseteq S_j(t+1)$, which is a special case of the following lemma.

Lemma 5.49. *If H is a spanning $\mathcal{C}$-aperiodic core and all communication digraphs respect $\mathcal{C}$, then $S_j(t_1) \subseteq S_j(t_2)$ whenever $t_2 - t_1 \geq \operatorname{ind}(H)$ and j is a leader of $\mathcal{C}$.*

Proof. Let $i \in S_j(t_1)$. Since all communication digraphs respect the clustering, i's leader l_i appears in some earlier set: $l_i \in S_j(t_1')$ with $t_1' \leq t_1$.

Because H is l_i-aperiodic and $t_2 - t_1' \geq \operatorname{ind}(H)$, there exists a walk of length $t_2 - t_1'$ from i to l_i in H by Theorem 2.1 and the definition of $\operatorname{ind}(H)$. The fact that H is a sub-digraph of all $G\big(A(\tau)\big)$ shows that $P_{i,l_i}(t_2, t_1')$ is positive by Lemma 5.45.

Hence

$$P_{i,j}(t_2) = \sum_k P_{i,k}(t_2, t_1') \cdot P_{k,j}(t_1') \geq P_{i,l_i}(t_2, t_1') \cdot P_{l_i,j}(t_1') \tag{5.54}$$

is positive, which shows $i \in S_j(t_2)$. □

The following lemmas are used to lower bound the steps need until $S_j(t) = [N]$.

Lemma 5.50. *If H is a spanning $\mathcal{C}$-aperiodic core, all communication graphs respect $\mathcal{C}$, j is a leader of $\mathcal{C}$, $t \geq \operatorname{ind}(H)$, and $G(A(t+1))$ is j-coordinated, then either $S_j(t) = [N]$ or $S_j(t+1) \setminus S_j(t) \neq \emptyset$.*

Proof. The hypothesis that $t \geq \operatorname{ind}(H)$ guarantees that $j \in S_j(t)$ by Lemma 5.49. Every node has a path to j, and hence to $S_j(t)$, in $G(A(t+1))$. Now, if $S_j(t) \neq [N]$, there is some $i \in [n] \setminus S_j(t)$ that has an outgoing neighbor k_0 in $S_j(t)$, i.e., $A_{i,k_0}(t+1) > 0$. The condition $k_0 \in S_j(t)$ means $P_{k_0,j}(t) > 0$ and hence

$$P_{i,j}(t+1) = \sum_k A_{i,k}(t+1) \cdot P_{k,j}(t-1) \geq A_{i,k_0}(t+1) \cdot P_{k_0,j}(t) > 0 \ , \tag{5.55}$$

which shows $i \in S_j(t+1)$. □

Lemma 5.51. *Let H be a spanning $\mathcal{C}$-aperiodic core, all communication graphs respect $\mathcal{C}$ and j be a leader of $\mathcal{C}$. If l is any leader of some cluster C of $\mathcal{C}$ and $l \in S_j(t)$, then $C \subseteq S_j(t + \operatorname{ind}(H))$.*

Proof. Because C is l-aperiodic and l-coordinated, we have $C \subseteq S_l(\tau)$ for all $\tau \geq \operatorname{ind}(C)$. Because $\operatorname{ind}(H) \geq \operatorname{ind}(C)$, the lemma follows with an application of Lemma 5.45. □

Set $t_m = m \cdot (\operatorname{ind}(H) + 1)$. For $m \geq 1$, let j_m be a leader of the digraph $G(A(t_m))$ and also of $\mathcal{C}$. Lemma 5.49 specialized to $s = t_{m-1}$ and $t = t_m - 1 = t_{m-1} + \operatorname{ind}(H)$ gives $S_j(t_m - 1) \supseteq S_j(t_{m-1})$ for all leaders j and all $m \geq 1$. Lemma 5.50 applied to $t = t_m$ and $j = j_m$ gives: $S_{j_m}(t_m) \supsetneq S_{j_m}(t_{m-1})$ if $S_{j_m}(t_m - 1) \neq [N]$.

If $m = (n-1)^2 = (n-2)n + 1$, then some $j_0 \in [n]$ appears at least $n - 1$ times in the sequence of leaders $j_1, j_2, \ldots, j_m$. By the above and Lemma 5.51, it is hence $S_{j_0}(t_m) = [N]$, which shows the theorem.

5.4.7 Dynamic Communication Digraphs with Fixed Leader

In this subsection, we assume a *fixed* leader in every communication digraph and are able to show a tighter bound on the rate of convergence. The case of strongly connected communication digraphs is a special case.

Theorem 5.52. *A synchronous setting with averaging matrices $A(1), A(2), \ldots$ with a $\mathcal{C}$-aperiodic spanning core H and minimal confidence α reaches asymptotic consensus if every communication digraph respects $\mathcal{C}$ and there is an agent j_0 such that every communication digraph is j_0-coordinated. Moreover, the rate of convergence is at most*

$$1 - \alpha^{(n-1)(\mathrm{ind}(H)+1)} / (n-1)(\mathrm{ind}(H)+1) \tag{5.56}$$

where n is the number of clusters in $\mathcal{C}$.

Corollary 5.53. *A B-bounded setting with averaging matrices $A(1), A(2), \ldots$ with self-confidence and minimal confidence α reaches asymptotic consensus if there is an agent j_0 such that every communication digraph is j_0-coordinated. Moreover, the rate of convergence is at most $1 - \alpha^{(n-1)B} / (n-1)B$.*

Corollary 5.53, without the explicit bound on the rate of convergence is included in Theorem 5.9.

We use the notation of the previous subsection. The theorem follows similarly by noticing that, in this case, $j_m = j_0$ for all $m \geq 1$ and hence j_0 appears $n-1$ times in the sequence of leaders $j_1, j_2, \ldots, j_{n-1}$.

5.4.8 Completely Reducible Communication Digraphs

We now show that one can replace the assumption of coordination by the assumption of completely reducibility at every time step and eventual weak connectivity.

Theorem 5.54. *A synchronous setting with averaging matrices $A(1), A(2), \ldots$ with a $\mathcal{C}$-aperiodic spanning core H and minimal confidence α reaches asymptotic consensus if every communication digraph respects $\mathcal{C}$, all cluster communication digraphs are completely reducible, and the digraph G_∞ formed by all edges that appear in infinitely many cluster communication digraphs is weakly connected.*

Corollary 5.55. *A B-bounded setting with averaging matrices $A(1), A(2), \ldots$ with self-confidence and minimal confidence α reaches asymptotic consensus if every communication digraph is completely reducible and the digraph G_∞ of edges that appear in infinitely many communication digraphs is weakly connected.*

Corollary 5.55 for synchronous settings is Theorem 5.7.

We prove this theorem in the rest of this subsection. We do not use the exact same proof strategy as in the previous subsection: We show the existence of a T such that

$$\|P(T)\|_\perp \leq 1 - \alpha^{n(\mathrm{ind}(H)+1)} \ . \tag{5.57}$$

This suffices to show the theorem because the conditions in the theorem are time-invariant and repeated application thus shows that $\|P(t)\|_\perp \to 0$. Even though we cannot bound T with the hypotheses of the theorem, we *can* bound the semi-norm uniformly, which is critical for the proof to work. Theorem 5.40 then concludes the proof.

We first show that G_∞ is completely reducible. For that, we show the following lemma.

Lemma 5.56. *Every union of completely reducible digraphs is completely reducible.*

Proof. Let $\mathcal{G}$ be a set of completely reducible digraphs and let $H = \bigcup \mathcal{G}$ be their union. Let i and j be two nodes in H and suppose that there exists a path P from i to j in the union digraph H. We will show that there then exists a path from j to i in H. This is trivial if $i = j$ so suppose the contrary, i.e., that P is nonempty.

Let $i_0, i_1, \ldots, i_n$ be P's sequence of nodes. For every $1 \leq k \leq n$, the edge e_k is in some digraph $G \in \mathcal{G}$. Now, because G is completely reducible, there exists a path P_k in G from e_k to e_{k-1}. But then the composite walk $P_n \cdot P_{n-1} \cdots P_1$ is a walk in H from j to i. □

Hence G_∞ is completely reducible because Lemma 5.56 shows that

$$G_\infty = \lim_{T \to \infty} \bigcup_{t \geq T} G(A(t)) \tag{5.58}$$

is a decreasing limit of a sequence of completely reducible digraphs. Because all digraphs are finite, this sequence is eventually constant. Hence its limit G_∞ is equal to one of the sequence's elements and hence completely reducible.

The next lemma captures the essence of the complete reducibility assumption: If $S_j(t)$ does not change, then $\mu_j(t)$ does not decrease. Together with the weak monotonicity of Lemma 5.49 and eventual connectivity, we are able to show the theorem.

Lemma 5.57. *Under the hypotheses of Theorem 5.54, if j is a leader of $\mathcal{C}$ and $S_j(t) = S_j(t+1)$, then $\mu_j(t+1) \geq \mu_j(t)$.*

Proof. Let $P_{i,j}(t+1)$ be positive, i.e., $i \in S_j(t+1) = S_j(t)$. By definition of $S_j(t)$, we have

$$P_{i,j}(t+1) = \sum_{k \in S_j(t)} A_{i,k}(t+1) \cdot P_{k,j}(t) \ . \tag{5.59}$$

Because $S_j(t) = S_j(t+1)$, we derive that $A_{i,k}(t+1)$ is zero whenever $i \notin S_j(t)$ and $k \in S_j(t)$. Because every node of a cluster is leader-coordinated, every the nodes of a cluster are either all in $S_j(t)$ or all outside of $S_j(t)$. Hence, because the cluster digraph $A(t+1)$ is completely irreducible, we also have that $A_{i,k}(t+1)$ is zero whenever $i \in S_j(t)$ and $k \notin S_j(t)$.

By assumption, we have $i \in S_j(t)$, and hence by the above and by stochasticity of $A(t+1)$:

$$1 = \sum_k A_{i,k}(t+1) = \sum_{k \in S_j(t)} A_{i,k}(t+1) \tag{5.60}$$

Because $P_{k,j}(t) \geq \mu_j(t)$ for all $k \in S_j(t)$, combination of Equations (5.59) and (5.60) yields $P_{i,j}(t+1) \geq \mu_j(t)$. □

Choose any leader j_0 of $\mathcal{C}$. For every $i \in [n]$, let t_i be the least nonnegative integer such that $C_i \subseteq S_{j_0}(t_i)$. All t_i are well-defined as G_∞ is strongly connected. By permuting indices, we can assume without loss of generality that $t_1 \leq t_2 \leq \cdots \leq t_n$. Because $P(0)$ is the identity matrix, we have $S_{j_0}(0) = \{j_0\}$ and hence $t_1 = 0$.

We inductively show

$$\mu_{j_0}(t_m) \geq \alpha^{(m-1)(\mathrm{ind}(H)+1)} \tag{5.61}$$

for all $1 \leq m \leq n$. This is true for $m = 1$. To prove the inductive step, we distinguish two cases: (A) $t_m - t_{m-1} < \mathrm{ind}(H)$ and (B) $t_m - t_{m-1} \geq \mathrm{ind}(H)$.

In case (A), we have

$$\mu_{j_0}(t_m) \geq \alpha^{t_m - t_{m-1}} \cdot \mu_{j_0}(t_{m-1}) \geq \alpha^{(m-1)(\mathrm{ind}(H)+1)} \tag{5.62}$$

by the induction hypothesis.

In case (B), we have $S_{j_0}(t) = S_{j_0}(t_{m-1})$ for all t with $t_{m-1} + \mathrm{ind}(H) \leq t \leq t_m - 1$ by Lemma 5.49 and the definition of t_m. Repeated application of Lemma 5.57 hence yields $\mu_{j_0}(t_m - 1) \geq \mu_{j_0}\big(t_{m-1} + \mathrm{ind}(H)\big)$. We thus have

$$\begin{aligned} \mu_{j_0}(t_m) &\geq \alpha \cdot \mu_{j_0}(t_m - 1) \geq \alpha \cdot \mu_{j_0}(t_{m-1} + \mathrm{ind}(H)) \\ &\geq \alpha^{\mathrm{ind}(H)+1} \cdot \mu_{j_0}(t_{m-1}) \geq \alpha^{(m-1)(\mathrm{ind}(H)+1)} \end{aligned} \tag{5.63}$$

by the induction hypothesis.

In particular, we have shown Equation (5.61) for $m = n$. Now set $T = t_n + \mathrm{ind}(H)$. By Lemmas 5.49 and 5.51, $S_{j_0}(T) = [N]$ for all and $\mu_{j_0}(T) \geq \alpha^{n(\mathrm{ind}(H)+1)}$. This concludes the proof of the theorem.

Part II

Glitch Propagation

Chapter 6

Glitch Propagation in Digital Circuits

6.1 Introduction

Binary value-domain models that allow to model glitch propagation have always been of interest, especially in asynchronous design [89]: Pure delay channels and inertial delay channels, which propagate input pulses with some constant delay only when they exceed some minimal duration, are still the basis of most digital timing analysis approaches and tools. The tremendous advances in digital circuit technology, in particular increased speeds and reduced voltage swings, raised concerns about the accuracy of these models [11]. For example, neither pure nor inertial delay models can express the well-known phenomenon of propagating glitches that decay from stage to stage, which is particularly important for analyzing high-frequency pulse trains or oscillatory metastability [67].

At the same time, the steadily increasing complexity of contemporary digital circuits fuels the need for fast digital timing analysis techniques: Although accurate Spice models, which facilitate very precise analog-level simulations, are usually available for those circuits, the achievable simulation times are prohibitive. Refined digital timing analysis models like the PID model proposed by Bellido-Díaz et al. [11], which is both fast and more accurate, are hence very important from a practical perspective [12].

The interest in binary models that faithfully model glitch propagation and even metastability has also been stimulated recently by the increasing importance of incorporating fault-tolerance in circuit design [33]: Reduced voltage swings and smaller critical charges make circuits more susceptible to particle hits, crosstalk, and electromagnetic interference [46, 68]. Since single-event transients, caused by an ionized particle hitting a reverse-biased transistor, just manifest themselves as short glitches, accurate propagation models are important for assessing soft error rates, in particular, for asynchronous circuits. After all, if system-level fault-tolerance techniques like triple modular redundancy are used for transparently masking value failures, the only remaining issue is timing failures, among which glitches are the most problematic ones.

For example, the DARTS Byzantine fault-tolerant distributed clock generation [45] employs standard asynchronous circuit components, like micropipelines [85], which store clock ticks received from other nodes; a new clock tick is generated when sufficiently many micropipelines are non-empty. Clearly, since any "wait-for-all" mechanism may deadlock in the presence of faulty components, handshaking was replaced by threshold logic in conjunction with some bounded delay assumptions. This way, DARTS can tolerate arbitrary behavior

of Byzantine faulty nodes, except for the generation of pulses with a duration that drive the Muller C-elements of a pipeline into metastability. Analyzing the propagation of such pulses along a pipeline is thus important in order to assess the achievable resilience against such threats [44]. The situation is even worse in case of self-stabilizing algorithms [37], which must be able to recover from an arbitrary initial/error state: Neither handshaking nor any bounded delay condition can be resorted to during stabilization in an algorithm like the one presented by Dolev et al. [36]. Consequently, glitches and the possibility of metastability cannot be avoided.

As a consequence, discrete-value circuit models, analysis techniques and supporting tools for a fast but nevertheless accurate glitch and metastability propagation analysis will be a key issue in the design of future VLSI circuits. We will rigorously prove that none of the existing binary-value candidate models proposed in the past captures glitch propagation adequately. We also propose a new model that does not suffer from the properties we use to show the inadequacy of the existing models.

6.2 State of the Art

Unger [89] proposed a general technique for deriving asynchronous sequential switching circuits that can cope with unrelated input signals. It assumes signals to be binary valued, and requires the availability of combinational circuit elements, as well as pure and inertial delay channels.

Bellido-Díaz et al. [11] proposed the PID model and justified its appropriateness both analytically and by comparing the model predictions against Spice simulation results. The results confirm very good accuracy even for such challenging scenarios as long chains of gates and ring oscillators.

Marino [66] showed that the problem of building a synchronizer can be reduced to the problem of building an inertial delay channel. The reduction circuit only makes use of combinational gates and pure delay channels in addition to inertial delay channels. Marino further shows, in a continuous value signal model, that for a set of standard designs of inertial delay channels, input pulses exist that produce outputs violating the requirements of inertial delay channels. Barros and Johnson [10] extended this work, by showing the equivalence of arbiter, synchronizer, latch, and inertial delay channels.

Marino [67] developed a general theory of metastable operation, and provided impossibility proofs for metastability-free synchronizers and arbiter circuits for several continuous-value circuit models. Branicky [16] proved the impossibility of time-unbounded deterministic and time-invariant arbiters modeled as ordinary differential equations. Mendler and Stroup [69] considered the same problem in the context of continuous automata.

Brzozowski and Ebergen [19] formally proved that, in a model that uses only binary values, it is impossible to implement Muller C-Elements (among other basic state-holding components used in (quasi) delay-insensitive designs) using only zero-time logical gates interconnected by wires without timing restrictions.

6.3 Short Pulse Filtration

There exist a number of devices, theoretical or physical, that seek to suppress glitches in circuits. These include latches, synchronizers, and inertial delay (ID) devices. For all of these,

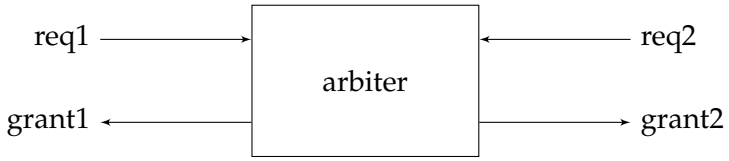


Figure 6.1: An arbiter

there exist various formal definitions. Barros and Johnson [10] proved equivalence of these devices for certain definitions. We choose to focus on a one-shot version of ID devices to study adequacy of binary circuit models. An ID device has one input and one output. It is required to propagate pulses whose length is above some threshold, and to suppress them if their length is below some other, not necessarily equal, threshold. The reason for its definition is that short pulses, which we dub as *glitches*, can cause subsequent devices to enter a metastable state.

Another well studied object in the domain of digital circuits is the *arbiter*. Its job is to arbitrate access to a shared resource. In its simplest version, it has two inputs and two outputs. The two inputs are called *request lines*, denoted req1 and req2. The two outputs are called the *grant lines*, denoted grant1 and grant2. Figure 6.1 schematically depicts an arbiter. The basic safety requirement is that the grant lines should never both be 1 at the same time, which guarantees mutual exclusion. With regard to the liveness requirements, there are a number of different possible definitions. They differ mainly in the delay from the time of a request until the arbiter reaches a decision and grants access. This time may be required to be bounded or not. It turns out that the reaction time of physical arbiter circuits seems to be larger when the time distance between the requests on the requests is smaller. Definitions moreover differ in whether a decision has to be reached if both request lines become activated at exactly the same time. For a definition with bounded reaction time, it was shown by Barros and Johnson [10] that an arbiter is equivalent to an ID device.

Although the experimental validation of the PID model [11] showed good accuracy for the evaluated examples, the question of the general ability of such a model to actually capture the behavior of real physical circuits remained open. And indeed, any *bounded single-history channel* fails to do so in case of the simple Short Pulse Filtration (SPF) problem. The SPF problem requires a circuit to capture a single input pulse if its duration is long enough, and suppress it otherwise, without generating any output glitches. An SPF is hence a one-shot ID device.

A bounded single-history channel is characterized by a bounded channel delay function $\delta(T)$ that may depend on the input-to-previous-output transition time T, i.e., may also take into account the previous output transition time. Pure delay, inertial delay and PID channels all belong to this class of models.

Binary circuit models based on channels with constant $\delta(T)$, i.e., pure delays, do not allow to solve SPF in unbounded time, although there is a simple physical circuit that achieves this. Using channels with non-constant $\delta(T)$, including inertial delays and PID channels, on the other hand, allows to design circuits that solve SPF in bounded time, which contradicts the impossibility of building such circuits physically [67]. Therefore, none of the existing binary circuit models can faithfully capture glitch propagation in real circuits.

We propose a class of channel models that does not suffer from this deficiency: Like bounded single history channels, our *involution channels* involve channel delay functions $\delta(T)$

that may depend on the input-to-previous-output time T. However, unlike bounded single-history channels, we do not assume $\delta(T)$ to be bounded from below but only from above: To support different input thresholds, different instances $\delta_{\downarrow}(T)$ resp. $\delta_{\uparrow}(T)$ can be employed for falling resp. rising transitions, leading to $-\delta_{\downarrow}(-\delta_{\uparrow}(T)) = T$. The case of ordinary involutions corresponds to a 50% threshold for both rising and falling transitions., i.e., functions that form their own inverse.

In the rest of this section, we give an overview of the results we prove on SPF. In Section 6.4, we define the Short Pulse Filtration (SPF) problem in the physical circuit model of Marino and recall the behavior of physical circuits with respect to SPF. That is, we show that unbounded SPF is solvable with physical circuits while bounded SPF is not.

In Chapter 7, we present a generic binary value-domain model for digital clocked and clockless circuits, and introduce the SPF problem. Our generic model comprises zero-time logical gates interconnected by channels that encapsulate model-specific propagation delays and related decay effects. Non-zero time logical gates can be expressed by appending channels with delay at the gate's inputs and outputs. The simplest channel is a pure delay channel, which propagates its input signal with a fixed delay and without any decay, i.e., a pulse has the same duration at the channel's input and output. We then turn our attention to a generalization of constant delay channels, termed *bounded single-history channels*, which are FIFO channels with a generalized delay function that also takes into consideration the last output transition. We distinguish between *forgetful* and *non-forgetful* single-history channels, depending on their behavior when a pulse disappears at the output due to decay effects. All existing binary models we are aware of can be expressed as single-history channels with specific delay functions: A pure delay channel (P) as either a forgetful or non-forgetful single-history channel, a classical inertial delay channel (I) as a forgetful single-history channel, and the channel model proposed by Bellido-Díaz et al. [11] (PID), which additionally has a decay component, as a non-forgetful single-history channel. We also define the *involution channels* and use a simple analog channel model to motivate why involutions are good candidates for suitable delay functions. It also reveals that the standard first-order model used, e.g., in [77] actually gives an instance of general involution channels.

In Chapter 8, we prove that even unbounded SPF is unsolvable when only pure, i.e., constant delay channels are available. This is in contrast with the solvability result with physical circuits of Section 6.4. We also provide a simple circuit that solves unbounded SPF with involution channels, along with a correctness proof. Finally, we prove that weakening SPF to eventual SPF fails to witness the above modeling mismatch: Eventual SPF can by solved both with constant delay and physical channels.

In Chapter 9, we show that bounded SPF is impossible to solve with involution channels. In a nutshell, our proof inductively constructs an execution that can determine the final output only after some unbounded time. We then prove that bounded SPF is solvable if just a single forgetful or non-forgetful bounded single-history channel with non-constant delay is available. However, this is again in contradiction with the result of Section 6.4 showing impossibility of bounded SPF with physical circuits.

Figure 6.2 summarizes our (un)solvability results. Our results reveal that involution channels indeed allow to solve SPF precisely when this is possible in physical circuits, while all other existing binary models do not. We see this as a hint that involution channels are better adapted for fast simulations of digital circuits when wanting to take glitch phenomena into account.

	constant	forgetful	non-forgetful	involution	physical
bounded SPF	X	✓	✓	X	X
SPF	X	✓	✓	✓	✓
eventual SPF	✓	✓	✓	✓	✓

Figure 6.2: Solvability (✓) and unsolvability (X) results for the various channels

6.4 Short Pulse Filtration in Physical Systems

In this section, we will introduce the SPF problem in the model of Marino [67] and use the classic results obtained for bistable elements to determine the solvability/unsolvability border of the SPF problem for real (physical) circuits.

The model of Marino considers circuits which process signals with both continuous value domain and continuous time domain. Accordingly, we assume (normalized) signal voltages to be within $[0, 1]$, and denote by $L_0 = [0, l_0]$ resp. $L_1 = [l_1, 1]$, with $0 < l_0 < l_1 < 1$, the signal ranges that are interpreted as logical 0 resp. logical 1 by a circuit.

A physical circuit with a single input and a single output *solves Short Pulse Filtration (SPF)* if it fulfills the following requirements:

(i) If the input signal is constantly logical 0, then so is the output signal.

(ii) There exists an input signal such that the output signal attains logical 1 at some point in time.

(iii) There exists some fixed $\varepsilon > 0$ such that, if the output signal is not interpreted as logical 1 at two points in time t and t' with $t' - t < \varepsilon$, then it is not logical 1 at any time in between t and t'. Informally, this condition prohibits output signals that may be interpreted as pulses (see Section 6.3) with a duration less than ε.

A physical circuit *solves bounded SPF* if additionally:

(iv) There exists a time T such that if the input signal switches to logical 1 by time t, then the output signal value is either logical 0 or logical 1 at time $t + T$ and remains logical 0 respectively logical 1 thereafter.

We will next argue why there is no physical circuit that solves bounded SPF, but that there are physical circuits solving unbounded SPF.

6.4.1 Unsolvability of Bounded Short Pulse Filtration

The proof is by reduction to the non-existence of a physical bistable storage element that stabilizes within bounded time in the model of Marino. A *single-input bistable element* is a physical circuit with a single input and a single output that fulfills properties (i) and (ii) of SPF as well as:

(iii') If the output is logical 1 at some time t, it also remains logical 1 at all times larger than t.

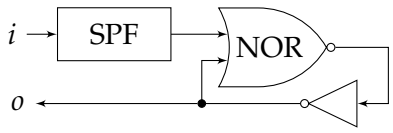


Figure 6.3: Building a bistable storage element from a circuit solving SPF

For a *single-input bistable element stabilizing within bounded time*, additionally (iv) has to hold.

The following Corollary 6.1, which proves the non-existence of a single-input bistable element that stabilizes within bounded time, follows immediately from Theorem 3 in [67].

Corollary 6.1. *There is no single-input bistable element stabilizing within bounded time.*

Now assume, for the sake of a contradiction, that there existed a physical circuit solving bounded SPF and consider the circuit shown in Figure 6.3, with the NOR's initial output equal to 1 and the inverter's initial output equal to 0 at time $t = 0$.

It is not difficult to prove that this circuit implements a single-input bistable element stabilizing within bounded time: In case the input signal i is always logical 0, the SPF's output signal will always be logical 0 due to property (i) of the SPF. Thus the circuit shown in Figure 6.3 will always drive a logical 0 at its output, which confirms property (i) for the bistable element.

Now let u be an input pulse that makes the SPF circuit produce a logical 1 at its output. Letting t' be the first time the SPF circuit drives a logical 1 at its output, its output must remain logical 1 within $[t', t' + \varepsilon]$ for some $\varepsilon > 0$ due to property (iii) of the SPF. Assuming that the signal propagation delay of the NOR gate and the inverter is short enough for the inverter's output to reach a logical 1 before time $t' + \varepsilon$, the NOR gate will subsequently drive a logical 0 on its output forever, irrespective of the output of the SPF circuit. The circuit's output signal o will hence continuously remain logical 1 once it switched to logical 1, which also confirms properties (ii) and (iii') of the bistable element.

Due to the use of a circuit solving bounded SPF in the compound circuit, we further obtain that there exists some $T > 0$ such that, for any input pulse u' that switches to logical 1 by time t, the circuit shown in Figure 6.3 produces a logical 1 by time $t + T$, a contradiction to the non-existence of a single-input bistable element stabilizing in bounded time. We hence obtain:

Theorem 6.2. *No physical circuit solves bounded SPF.*

6.4.2 Solvability of Unbounded Short Pulse Filtration

To show the existence of a circuit solving unbounded SPF, we make use of a circuit known as a metastability filter (see, e.g., [60, p. 40]). According to Marino [67], pulses of arbitrary length may drive the internal state of every storage loop (including the one shown in Figure 6.3) into a metastable region for an unbounded time. A circuit may hence produce an output signal within some region of metastable output values $[v_M^-, v_M^+] \subset [0, 1]$ during an unbounded time, where the values v_M^-, and v_M^+ depend on technology parameters. However, since it is possible to compute safe bounds V_M^-, and V_M^+ such that $[v_M^-, v_M^+] \subset [V_M^-, V_M^+] \subset [0, 1]$, a continuously valid output signal can be produced by means of a subsequent high-threshold buffer: By

connecting the output o of Figure 6.3, ignoring the SPF block, to the input of a (high-threshold) buffer, which maps input signal values within $[0, B_M^-]$ to output signal values that are logical 0, and input values within $[B_M^+, 1]$ to output values that are logical 1, where $V_M^+ < B_M^-$, we obtain a physical circuit that solves (unbounded) SPF. Hence:

Theorem 6.3. *There is a physical circuit that solves unbounded SPF.*

Chapter 7

Binary Circuit Model

7.1 Basics: Signals, Circuits, Executions, Problem Definition

Since the purpose of our work is to replace analog models by a purely digital model, we will now formally define the binary-value continuous-time circuit model we use. It unifies all models of this sort that we are aware of.

7.1.1 Signals

A *falling transition* at time t is the pair $(t, 0)$, a *rising transition* at time t is the pair $(t, 1)$. A *signal* is a (finite or infinite) list of alternating transitions such that

S1) the initial transition is at time $-\infty$; all other transitions are at times $t \geq 0$.

S2) the transition times are strictly increasing.

S3) if there are infinitely many transitions in the list, then the set of transition times is unbounded.

To every signal s corresponds a function $\mathbb{R}_+ \to \{0, 1\}$ whose value at time t is that of the most recent transition. We follow the convention that the function already has the new value at the time of a transition, i.e., the function is constant in the half-open interval $[t_n, t_{n+1})$ if t_n and t_{n+1} are the times of two consecutive transitions. A signal is uniquely determined by such a function and its value at $-\infty$.

7.1.2 Circuits

Circuits are obtained by interconnecting a set of input ports and a set of output ports, forming the external interface of a circuit, and a set of combinational gates via channels. We constrain the way components are interconnected in a natural way, by requiring that any gate input, channel input and output port is attached to only one input port, gate output or channel output. Moreover, gates and channels must alternate on every path in the circuit.

Formally, a *circuit* is described by a directed graph with the following properties:

C1) The vertices are partitioned into *input ports*, *output ports*, *channels*, and *gates*.

C2) Input ports have no incoming edges and at least one outgoing edge.

C3) Output ports have exactly one incoming edge from a gate and no outgoing edges.

C4) Channels are nodes that have exactly one incoming and exactly one outgoing edge. Every channel is assigned a channel function, which maps the input to the output.

C5) Every gate is assigned a Boolean function $\{0,1\}^d \to \{0,1\}$, where d is the number of incoming edges.

C6) There is a fixed order on the incoming edges of every gate.

C7) Gates and channels alternate on every path in a circuit.

7.1.3 Executions

An execution of circuit C is an assignment of signals to vertices that respects the channel functions and Boolean gate functions.

Formally, an *execution* of circuit C is a collection of signals s_v for all vertices v of C such that the following properties holds:

E1) If i is an input port, then there are no restrictions on s_i.

E2) If o is an output port, then $s_o = s_v$, where v is the unique gate v associated with o.

E3) If c is a channel, then $s_c = f_c(s_v)$, where v is the unique incoming neighbor of c and f_c the channel function.

E4) If b is a gate with d incoming neighbors $v_1, \ldots, v_d$, ordered according to the fixed order of condition (C6) and gate function f_b, then for all times t,

$$s_b(t) = f_b\big(s_{v_1}(t), s_{v_2}(t), \ldots, s_{v_d}(t)\big) \ . \tag{7.1}$$

7.1.4 Short Pulse Filtration.

A *pulse* of length Δ at time T has initial value 0, one rising transition at time T, and one falling transition at time $T + \Delta$.

A signal *contains a pulse* of length Δ at time T if it contains a rising transition at time T, a falling transition at time $T + \Delta$ and no transition in between.

A circuit *solves Short Pulse Filtration (SPF)* if it fulfills the following conditions:

F1) It has exactly one input port and exactly one output port. *(Well-formedness)*

F2) If the input signal is zero, then so is the output signal. *(No generation)*

F3) There exist an input pulse such that the output signal is not the zero signal. *(Nontriviality)*

F4) There exists an $\varepsilon > 0$ such that the output signal never contains a pulse of length less than ε. *(No short pulses)*

A circuit *solves bounded SPF* if additionally the following condition holds:

F5) There exists a $K > 0$ such that the last output transition is before time $T + K$ if T is the time of the last input transition. *(Bounded stabilization time)*

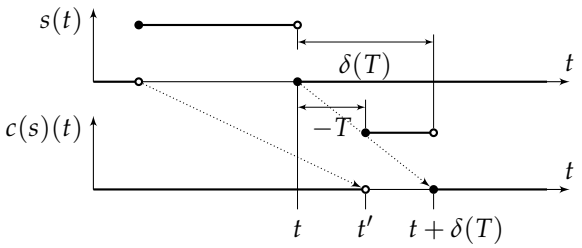


Figure 7.1: Input/output signal of a single-history channel, involving the input-to-previous-output delay T and the resulting output-to-input delay $\delta(T)$

7.2 Bounded Single-History Channels

This section formally introduces the notion of bounded single-history channels in the binary circuit model. They are a generalization of constant delay channels that covers all channel models for binary circuit models in the existing literature that we are aware of.

Intuitively, a bounded single-history channel propagates each event, occurring at time t, of the input signal to an event at the output happening after some bounded *output-to-input* delay $\delta(T)$, which depends on the *input-to-previous-output* delay $T = t - t'$. Note that T is positive if the channel delay is short compared to the input signal transition times, and negative otherwise. Figure 7.1 illustrates this relation and the involved delays. In case FIFO order would be invalidated, i.e., $t + \delta(T) \leq t'$, such that the next output event would not occur after the previous one, both events cancel.

There exist two variants of bounded single-history channels in the literature, depending on whether the time of an canceled event is remembered or not. We dub these two variants *forgetful* and *non-forgetful* bounded single-history channels, which we both formally define below. At the end of this section, we give a list of channel models that are special cases of our definition of bounded single-history channels.

Formally, a *bounded single-history channel* c is characterized by an *initial value* $I \in \{0,1\}$, a nondecreasing *delay function* $\delta : \mathbb{R} \to \mathbb{R}$ such that $\delta(\infty) = \delta_\infty = \lim_{T\to\infty} \delta(T)$ is finite and positive, and the fact whether it is forgetful or not. We detail the channel behavior in the next two subsections.

7.2.1 Forgetful Single-History Channels

This class of channels includes the classical inertial delay channels as used, for example, in VHDL simulators [5].

Their behavior is defined by the following algorithm: Let s be a signal. In case the channel's initial value I is equal to the initial value of s, or there is an event at time 0 in the event list of s, let the channel's *input list* $\big((t_n, x_n)\big)_n$ be the event list of s. Otherwise, let the channel's input list be the event list of s with an additional event at time 0 and value equal to the initial value of s. The algorithm iterates the input list and updates the *output list*, which will define the channel's output signal $c(s)$.

Initially, let $(-\infty, I)$ be the sole element of the output list. In its n^{th} iteration the algorithm considers input event (t_n, x_n) and modifies the output list accordingly:

1. Denote by (t'_n, x'_n) the last event in the output list. If $x_n = x'_n$, then input event (t_n, x_n) has no effect: Proceed to the $(n+1)^{\text{th}}$ iteration.

2. Otherwise, let $T_n = t_n - t'_n$ be the difference of input and previous-output event times. Note that $T_n = \infty$ is possible. In this case $\delta(T_n) = \delta(\infty) = \lim_{T\to\infty} \delta(T)$, which is finite by assumption.

   If $t_n + \delta(T_n) > t'_n$, then add the event $\big(t_n + \delta(T_n), x_n\big)$ to the output list.

   If $t_n + \delta(T_n) \leq t'_n$, then delete the event (t'_n, x'_n) from the output list.

Note that the output sequence's first event is always $(-\infty, I)$, all other events have positive times (since $\delta(\infty) > 0$), its sequence of event times is strictly increasing, and its sequence of values is alternating.

If the input list is finite, the algorithm halts. If not, the output sequence nonetheless stabilizes in the sense that, for every time t, there exists some N such that all iterations with $n \geq N$ make no changes to the output sequence at times $\leq t$. The next lemma (Lemma 7.2) proves this property and makes the limit output list as n tends to infinity well-defined. So, even if the input list is infinite, there exists a well-defined (infinite) output list S that is the result of the described algorithm. The channel's output signal $c(s)$ is then defined by event list S:

Definition 7.1. For input signal s, the output signal $c(s)$ of the forgetful bounded single-history channel c is the signal whose event list is the list S as defined by the above algorithm.

Lemma 7.2. *Denote by S_n the output list after the n^{th} iteration of the forgetful channel algorithm, and by $S_n|t$ its restriction to the events at times at most t. For all t there exists an N such that $S_n|t$ is constant for all $n \geq N$.*

Proof. The lemma is trivial if the input list is finite, so we assume it to be infinite.

Because the sequence of input event times (t_n) tends to infinity, there exists an N such that

$$t_N \geq \max\big(t\,,\, t - \delta(-\delta(\infty))\big)\ . \tag{7.2}$$

We show by induction that $S_n|t = S_N|t$ for all $n \geq N$. This is trivial for $n = N$, so let $n > N$. Then $t_n > t_N$.

Let (t'_n, x'_n) be the last element in S_{n-1}, and $T_n = t_n - t'_n$. The case $x_n = x'_n$ is trivial, so let $x_n \neq x'_n$. We distinguish two cases, depending on whether $\delta(T_n) > -T_n$ or not:

Case 1: $\delta(T_n) > -T_n$. Because δ is nondecreasing, $\delta(T_n) \leq \delta(\infty)$, and hence $T_n > -\delta(\infty)$ and also $\delta(T_n) \geq \delta(-\delta(\infty))$. This implies $t_n + \delta(T_n) > t_N + \delta(-\delta(\infty)) \geq t$ by using (7.2). Hence $S_n|t = S_{n-1}|t = S_N|t$ by the induction hypothesis.

Case 2: $\delta(T_n) \leq -T_n$. We show that $t'_n > t$ by contradiction: Let $t'_n \leq t$. Then $T_n = t_n - t'_n > t_N - t \geq 0$, by using (7.2). From $\delta(\infty) > 0$, we thus obtain $T_n > -\delta(\infty)$. Hence $\delta(T_n) \geq \delta(-\delta(\infty))$ by monotonicity of δ. By assumption, $\delta(-\delta(\infty)) \leq \delta(T_n) \leq -T_n = t'_n - t_n$, which implies $t_n \leq t'_n - \delta(-\delta(\infty))$, i.e., $t_N < t - \delta(-\delta(\infty))$. This is a contradiction to (7.2), which shows that $t'_n > t$. Hence $S_n|t = S_{n-1}|t = S_N|t$ by the induction hypothesis. □

7.2.2 Non-Forgetful Single-History Channels

The PID channel introduced by Bellido-Díaz et al. [11] is not covered by the above forgetful bounded single-history channels, since it has been designed to reasonably match analog RC

waveforms: Analog signals like exponential functions do not "forget" sub-threshold pulses. Hence, they cannot be modeled via delay functions $\delta(T)$ that depend on the input-to-previous output delay T. To also cover the PID model, we hence introduce non-forgetful bounded single-history channels, the delay function of which may also depend on the last canceled event.

The output-eventlist generation algorithm for non-forgetful channels thus maintains an additional variable r, which, in each iteration, contains the time of the *potential output event* considered in the last iteration. Note that this approach was already used in the PID-channel-model by Bellido-Díaz et al. [11, Fig. 13]. Similar to the forgetful case, it determines the output signal $c(s)$ of a non-forgetful bounded single-history channel c, given input signal s with input event list $\big((t_n, x_n)\big)_n$ as follows:

Initially, the output list contains the sole element $(-\infty, I)$ and $r = r_{-1} = -\infty$. In its n^{th} iteration, the algorithm considers input event (t_n, x_n) and modifies the output list accordingly:

1. Denote by (t'_n, x'_n) the last event in the output list. If $x_n = x'_n$, then input event (t_n, x_n) has no effect: Proceed to the $(n+1)^{\text{th}}$ iteration.

2. Otherwise, let $T_n = t_n - r_{n-1}$ be the difference of input and most recent potential output event times and set $r_n = t_n + \delta(T_n)$.

   If $t_n + \delta(T_n) > r_{n-1}$, then add the event $\big(t_n + \delta(T_n), x_n\big)$ to the output list.

   If $t_n + \delta(T_n) \leq r_{n-1}$, then delete the event (t'_n, x'_n) from the output list.

We first show that if event (t'_n, x'_n) is deleted in the n^{th} iteration, then $r_{n-1} = t'_n$: Assume by contradiction that this is not the case, and let n be the first iteration where the statement is violated. Then it must hold that $n \geq 2$, as in iteration $n-2$ some event (τ, x_{n-2}) must have been added to the output list that was deleted in iteration $n-1$, due to $\tau' = t_{n-1} + \delta(T_{n-1}) \leq r_{n-2} = \tau$. Furthermore, in iteration n, our assumption of deleting some event with a time different from $r_{n-1} = \tau'$ implies $\tau'' = t_n + \delta(T_n) \leq \tau'$. However, from $t_{n-1} < t_n$, $\tau \geq \tau'$ and monotonicity of δ, $t_{n-1} + \delta(t_{n-1} - \tau) < t + \delta(t - \tau')$, i.e., $\tau' < \tau''$, which provides the required contradiction.

Thus, an event is either deleted in the next iteration, or never deleted. The output sequence's first event $(-\infty, I)$ is obviously never deleted.

By analogous arguments, one can show that the sequence of event times is strictly increasing, with an alternating sequence of values. Unlike in the case of forgetful channels, however, the event list generation algorithm may produce events with *finite* negative times that will be removed from the final output. In case the input list is finite, the algorithm clearly halts. If not, we again have the same stabilization property as for forgetful bounded single-history channels, which we will provide in Lemma 7.4 below. Thus the algorithm's final output list S is again well-defined and we can define:

Definition 7.3. For input signal s, the output signal $c(s)$ of the forgetful bounded single-history channel c is the signal whose event list is the list S as defined by the above algorithm. after deleting all events with finite negative times and the first non-negative time event if its value is equal to the channel's initial value I.

Lemma 7.4. *Denote by S_n the output list after the n-th iteration of the forgetful channel algorithm, and by $S_n|t$ its restriction to the events at times at most t. For all t, there exists an N such that $S_n|t$ is constant for all $n \geq N$.*

Proof. The lemma follows from the fact that an event can only be deleted one iteration after it was added to the output list, and the fact that in each iteration n, $T_n > -\delta(\infty)$ and thus $t_n + \delta(T_n)$ is lower bounded by $t_n + \lim_{t\to 0^+} \delta(-\delta(\infty) + t)$. □

7.2.3 Examples of Single-History Channels

Below, we summarize how the existing binary-value models are mapped to bounded single-history channels:

1) A classic *pure-delay channel* is a bounded single-history channel whose delay function δ is constant and positive. The behavior of a pure-delay channel does not depend on whether it is forgetful or not.

2) An inertial channel is a forgetful bounded single-history channel whose delay function δ is of the form
$$\delta(T) = \begin{cases} \delta_0 & \text{if } T > T_0 \\ -T_0 & \text{if } T \leq T_0 \end{cases} \tag{7.3}$$
for parameters $\delta_0 > 0$ and $T_0 > -\delta_0$. An inertial channel filters an incoming pulse if and only if its pulse length is less or equal to $T_0 + \delta_0$; otherwise, it is forwarded with delay δ_0.

3) The PID-channels of Bellido-Díaz et al. [11] are non-forgetful with delay function
$$\delta(T) = t_{p0} \cdot \left(1 - e^{-(T-T_0)/\tau}\right) \tag{7.4}$$
for (measured) positive parameters t_{p0}, τ, and T_0. Note that $\delta(T_0) = 0$, $\lim_{t\to\infty} \delta(T) = t_{p0}$, and $\frac{d\delta(T)}{dT}|_{T=0} = t_{p0}/\tau$ here.

7.3 Involution Channels

Intuitively, a channel propagates each transition at time t of the input signal to a transition at the output happening after some *output-to-input* delay $\delta(T)$, which depends on the *input-to-previous-output* delay $T = t - t'$. Note that T is positive if the channel delay is short compared to the input signal transition gaps, and negative otherwise. Figure 7.1 illustrates this relation and the involved delays.

Formally, an *involution channel* is characterized by an *initial value* $I \in \{0,1\}$ and two increasing concave *delay functions* $\delta_\uparrow : (-\delta_\infty^\downarrow, \infty) \to (-\infty, \delta_\infty^\uparrow)$ and $\delta_\downarrow : (-\delta_\infty^\uparrow, \infty) \to (-\infty, \delta_\infty^\downarrow)$ such that both $\delta_\infty^\uparrow = \lim_{T\to\infty} \delta^\uparrow(T)$ and $\delta_\infty^\downarrow = \lim_{T\to\infty} \delta^\downarrow(T)$ are finite and
$$-\delta_\uparrow\big(-\delta_\downarrow(T)\big) = T \text{ and } -\delta_\downarrow\big(-\delta_\uparrow(T)\big) = T \tag{7.5}$$
for all applicable T. All such functions are necessarily continuous and strictly increasing. For simplicity, we will also assume them to be differentiable; δ being concave thus implies that its derivative δ' is decreasing.

The behavior of involution channels is defined as follows:
Initialization: If the channel's initial value I is different from the initial value X of the channel input signal s and s has no transition at time 0, add the transition $(0, X)$ at time 0 to s. If multiple channels share a common input signal, as depicted in Figure 7.2, we require that

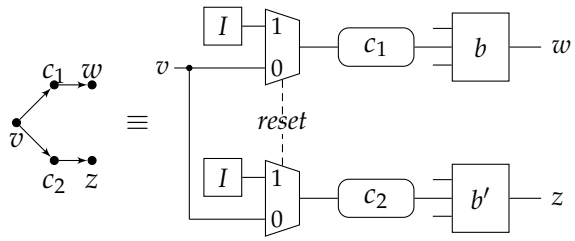


Figure 7.2: A circuit (graph) with vertex v (being an input or a gate), gates w, z, and channels c_1 and c_2 (on the left) and the physical equivalent (on the right). Both channels must have the same initial value I; b and b' are the Boolean functions assigned to gates w and z, respectively

they all have the same initial value I. This is without loss of generality, as one can always replicate the input signal.
Output transition generation algorithm: Let $t_1, t_2, \ldots$ be the times of the transitions of s, and set $t_0 = -\infty$ and $\delta_0 = 0$.

- *Iteration:* Determine the tentative list of *pending* output transitions: Recursively determine the input-to-output delay for the input transition at time t_n by setting $\delta_n = \delta_\uparrow(t_n - t_{n-1} - \delta_{n-1})$ if t_n is a rising transition and $\delta_n = \delta_\downarrow(t_n - t_{n-1} - \delta_{n-1})$ if it is falling. The n^{th} and m^{th} pending output transitions *cancel* if $n < m$ but $t_n + \delta_n \geq t_m + \delta_m$. In this case, we mark both as canceled.

- *Return:* The channel output signal $c(s)$ has initial value I and contains every pending transition at time $t_n + \delta_n$, provided it has not been marked as canceled.

Definition 7.5. An involution channel is *strictly causal* if $\delta_\uparrow(0) > 0$, which is equivalent to the condition $\delta_\downarrow(0) > 0$ due to (7.5).

The next lemma identifies an important parameter $\delta_{\min}$ of a strictly causal involution channel, which gives its minimal pure delay.

Lemma 7.6. *A strictly causal involution channel has a unique $\delta_{\min}$ defined by $\delta_\uparrow(-\delta_{\min}) = \delta_{\min} = \delta_\downarrow(-\delta_{\min})$, which is positive.*
For the derivative, $\delta'_\uparrow(-\delta_\downarrow(T)) = 1/\delta'_\downarrow(T)$ and hence $\delta'_\uparrow(-\delta_{\min}) = 1/\delta'_\downarrow(-\delta_{\min})$.

Proof. Set $f(T) = -T + \delta_\uparrow(-T)$. This function is continuous and strictly decreasing, since $\delta_\uparrow$ is continuous and nondecreasing. Because $f(0) = \delta_\uparrow(0)$ is positive and the limit of $f(T)$ as $T \to \delta_\infty^\uparrow$ is $-\infty$, there exists a unique $\delta_{\min}$ between 0 and $\delta_\infty^\uparrow$ for which $f(\delta_{\min}) = 0$. Hence, $\delta_\uparrow(-\delta_{\min}) = \delta_{\min}$. The second equality follows from $\delta_{\min} = \delta_\downarrow(-\delta_\uparrow(-\delta_{\min})) = \delta_\downarrow(-\delta_{\min})$ according to (7.5).
The second part of the lemma follows by differentiating Equation (7.5). □

We next show that $\delta_{\min}$ indeed deserves its name: A particular consequence of the following lemma is that the channel delay for any non-canceled transition is at least $\delta_{\min}$.

Lemma 7.7. *The n^{th} and $(n+1)^{th}$ pending output transitions cancel if and only if $t_{n+1} \leq t_n + \delta_n - \delta_{\min}$.*

Proof. Let δ be either $\delta_\uparrow$ or $\delta_\downarrow$, depending on whether t_{n+1} is a rising or falling transition. By definition, the two transitions cancel if and only if

$$\delta_{n+1} = \delta(t_{n+1} - t_n - \delta_n) \leq -(t_{n+1} - t_n - \delta_n) \ . \tag{7.6}$$

Set $T = t_{n+1} - t_n - \delta_n$. By Lemma 7.6, equality holds in (7.6) if and only if $T = -\delta_{\min}$. Because the left-hand side of (7.6) is increasing in T and the right-hand side is strictly decreasing in T, (7.6) is equivalent to $T \leq -\delta_{\min}$, which in turn is equivalent to $t_{n+1} \leq t_n + \delta_n - \delta_{\min}$. □

In the remainder, we assume all channels to be strictly causal involution channels.

7.4 Constructing Executions of Circuits with Strictly Causal Involution Channels

The definition of an execution of a general circuit as given in Section 7.1 is "existential", in the sense that it only allows to check for a given collection of signals whether it is an execution or not. And indeed, in general, circuits may have no execution or may have several different executions. By contrast, in case of circuits involving strictly causal involution channels only, executions are unique and can be constructed iteratively: We give a deterministic construction algorithm below.

Given a circuit C with strongly causal involution channels, let $(s_i)_{i\in\mathcal{I}}$ be any collection of signals for all the input ports $\mathcal{I}$; E_i denotes s_i's corresponding transition list. Without loss of generality, we can assume that all output ports are driven by gates and identify the output port with the output of its driving gate. The channel with predecessor x (an input port or a gate output) and successor y (a gate input) is denoted by the tuple (x, y). The algorithm iteratively generates the list of transitions E_σ of (the output of) every vertex σ in the circuit, and hence the corresponding signal $s_\sigma(t)$. In the course of the execution of this algorithm, a subset of the generated transitions will be marked *fixed*: Non-fixed transitions could still be canceled by other transitions later on, fixed transitions will actually occur in the constructed execution.

The detailed algorithm is as follows:
Initialization: For all channels (v, w) in C, $E_{(v,w)} = ((-\infty, I))$ initially, with I being the initial value of channel (v, w). According to the implicit reset of our channels introduced in Section 7.3, the transition $(0, X)$ is also added to $E_{(v,w)}$ if the initial transition $(-\infty, X)$ of E_v satisfies $X \neq I$. Note that this is well-defined also in case of channels (v, w) and (v, w') attached to the same v, as we require $E_{(v,w)} = E_{(v,w')}$ in this case. For a gate v, $E_v = ((-\infty, X))$ initially, where X is the value of the Boolean function corresponding to v applied to the values of the initial transitions in E_σ for all of v's predecessors σ. The zero-input gates 0 and 1 used for generating constant 0 and constant 1 signals have $E_0 = ((-\infty, 0)$ and $E_1 = ((-\infty, 1)$, respectively. Initially, all transitions at $-\infty$ are fixed and all others are not.
Iteration: If there is no non-fixed transition left, terminate with the execution made up by all fixed transitions. Otherwise, let $t \geq 0$ be the smallest time of a non-fixed transition.

(i) Mark all transitions at t fixed.

(ii) For each newly fixed transition from step (i), occurring in E_σ where σ is a predecessor of a gate v: If signal s_v's current value $s_v(t) = X$ differs from the value of v's Boolean

function applied to the values $s_{\sigma'}(t)$ for all of v's predecessors σ' (which also include σ), add the transition $(t, 1 - X)$ to E_v and mark it fixed.

(iii) For each newly fixed transition $(t, x) \in E_v$ from steps (i) or (ii), occurring in E_v of a gate output or an input port: For each successor channel (v, w) of v, apply the iteration step of (v, w)'s transition generation algorithm with input list E_v, output list $E_{(v,w)}$, and current input transition (t, X). If this leads to a cancellation in $E_{(v,w)}$, remove both canceling and canceled transition from the list. Lemma 7.10 will show that no fixed transition will ever be removed this way.

We will now show that this algorithm indeed constructs an execution of C. Let t_ℓ be the smallest finite time of non-fixed transitions at the beginning of iteration $\ell \geq 1$ of the algorithm, and denote by $\delta_{\min}^C > 0$ the minimal $\delta_{\min}$ of all channels in circuit C. By slight abuse of notation, we omit C when it is clear from the context.

Lemma 7.8. *For all iterations $\ell \geq 1$, (a) no transition (t, X) with $t \neq t_\ell$ is newly marked fixed in the iteration, (b) a transition (t, X) added during and not removed by the end of iteration ℓ either has time $t = t_\ell$ or $t > t_\ell + \delta_{\min} > t_\ell$, and (c) every transition at time t_ℓ is fixed at the end of the iteration.*

Proof. Statement (a) is implied by the fact that transitions are only marked fixed in step (i) and (ii), which act on transitions at time t_ℓ only.

For (b), assume by contradiction that a transition (t, X) with $t \leq t_\ell + \delta_{\min}$ but different from t_ℓ was added in iteration ℓ and still exists at the end of iteration ℓ. Such a transition can only be added via step (iii). For the respective channel algorithm with delay function δ, $\delta(t_\ell - t') \leq \delta_{\min}$ must have held, where t' is the time of the channel's last output transition. From Lemma 7.6, we deduce that this implies $t_\ell \leq t' - \delta_{\min}$. By Lemma 7.7, this leads to a cancellation and hence removal of (t, X), which provides the required contradiction.

For (c), assume by contradiction that, at the end of iteration ℓ, there exists a non-fixed transition (t_ℓ, X). Since step (i) marks all transitions at time t_ℓ fixed and (ii) adds only fixed transitions at time t_ℓ, the non fixed transition must have been newly added in step (iii). However, from (b), we know that this requires $t > t_\ell + \delta_{\min} > t_\ell$, a contradiction. □

From an inductive application of Lemma 7.8, we obtain that the sequence of iteration start times $(t_\ell)_{\ell \geq 1}$ is strictly increasing without bound:

Lemma 7.9. *For all iterations $\ell > 1$, $t_\ell - t_{\ell-1} > 0$. If t_ℓ does not involve an input transition, then $t_\ell - t_{\ell-1} > \delta_{\min}$.*

Proof. By Lemma 7.8 (b), $t_{\ell+1}$ is larger than $t_\ell + \delta_{\min}$, provided no input transition occurs earlier. As we do not allow Zeno behavior of input signals, $t_\ell - t_{\ell-1} > 0$ is guaranteed also in the latter case. □

The following lemma proves that the generated event lists are well-defined, in the sense that no later iteration can remove events that may have generated causally dependent other events already.

Lemma 7.10. *No fixed transition is canceled in any iteration.*

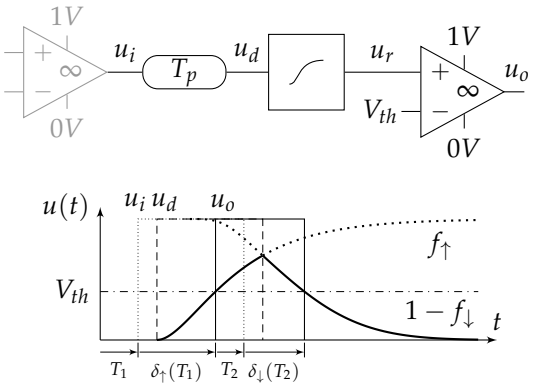


Figure 7.3: Simple analog channel model

Proof. Assume by contradiction that some iteration $\ell \geq 2$ is the first in which a fixed transition is canceled. From Lemma 7.9, it follows that such a transition can only be canceled in step (iii). Thus, there exists a transition at time t_ℓ that generated a new transition at some time t that results in the cancellation of a fixed transition at time t'. Lemma 7.7 implies that $t_\ell - t' \leq -\delta_{\min} < 0$ in this case. By Lemma 7.8, however, $t \leq t' < t_\ell$ and thus $t_\ell - t' > 0$, which provides the required contradiction. □

We are now ready for the main result of this section, which asserts the existence of a unique execution of our circuit C:

Theorem 7.11. *At the end of iteration $\ell \geq 1$, the collection of signals s_v corresponding to E_v, v in G, restricted to time $[-\infty, t_\ell]$ is the unique execution of circuit C restricted to time $[-\infty, t_\ell]$. If the algorithm terminates at the beginning of iteration ℓ, then this collection of signals is the unique execution of circuit C.*

Proof. From Lemma 7.9, we deduce that for all times $t \geq 0$, there is an iteration $\ell \geq 1$ such that $t_\ell > t$ or the algorithm terminates. From Lemma 7.10, we further know that in both cases the algorithm does not add transitions with times less or equal to t. Uniqueness of the execution follows from the fact that the construction algorithm is deterministic. □

7.5 Specific Class of Involution Channels: Exp-Channels

In order to motivate why involutions are promising candidates for suitable δ-functions, consider the simple analog channel model depicted in Figure 7.3. This well-known model, see, e.g., [77] for an instance, consists of a (pure) delay element with delay T_p, a slew rate limiter and a comparator, all of which are idealized. The circuit input u_i, coming from the comparator of the previous stage, hence takes on the value 0 or 1 (Volt) and switches between those two values with infinite slope. The unit (Volt) will be omitted subsequently. Both $u_i(t)$ and the output $u_d(t)$ of the delay element can hence be viewed as binary-valued signals. The slew rate limiter replaces the infinite-slope transitions of u_d with the predefined slew rate functions $f_\uparrow$ for the rising edge and $1 - f_\downarrow$ for the falling edge on its output u_r. These functions (collectively termed "f" below) must have the following properties: $f(0) = 0$, $\lim_{t\to\infty} f(t) = 1$, and

f is strictly increasing and continuous. Finally, the comparator discretizes u_r by comparing its value to a threshold V_{th}, thereby generating output signal u_o.

In order to analyze the behavior of such a channel, we consider its input $u_i(t)$ to be a signal made up of a sequence of alternating transitions $(t_n, x_n)_{n\geq 0}$: t_n is the time of the n-th transition, and $x_n = 0$ resp. $x_n = 1$ identifies it to be falling resp. rising; the initial transition occurs at time $t_0 = -\infty$ for convenience. Let $(\hat{t}_n)_{n\geq 0}$ be the corresponding sequence of switching times of the pure delay output u_d, i.e., $\hat{t}_n = t_n + T_p$ for $n \geq 0$. Note that $\hat{t}_0 = -\infty$, and assume for simplicity that the initial values of u_i and u_d are $x_0 = \hat{x}_0 = 0$. Then, $\forall k \in \mathbb{N}_0, \forall t \in [\hat{t}_{2k}, \hat{t}_{2k+1}) : u_d(t) = 0$ and similarly $u_d(t) = 1$ when $t \in [\hat{t}_{2k+1}, \hat{t}_{2k+2})$. The slew rate limiter replaces the infinite-slope transitions with instances of $f_\uparrow$ and $1 - f_\downarrow$ as follows: $\forall k \in \mathbb{N}_0, \forall t \in [\hat{t}_{2k}, \hat{t}_{2k+1}) : u_r(t) = 1 - f_\downarrow(t - \hat{t}_{2k} + \theta_{2k})$ and similarly $u_r(t) = f_\uparrow(t - \hat{t}_{2k+1} + \theta_{2k+1})$ when $t \in [\hat{t}_{2k+1}, \hat{t}_{2k+2})$. The sequence $(\theta_n)_{n\geq 0}$ is defined implicitly, by requesting continuity of $u_r(t)$ for all t. More explicitly, this requires $\theta_0 = 0$ and, in case of $n = 2k + 1$, i.e., a rising transition at $\hat{t}_n$, $\theta_n = f_\uparrow^{-1}(u_r(\hat{t}_n))$. Substituting $u_r(\hat{t}_n)$, we get $\theta_n = f_\uparrow^{-1}(1 - f_\downarrow(\hat{t}_n - \hat{t}_{n-1} + \theta_{n-1}))$. For falling transitions, the formula is the same with $f_\uparrow$ and $f_\downarrow$ swapped; note that the inverse of $1 - f(x)$ is just $f^{-1}(1 - x)$.

The comparator again produces the "binary-valued" output signal $u_o(t)$, by comparing u_r to a threshold voltage $V_{th} \in (0, 1)$. Knowing that u_r is composed of alternating instances of the bijective functions $f_\uparrow$ and $1 - f_\downarrow$, there exist unique $\Delta_\uparrow = f_\uparrow^{-1}(V_{th})$ and $\Delta_\downarrow = f_\downarrow^{-1}(1 - V_{th})$ such that $f_\uparrow(\Delta_\uparrow) = V_{th}$ and $1 - f_\downarrow(\Delta_\downarrow) = V_{th}$. Therefore, we can derive $u_o(t)$ directly from u_r, by generating rising transitions at time $t_{2k+1} - \theta_{2k+1} + \Delta_\uparrow$ and falling transitions at $t_{2k} - \theta_{2k} + \Delta_\downarrow$. Note carefully that the resulting output transition times need not be strictly increasing any more, which results in cancellation of transitions.

The overall input to output behavior of the channel for any rising input transition $(t_n, 1)$ on u_i can now be stated as follows: $\hat{t}_n = t_n + T_p$ is mapped to the instance $f_\uparrow(t - t_n - T_p + \theta_n)$ in the slew rate limiter, from which the comparator generates the corresponding transition of u_o at time $t'_n = t_n + T_p - \theta_n + f_\uparrow^{-1}(V_{th})$. Substituting for θ_n, we get $t'_n = t_n + T_p - f_\uparrow^{-1}(1 - f_\downarrow(t_n - t_{n-1} + \theta_{n-1})) + f_\uparrow^{-1}(V_{th})$. Utilizing the input-to-previous-output transition time $T = t_n - t_{n-1} - T_p + \theta_{n-1} - f_\downarrow^{-1}(1 - V_{th})$, we obtain $t'_n = t_n + T_p - f_\uparrow^{-1}(1 - f_\downarrow(T + T_p + f_\downarrow^{-1}(1 - V_{th}))) + f_\uparrow^{-1}(V_{th})$. The output-to-input delay $\delta_\uparrow(T) = t'_n - t_n$ (and $\delta_\downarrow(T)$, which is obtained analogously) is thus:

$$\begin{aligned} \delta_\uparrow(T) &= T_p - f_\uparrow^{-1}(1 - f_\downarrow(T + T_p + f_\downarrow^{-1}(1 - V_{th}))) + f_\uparrow^{-1}(V_{th}) \\ \delta_\downarrow(T) &= T_p - f_\downarrow^{-1}(1 - f_\uparrow(T + T_p + f_\uparrow^{-1}(V_{th}))) + f_\downarrow^{-1}(1 - V_{th}) \end{aligned} \tag{7.7}$$

If $V_{th} = 0.5$ is plugged into the definitions above, we obtain $\delta_\uparrow(T) = \delta_\downarrow(T) = \delta(T)$. By plugging in $-\delta(T)$ into the resulting definition of $\delta(T)$, it is easy to verify that $-\delta(T)$ is indeed an involution.

We conclude this section with the observation that the model used in [77], which uses a first-order RC low-pass filter for the slew rate limiter, is actually a particular simple instance of an involution channel: It produces transitions with $f_\uparrow(t) = f_\downarrow(t) = f(t) = 1 - e^{-(t/\tau)}$, where τ is the RC constant. The inverse is $f^{-1}(u) = -\tau \ln(1 - u)$, which leads to the following δ-functions:

$$\delta_\uparrow(T) = \tau \ln(1 - e^{-(T + T_p - \tau \ln(V_{th}))/\tau}) + T_p - \tau \ln(1 - V_{th}) \tag{7.8}$$

$$\delta_\downarrow(T) = \tau \ln(1 - e^{-(T + T_p - \tau \ln(1 - V_{th}))/\tau}) + T_p - \tau \ln(V_{th}) \tag{7.9}$$

In the sequel, we call these channels *exp-channels*.

Lemma 7.12. *An exp-channel is strictly causal if and only if $T_p > 0$. For exp-channels, $\delta_{\min} = T_p$.*

Chapter 8

Unbounded Short Pulse Filtration in Binary Models

8.1 Unsolvability of Unbounded Short Pulse Filtration with Constant Delay Channels

In this section, we show that no circuit whose channels are all positive constant delay channels solves SPF. The idea of the proof is to exploit the fact that the value of the output signal of the circuit at each time t only depends on a *finite* number of values of the input signal at times t' between 0 and t.

Calling each such time t' a *measure point* for time t, we show that indeed only a finite number of measure points exists for time t, i.e., the circuit cannot distinguish two different input signals that do not differ in the input signal values at the measure points for time t: For both such input signals, the output signal must have the same value at time t. Combining that indistinguishability result with a shifting argument of the input signal allows us to construct an arbitrary short pulse at the output of the circuit, a contradiction to property (F4) of Short Pulse Filtration.

8.1.1 Dependence Graphs

For each constant delay circuit with a single input port and a single output port, we introduce its *dependence graph*, which describes the way the output signals may depend on the input signals.

Let $C = (G, I, O, c, m)$ be a circuit with constant delay channels, a single input port i, and a single output port o. For every channel $c_{u,v}$ of C, denote by $\delta(u, v)$ its delay parameter δ and by $x(u, v)$ its initial value. The *dependence graph* $DG(t)$ of C at time t is a directed graph with vertices (v, τ), where v is a vertex in G and τ a time. It is defined as follows:

- The pair $(o, 0)$ is a vertex of $DG(t)$.
- If (v, τ) is a vertex of $DG(t)$ and (u, v) is an edge in G such that $\tau + \delta(u, v) \leq t$, then the pair $\big(u, \tau + \delta(u, v)\big)$ is also a vertex of $DG(t)$ and there is an edge in $DG(t)$ from $\big(u, \tau + \delta(u, v)\big)$ to (v, τ).

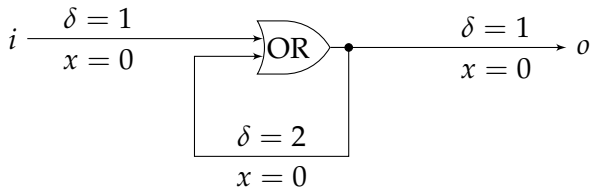


Figure 8.1: Example circuit

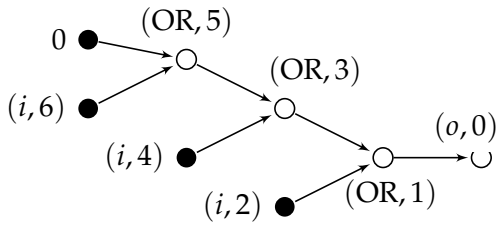


Figure 8.2: Example dependence graph $DG(6)$

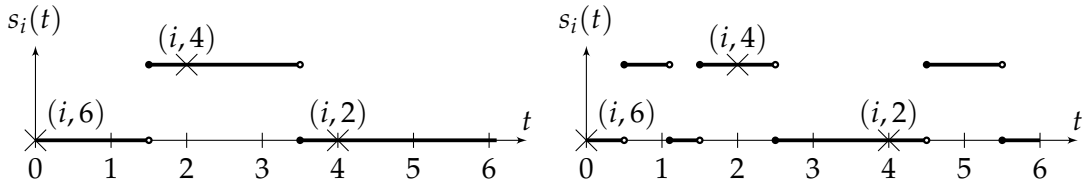


Figure 8.3: Input pulses with measure points ($\times$), labeled with the corresponding input leaf names

- If (v, τ) is a vertex of $DG(t)$ and (u, v) is an edge in G such that $\tau + \delta(u, v) > t$, then $c_{u,v}$'s initial value $x(u, v)$ is a vertex of $DG(t)$ and there is an edge in $DG(t)$ from $x(u, v)$ to (v, τ).

Because all $\delta(u, v)$ are strictly positive, the dependence graphs are finite and acyclic. A vertex of $DG(t)$ without incoming neighbors is a *leaf*, all others *intermediate vertices*. A vertex of the form (i, τ), with $i \in I$, is an *input leaf* and we call the time $t - \tau$ the corresponding *measure point* for time t. If $DG(t) = DG(\tilde{t})$, then the measure points for t are exactly the measure points for $\tilde{t}$ shifted by the difference $t - \tilde{t}$. All leaves of $DG(t)$ are either input leaves or elements of $\{0, 1\}$ (initial values of channels).

As an example, consider the circuit shown in Figure 8.1. The dependence graph $DG(6)$ is shown in Figure 8.2. Leaves are depicted as filled nodes, while intermediate nodes are empty. From the construction of the graph, we immediately see that in each execution the output signal value $s_o(6)$ only depends on the (input) signal values $s_i(4)$, $s_i(2)$, and $s_i(0)$. Thus, in particular, $s_o(6)$ is the same for both input signals depicted in Figure 8.3.

Generalizing the observations from the example, we thus observe:

Lemma 8.1. *The value of the output signal at time t only depends on the values of the input signal at the measure points for time t, according to $DG(t)$.*

Furthermore, if $DG(t) = DG(\tilde{t})$ and the values of input signals s_i and $\tilde{s}_i$ coincide at the respective measure points for t and $\tilde{t}$, then the respective output signals fulfill $s_o(t) = \tilde{s}_o(\tilde{t})$.

Proof. For a path π in G, denote by $\delta(\pi)$ the sum of delays $\delta(u,v)$ over all edges (u,v) of π. For every vertex v of G and every time $t \in \mathbb{R}_+$, let $\mathcal{P}(\to y, t)$ be the set of maximum length paths π ending in v such that $\delta(\pi) \leq t$.

It is clear, by the channel algorithm, that the value of $s_v(t)$ is uniquely determined by the collection of values $s_u\big(t - \delta(\pi)\big)$ where u is the start vertex of $\pi \in \mathcal{P}(\to v, t)$. Moreover, by maximality of π, if $u \neq i$, then $s_u\big(t - \delta(\pi)\big)$ only depends on the initial values of channels of incoming edges to u. Hence $s_v(t)$ is uniquely determined by the collection of values $s_i\big(t - \delta(\pi)\big)$ where $\pi \in \mathcal{P}(\to y, t)$ starts at i. This holds in particular for $v = o$. □

This lemma immediately shows that circuits with positive constant delay channels have unique executions:

Lemma 8.2. *If C is a circuit with only constant delay channels, then for all assignments of input signals $(s_i)_{i \in I}$ there exists a unique execution of C extending this assignment.*

Due to the fact that there are only finitely many measure points for a given time t, they are discrete and hence there is always a small margin until a new measure point appears:

Lemma 8.3. *For every time $t \in \mathbb{R}_+$, there exists an $\varepsilon > 0$ such that $DG(t) = DG(t + \varepsilon')$ for all $0 \leq \varepsilon' \leq \varepsilon$.*

Proof. Let $\varepsilon > 0$ be smaller than all positive values of the form $\delta(u,v) + \tau - t$ where (v,τ) is an intermediate vertex of $DG(t)$ and (u,v) is an edge in G. If no such intermediate vertex or edge exists, choose $\varepsilon > 0$ arbitrarily.

Let (v,τ) be an intermediate vertex of $DG(t)$ and (u,v) be an edge in G. If $t + \varepsilon - \tau < \delta(u,v)$, then clearly $t - \tau < \delta(u,v)$, because $\varepsilon > 0$. On the other hand, if $t - \tau < \delta(u,v)$, then $\delta(u,v) + \tau - t$ is positive and hence $\delta(u,v) > t + \varepsilon - \tau$ by choice of ε. Thus, the conditions $t - \tau < \delta(u,v)$ and $t + \varepsilon - \tau < \delta(u,v)$ are equivalent. This shows that the two dependence graphs $DG(t)$ and $DG(t + \varepsilon)$ and hence all dependence graphs in between are equal. □

8.1.2 Unsolvability Proof

Assume by contradiction that C solves SPF. By the nontriviality property (F3), there exists an input pulse such that the corresponding output signal is non-zero, i.e., there exists an input pulse of some length and a time t such that the corresponding output signal's value at time t is 1.

By Lemma 8.3, there exists an $\varepsilon > 0$ such that $DG(t) = DG(t + \varepsilon)$. We may choose ε arbitrarily small, in particular strictly smaller than all differences of distinct measure points for time t.

Clearly, $DG(\tilde{t}) = DG(t)$ for all times $\tilde{t}$ between t and $t + \varepsilon$, in particular, for $\tilde{t} = t + \varepsilon/2$. Denote by Δ the infimum of input pulse lengths (where all pulses start at the same time) such that the corresponding output signal's value at time $\tilde{t}$ is 1. This infimum is finite by the choice of t and $\tilde{t}$. There hence exists an input pulse p with the above property of length at most $\Delta + \varepsilon/4$. We show that its corresponding output signal s_p contains a pulse of length strictly less than ε, in contradiction to the no short pulses property (F4).

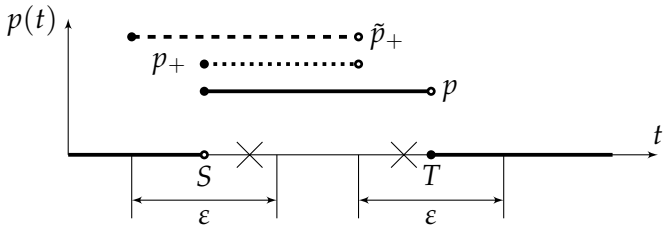


Figure 8.4: Input pulse p, together with its derived pulses p_+ and $\tilde{p}_+$, and measure points for time $\tilde{t}$

Denote by S the time of p's rising transition and by T the time of p's falling transition. Now let p_+ be the pulse whose rising transition is at time S and whose falling transition is at time $T - \varepsilon/2$. If $S \geq T - \varepsilon/2$, then let p_+ be the zero signal instead. The length of p_+ is either strictly less than Δ or it is the zero signal. Hence, by the definition of the no generation property (F2), its corresponding output signal's value at time $\tilde{t}$ is 0. This implies that there exists a measure point for time $\tilde{t}$ within $[T - \varepsilon/2, T)$, because p and p_+ coincide everywhere else (see marked measure point on the right in Figure 8.4).

Because we chose ε to be smaller than all differences of distinct measure points for time t (and hence also for time $\tilde{t}$), we see that there is no measure point for $\tilde{t}$ in the interval $[T, T + \varepsilon/2)$.

Likewise, by defining p_- as the pulse with rising transition at time $S + \varepsilon/2$ and falling transition at time T, we infer that there is one measure point for time $\tilde{t}$ in the interval $[S, S + \varepsilon/2)$ and there is no measure point for $\tilde{t}$ in the interval $[S - \varepsilon/2, S)$ (see Figure 8.4).

Now consider the pulse $\tilde{p}_+$ generated by shifting pulse p into the past by $\varepsilon/2$, i.e., $\tilde{p}_+$'s rising transition is at time $S - \varepsilon/2$ and its falling transition is at $T - \varepsilon/2$. Because $\tilde{p}_+$ coincides with p_+ at all measure points for $\tilde{t}$, the output signal $s^{\tilde{p}_+}$ corresponding to $\tilde{p}_+$ has value 0 at time $\tilde{t}$. Because $DG(\tilde{t}) = DG(\tilde{t} + \varepsilon/2)$, the second part of Lemma 8.1 shows that $s^{\tilde{p}_+}(\tilde{t} + \varepsilon/2) = 0$.

Likewise, by considering p shifted into the future by $\varepsilon/2$, we see that also $s^{\tilde{p}_+}(\tilde{t} - \varepsilon/2) = 0$. But because $s^p(\tilde{t}) = 1$, this shows that the output signal s^p contains a pulse of length strictly less than ε. Since ε can be chosen arbitrarily small, this concludes the proof.

8.2 Solvability of Unbounded Short Pulse Filtration with Involution Channels

In this section, we show that unbounded SPF is solvable in our circuit model with strictly causal involution channels. We do this by verifying that the circuit shown in Figure 8.5, which consists of a fed back OR-gate and a high-threshold filter (implemented by a channel), indeed solves SPF. In order not to obfuscate the essentials, we restrict our attention to certain classes of involution channels. More specifically, in our proof, the channel in the feed-back loop must be strictly causal and symmetric, i.e., $\delta_\uparrow = \delta_\downarrow = \delta$. When using an exp-channel, for example, this implies a threshold $V_{th} = 0.5$. The channel implementing the high-threshold filter is assumed to be an exp-channel, because we have to adjust its parameters appropriately. However, the proof could be adapted to show the possibility of unbounded SPF with any given class of strictly causal involution channels.

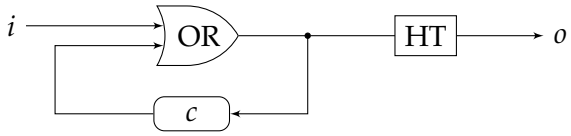


Figure 8.5: A circuit solving unbounded SPF, consisting of an OR-gate fed back by channel c, and a high-threshold filter HT

We consider a pulse of length Δ at time 0 at the input and reason about the behavior of the feed-back loop. Then, we show that this behavior can be translated to a legitimate SPF output by using a high-threshold buffer. We start by identifying two extremal cases: If Δ is very small, then the pulse is filtered by the channel in the feed-back loop, if it is very big, the pulse is captured by the storage loop, leading to a stable output 1.

Lemma 8.4. *If the input pulse's length Δ satisfies $\Delta \geq \delta_\infty$, then the output of the OR has a unique rising transition at time δ_∞.*

Proof. Assigning the channel output s_c a single rising transition at time δ_∞ is part of a consistent execution, in which the OR's output has a single rising transition at time 0. The lemma now follows from uniqueness of executions. □

Lemma 8.5. *If the input pulse's length Δ satisfies $\Delta \leq \delta_\infty - \delta_{\min}$, then the OR output contains only the input pulse.*

Proof. The input signal contains only two transitions: One at time $t_1 = 0$ and one at time $t_2 = \Delta \leq \delta_\infty - \delta_{\min}$. Since $\delta_1 = \delta_\infty$ and hence $t_2 = t_1 + \Delta \leq t_1 + \delta_1 - \delta_{\min}$, the two pending transitions of c's output cancel by Lemma 7.7, and no further transitions are generated afterwards. □

Now suppose that the input pulse length satisfies $\delta_\infty - \delta_{\min} < \Delta_0 < \delta_\infty$. For these pulse lengths Δ_0, the output signal will contain a series of pulses of lengths $\Delta_0, \Delta_1, \Delta_2, \ldots$ For all but one Δ_0, this series will turn out to be finite and the output signal will either eventually be 0 or eventually 1. To compute these pulse lengths, we define the auxiliary function

$$f(\Delta) = \delta\big(\Delta - \delta(-\Delta)\big) + \Delta - \delta(-\Delta) \ , \tag{8.1}$$

which gives $\Delta_n = f(\Delta_{n-1})$ for all $n \geq 2$. To see this, note that Δ_{n-1} at the channel input is also present at the channel output, so the rising resp. falling transition is delayed by $\delta(-\Delta_{n-1})$ resp. $\delta(\Delta_{n-1} - \delta(-\Delta_{n-1}))$. The first generated pulse starts from a zero channel input and thus fulfills

$$\Delta_1 = \Delta_0 - \delta_\infty + \delta(\Delta_0 - \delta_\infty) \ . \tag{8.2}$$

The procedure stops if either $f(\Delta_n) \leq 0$ (pulse canceled; the output is constant 0 thereafter) or $f(\Delta_n) \geq \delta(0) > 0$ (pulse captured; the output is constant 1 thereafter).

The only case in which the procedure does not stop is if $f(\Delta_1) = \Delta_1$. There is a unique $\Delta_1 > 0$ with this property, denoted $\tilde{\Delta}_1$. By (8.1), it is also characterized by the relation $\delta(-\tilde{\Delta}_1) = 2\tilde{\Delta}_1$. Since $\delta(-\delta(0)) = 0$ by the involution property, we must have $\tilde{\Delta}_1 < \delta(0)$. Since $\Delta_1 \to \delta(0)$ as $\Delta_0 \to \delta_\infty$ and $\Delta_1 \to 0$ as $\Delta_0 \to \delta_\infty - \delta_{\min}$, there exists a unique Δ_0 such that $\Delta_1 = \tilde{\Delta}_1$. Denote it by $\tilde{\Delta}_0$.

The following lemma shows that the procedure indeed stops if and only if $\Delta_1 \neq \tilde{\Delta}_1$, and can be used to bound the number of steps until it stops.

Lemma 8.6. $|f(\Delta_1) - \tilde{\Delta}_1| \geq (1 + \delta'(0)) \cdot |\Delta_1 - \tilde{\Delta}_1|$ *if* $\Delta_1 > 0$.

Proof. We have

$$\begin{aligned} f'(\Delta_1) =& \big(1 + \delta'(-\Delta_1)\big) \cdot \delta'\big(\Delta_1 - \delta(-\Delta_1)\big) \\ &+ 1 + \delta'(-\Delta_1) \geq 1 + \delta'(0) \end{aligned} \tag{8.3}$$

because $\delta'(-\Delta_1) \geq \delta'(0)$ and $\delta'(T) > 0$ for all T as δ is concave and increasing. The mean value theorem of calculus now implies the lemma. □

Theorem 8.7. *The fed-back OR gate with a strictly causal symmetric involution channel has the following output when the input pulse has length* Δ_0*:*

- *If* $\Delta_0 > \tilde{\Delta}_0$*, then the output is eventually constant* 1*.*
- *If* $\Delta_0 < \tilde{\Delta}_0$*, then the output is eventually constant* 0*.*
- *If* $\Delta_0 = \tilde{\Delta}_0$*, then the output is a periodic pulse train with duty cycle 50%.*

Furthermore, the stabilization time in the first two cases is in the order of $\log 1/|\Delta_0 - \tilde{\Delta}_0|$.

Proof. If $\Delta_0 \geq \delta_\infty$ or $\Delta_0 \leq \delta_\infty - \delta_{\min}$, then Lemmas 8.4 and 8.5 show the theorem.

So let $\Delta_0 \in (\delta_\infty - \delta_{\min}, \delta_\infty)$. By Lemma 8.6, the number of generated pulses until the procedure stops is in the order of $\log 1/|\Delta_1 - \tilde{\Delta}_1|$. Setting $g(\Delta_0) = \Delta_0 - \delta_\infty + \delta(\Delta_0 - \delta_\infty)$, and applying the mean value theorem of calculus to this function, we see analogously as in the proof of Lemma 8.6 that

$$|\Delta_1 - \tilde{\Delta}_1| \geq \big(1 + \delta'(0)\big) \cdot |\Delta_0 - \tilde{\Delta}_0| \ . \tag{8.4}$$

Hence the number of generated pulses is in the order of $\log 1/|\Delta_0 - \tilde{\Delta}_0|$. Since both the length Δ_n of the occurring pulses and, by symmetry, the time between them is at most $\delta(0)$, we have the same asymptotic bound on the stabilization time. □

We now turn to the analysis of the high-threshold filter.

Lemma 8.8. *Let* c *be an exp-channel* c *with threshold* V_{th}*. Then there exists some* $\Delta > 0$ *such that every periodic pulse train with pulse lengths at most* Δ *and duty cycle (ratio of 1-to-0) at most* V_{th} *is mapped to the zero signal by* c*.*

Proof. Let $t_1, t_2, \ldots$ be the times of transitions in the input pulse train with duty cycle $\gamma \leq V_{th}$, i.e., $t_1 = 0$, $t_{2n+2} = t_{2n+1} + \Delta$, and $t_{2n+1} = t_{2n} + \Delta/\gamma$. We assume that Δ is smaller than both $-\tau \log(1 - V_{th})$ and a to-be-determined Δ_0. We inductively show that all pulses get canceled:

If $\Delta \leq -\tau \log(1 - V_{th})$, then the first pulse is canceled and $\delta_2 \leq T_p$. If $\delta_{2n} \leq T_p$, then

$$\begin{aligned} \delta_{2n+1} &= \delta_\uparrow(\Delta/\gamma - \delta_{2n}) \geq \delta_\uparrow(\Delta/V_{th} - T_p) \\ &= T_p - \tau \log(1 - V_{th}) + \tau \log(1 - V_{th} e^{-\Delta/V_{th}\tau}) \end{aligned} \tag{8.5}$$

Hence t_{2n+1} and $t_{2n+2} = t_{2n+1} + \Delta$ cancel if

$$\begin{aligned} \Delta &\leq \delta_{2n+1} - T_p \\ &= -\tau \log(1 - V_{th}) + \tau \log(1 - V_{th} e^{-\Delta/V_{th}\tau}) \ , \end{aligned} \tag{8.6}$$

which is equivalent to

$$h(\Delta) = V_{th} e^{-\Delta/V_{th}\tau} + (1 - V_{th}) e^{\Delta/\tau} \leq 1 \ . \tag{8.7}$$

It is $h(0) = 1$ and

$$h'(\Delta) = -\frac{1}{\tau} e^{-\Delta/V_{th}\tau} + \frac{1 - V_{th}}{\tau} e^{\Delta/\tau} \ , \tag{8.8}$$

in particular $h'(0) < 0$. There hence exists some $\Delta_0 > 0$ such that $h(\Delta) \leq 1$ for all $0 \leq \Delta \leq \Delta_0$. In particular, t_{2n+1} and t_{2n+2} cancel. Also,

$$\delta_{2n+2} = \delta(\Delta - \delta_{2n+1}) \leq \delta(-T_p) = T_p \tag{8.9}$$

because $h(\Delta) \leq 1$. We can hence continue the induction. □

By letting τ grow, one can even achieve the following result.

Lemma 8.9. *Let $\Delta > 0$ and $0 < \gamma < 1$. Then there exists an exp-channel with threshold $V_{th} = \gamma$ such that every periodic pulse train with pulse lengths at most Δ and duty cycle at most γ is mapped to the zero signal by c.*

Proof. We use the notation of the proof of Lemma 8.8. The unique root of $h'(\Delta)$ is equal to

$$\Delta_\tau = -\frac{\tau \log(1 - V_{th})}{1 + 1/V_{th}} \ , \tag{8.10}$$

which goes to infinity as $\tau \to \infty$. We can choose $\Delta_0 = \Delta_\tau$ because $h'(\Delta) \leq 0$ for all $0 \leq \Delta \leq \Delta_\tau$. Because also $-\tau \log(1 - V_{th})$ goes to infinity as $\tau \to \infty$, we can find, for any given Δ, some $\tau > 0$ such that both $\Delta \leq -\tau \log(1 - V_{th})$ and $\Delta \leq \Delta_\tau$. But for these Δ, all input pulse trains with pulse lengths Δ and duty cycle at most $V_{th} = \gamma$ get mapped to the zero signal. □

In particular, by choosing $\gamma = 0.6$ and Δ large enough such that the output of the feedback loop is already constant 1 at time $T + \Delta$ if the duty cycle in the loop passes 0.6 at time T, we show:

Theorem 8.10. *There is a circuit that solves unbounded SPF.*

8.3 Eventual Short Pulse Filtration with Constant Delay Channels

We proved that SPF is not solvable with constant delay channels. In this section, we consider the weaker eventual SPF problem, which drops the "no short pulses" requirement (F4) and replaces it with an eventual analogue (F4e): A circuit *solves eventual SPF* if conditions (F1)–(F3) and the following condition hold:

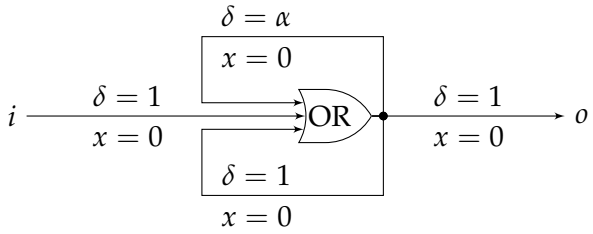


Figure 8.6: Circuit C_{ev} solving eventual SPF

F4e) There exists an $\varepsilon > 0$ and a $K > 0$ such that, in all executions with a pulse at time T as the input signal, the output signal does not contain a pulse of length less than ε after time $T + K$. *(Eventually no short pulses)*

We show that eventual SPF is solvable using only constant delay channels. More specifically, we prove that circuit C_{ev} in Figure 8.6 solves eventual SPF. The circuit contains a delay parameter α, which we will choose to be a positive irrational like $\alpha = \sqrt{2}$.

We will show that the circuit's output is eventually stable at 1 whenever the input is a pulse of positive length. We derive a bound on this stabilization time in terms of the input pulse length Δ. The bound is almost linear in $1/\Delta$: It is in the order of $O(\Delta^{-1-\varepsilon})$ for all $\varepsilon > 0$.

The measure points of circuit C_{ev} for time t are of the form $t - (\alpha k + \ell) - 2$, where k and ℓ are nonnegative integers. We can hence characterize the circuit's behavior with the following obvious lemma.

Lemma 8.11. *In every execution (s_v) of circuit C_{ev}, the following are equivalent: (i) $s_o(t) = 1$, and (ii) there exist nonnegative integers k and ℓ such that $s_i\big(t - (\alpha k + \ell) - 2\big) = 1$.*

We may restrict our considerations to input pulses starting at time 0. In the following, let the input signal s_i be a pulse of length $\Delta > 0$. We are looking for the *stabilization time*, which is the minimal time $T = T(\Delta)$ such that, for all $t \geq T$, we have $s_o(t) = 1$.

To prove finiteness and effective bounds on the stabilization time, we relate it to the number-theoretic concept of *discrepancy* of the sequence (αn) modulo 1 (see, e.g., [38]). The discrepancy compares the number of sequence elements in a given interval with their expected number if the elements were uniformly distributed.

For a given nonempty subinterval $(x, y] \subseteq (0, 1]$ and a given positive integer N, denote by $A(x, y; N)$ the number of αn's with $n \leq N$ that lie in the interval modulo 1: $\alpha n \in (x, y] + \mathbb{Z}$. The expected number of such αn's is $(y - x)N$. The discrepancy $D_N(\alpha)$ is then defined as the maximum difference between $A(x, y; N)$ and $(y - x)N$, formed over all nonempty subintervals $(x, y]$ of $(0, 1]$.

It is well-known that $D_N(\alpha)/N \to 0$ if and only if α is irrational. Also, if α has a bounded continued fraction expansion, then $D_N(\alpha) = O(\log N)$ and the constant can be computed [79]. This is, in particular, true for $\alpha = \sqrt{2}$.

Lemma 8.12. *Let $K = K(\Delta)$ be the least integer K such that for all real t there exists an integer k, $0 \leq k \leq K$, with $\alpha k \in (t - \Delta, t] + \mathbb{Z}$. Then, $T(\Delta) \leq \alpha \cdot K(\Delta) + \Delta + 2$.*

Proof. The lemma is trivial if $K = \infty$, so assume the contrary.

Let $t \geq \alpha K + \Delta + 2$. By the definition of K, there exists a k with $0 \leq k \leq K$ and an ℓ such that $t - \Delta - \ell - 2 < \alpha k \leq t - \ell - 2$, which is equivalent to $0 \leq t - (\alpha k + \ell) - 2 < \Delta$.

By Lemma 8.11, it remains to prove that ℓ is nonnegative. The inequality $t - (\alpha k + \ell) - 2 < \Delta$ is equivalent to $\ell > t - \Delta - \alpha k - 2$. Noting $-\alpha k \geq -\alpha K$ and $t \geq \alpha K + \Delta + 2$ shows $\ell > 0$ and concludes the proof. □

Lemma 8.13. *Let $0 < \Delta \leq 1$. If $D_N(\alpha)/N < \Delta/2$, then $K(\Delta) \leq N$.*

Proof. Suppose the contrary, i.e., that there exists a real t such that, for all $n \leq N$, we have $\alpha n \notin (t - \Delta, t] + \mathbb{Z}$. Let $0 < x < y \leq z < u \leq 1$ such that we can decompose the interval $(t - \Delta, t] + \mathbb{Z} = \big((x, y] + \mathbb{Z}\big) \cup \big((z, u] + \mathbb{Z}\big)$ modulo 1. None of the two intervals $(x, y]$ and $(z, u]$ contains an αn modulo 1 with $n \leq N$. Hence $A(x, y; N) = A(u, z; N) = 0$, which implies $2D_N(\alpha) \geq (y - x)N + (u - z)N = \Delta N$, a contradiction. □

Theorem 8.14. *Circuit C_{ev} solves eventual SPF if α is irrational. If $\alpha = \sqrt{2}$, then the stabilization time satisfies $T(\Delta) = O(\Delta^{-1-\varepsilon})$ as $\Delta \to 0$ for all $\varepsilon > 0$.*

Proof. (F1) is obviously fulfilled. Because all initial values of channels are 0, also (F2) holds. Because $D_N(\alpha)/N \to 0$ whenever α is irrational, for all $\Delta > 0$, there exists some N such that $D_N(\alpha)/N < \Delta/2$. Hence Lemma 8.13 and Lemma 8.12 show that $T(\Delta)$ is finite, which shows (F3) and (F4e).

We now prove the bound on the stabilization time. Let $\gamma = -1 - \varepsilon < -1$. There exists a $C_1 > 0$ such that $D_N(\alpha) \leq C_1 \log N$. Because $1 + 1/\gamma > 0$, there exists a $C_2 > 0$ such that $\log N < C_2 N^{1+1/\gamma}$. Thus if

$$N \geq \left(\frac{\Delta}{2C_1C_2}\right)^{\gamma} \tag{8.11}$$

then

$$\frac{D_N(\alpha)}{N} \leq \frac{C_1 \log N}{N} < C_1C_2N^{1/\gamma} \leq \frac{\Delta}{2}\ , \tag{8.12}$$

which, by Lemma 8.13, implies

$$K(\Delta) \leq \left(\frac{\Delta}{2C_1C_2}\right)^{\gamma} + 1 \tag{8.13}$$

for all $0 < \Delta \leq 1$. That is, $K(\Delta) = O(\Delta^{\gamma})$ as $\Delta \to 0$.

It is easy to see that $K(\Delta) \to \infty$ as $\Delta \to 0$. Hence Lemma 8.12 implies $T(\Delta) = O(K(\Delta))$ as $\Delta \to 0$, as claimed. □

Chapter 9

Bounded Short Pulse Filtration in Binary Models

9.1 Unsolvability of Bounded Short Pulse Filtration with Involution Channels

9.1.1 Continuity of Involution Channels

In this subsection, we prove that strictly causal involution channels are continuous in a certain sense that we will define precisely. For ease of exposition, we give the proof only in the case of symmetric channels, i.e., for the case that $\delta_\uparrow = \delta_\downarrow = \delta$.

We begin by noting that channels are monotone. To compare two signals, we write $s_1 \leq s_2$ if s_2 is 1 whenever s_1 is.

Lemma 9.1. *Let s_1 and s_2 be signals such that $s_1 \leq s_2$ and let c be a channel. Then $c(s_1) \leq c(s_2)$.*

We next define a distance for signals, for which channels will turn out to be continuous.

Definition 9.2. For a signal s and a time T, denote by $\mu_T(s)$ the combined amount in $[0, T]$ that s is 1. In more symbolic terms, $\mu_T(s)$ is the measure of the set $\{t \in [0, T] \mid s(t) = 1\}$.

For any two signals s_1 and s_2 and every T, we define their *distance up to time T* by setting $\|s_1 - s_2\|_T = \mu_T(|s_1 - s_2|)$.

With the next lemma, we identify an optimal choice for adding a pulse to the end of a signal when wanting to maximize μ_T. We will use it later when bounding the maximum impact an infinitesimally small pulse can have.

We use the shorthand notation $(x)_+$ to mean $\max\{x, 0\}$.

Lemma 9.3. *Let s be a signal that is eventually constant 0 and let c be a channel. Denote by t_n the time of the last (falling) transition in s and by δ_n its delay in the channel algorithm for c. Then the maximal $\mu_T(c(s'))$ among all s' obtained from s by adding one pulse of length Δ after time t_n is attained by the addition of the pulse at time $t_n + (\delta_n - \delta_{\min})_+$.*

Proof. We first show the lemma for $T = \infty$ and then extend the result to finite T. Let s'_γ be the addition of the pulse of length Δ to s at time $t_n + \gamma$.

For all $0 \leq \gamma \leq \delta_n - \delta_{\min}$, set

$$f(\gamma) = \gamma + \delta(\Delta - \delta(\gamma - \delta_n)) \ .$$

In the class of all s'_γ with $\gamma \leq \delta_n - \delta_{\min}$, the maximum of $\mu_\infty(c(s'_\gamma))$ is attained at the maximum of f. This is because the transition at time $t_n + \gamma$ cancels that at time t_n in this case. The derivative of f is equal to

$$f'(\gamma) = 1 - \delta'(\Delta - \delta(\gamma - \delta_n)) \cdot \delta'(\gamma - \delta_n) \ .$$

The condition $f'(\gamma) = 0$ is equivalent to $\delta'(\Delta - \delta(\gamma - \delta_n)) = 1/\delta'(\gamma - \delta_n)$, which is equivalent to $\delta(\Delta - \delta(\gamma - \delta_n)) = -(\gamma - \delta_2)$, or $\Delta = 0$. Hence $f'(\gamma)$ is never zero. Since $f'(\gamma) \to 1$ as $\gamma \to \infty$, the derivative of f is always positive, hence f is increasing. This shows that $\gamma = \delta_n - \delta_{\min}$ is a strictly better choice than any other γ in this class.

For the class of s'_γ with $\gamma \geq (\delta_n - \delta_{\min})_+$, we define the function

$$g(\gamma) = \Delta + \delta(\Delta - \delta(\gamma - \delta_n)) - \delta(\gamma - \delta_n) \ .$$

Since the transitions at t_n and $t_n + \gamma$ do not cancel in this class, the maximum of $\mu_\infty(c(s'_\gamma))$ is attained at the maximum of g. But it is easy to see, using the monotonicity of δ, that g is decreasing. The maximum of g is hence attained at $\gamma = (\delta_n - \delta_{\min})_+$.

We have shown that the choice $\gamma = \gamma_0 = (\delta_n - \delta_{\min})_+$ maximizes $\mu_\infty(c(s'_\gamma))$, which concludes the proof for $T = \infty$

Let now T be finite. Denote by T_0 the time of the last, falling, output transition in $c(s'_{\gamma_0})$. In this case, transitions of $c(s)$ and $c(s'_{\gamma_0})$ are the same except the last, falling, transition which is delayed from $t_n + \delta_n$ to T_0. We distinguish the two cases (a) $T \leq T_0$ and (b) $T > T_0$. In case (a), the last transition of $c(s)$ is delayed beyond T in $c(s'_{\gamma_0})$. Because all other transitions remain unchanged in all $c(s'_\gamma)$, the measure $\mu_T(c(s'_{\gamma_0}))$ is maximal among all $\mu_T(c(s'_\gamma))$ if $T \leq T_0$. In case (b), we have $\mu_T(c(s'_{\gamma_0})) = \mu_\infty(c(s'_{\gamma_0}))$. But because $\mu_T \leq \mu_\infty$ and $\mu_\infty(c(s'_{\gamma_0}))$ is maximal among all $\mu_\infty(c(s'_\gamma))$, so is $\mu_T(c(s'_{\gamma_0}))$ among all $\mu_T(c(s'_\gamma))$. □

We next effectively bound the maximum impact on μ_T that a set of pulses of small combined length can have.

Lemma 9.4. *Let s be a signal that is eventually constant 0 and let c be a channel. Then there exists a constant d such that the maximal $\mu_T(c(s'))$ among all s' obtained from s by adding pulses of combined length ε after the last transition of s is at most $\mu_T(c(s)) + d \cdot \varepsilon$.*

Proof. It suffices to show the lemma for $T = \infty$. Let $\varepsilon = \sum_{k=1}^{\infty} \varepsilon_k$. We add, one after the other, pulses of length ε_k after the last transition. We show that the maximum gain after adding K pulses is at most $\sum_{k=1}^{K} \varepsilon_k$.

Denote by t_n the last transition in s and by δ_n its delay. By Lemma 9.3, it is optimal to add the first pulse (of length ε_1) at time $t_n + (\delta_n - \delta_{\min})_+$; call the resulting signal s'_1.

We first assume $\delta_n - \delta_{\min} \geq 0$. Here, the two new transitions in s'_1 are $t_{n+1} = t_n + \delta_n - \delta_{\min}$ and $t_{n+2} = t_n + \delta_n - \delta_{\min} + \varepsilon_1$. Their corresponding delays are $\delta_{n+1} = \delta_{\min}$ and $\delta_{n+2} = \delta(\varepsilon_1 - \delta_{\min})$. By the mean value theorem of calculus and Lemma 7.6, we have

$$\delta_{n+2} - \delta_{n+1} = \delta(\varepsilon_1 - \delta_{\min}) - \delta(-\delta_{\min}) = \varepsilon_1 \cdot \delta'(\xi) \tag{9.1}$$

for some $-\delta_{\min} \leq \xi \leq \varepsilon_1 - \delta_{\min}$. Since δ' is increasing and $\delta'(-\delta_{\min}) = 1$, we hence deduce $0 \leq \delta_{n+2} - \delta_{n+1} \leq \varepsilon_1$. Thus $\mu_T(c(s'_1) - c(s)) = \varepsilon_1 + \delta_{n+2} - \delta_{n+1} \leq 2\varepsilon_1$. Since $\delta_{n+2} > \delta_{\min}$, we can continue this argument inductively.

If now $\delta_n - \delta_{\min} < 0$, then t_n is replaced by $t_n + \varepsilon_1$ in s'_1. This changes the measure by

$$\big(\delta(t_n - t_{n-1} + \varepsilon_1 - \delta_{n-1}) - \delta(t_n - t_{n-1} - \delta_{n-1})\big)_+ \ , \tag{9.2}$$

which is at most $\varepsilon_1 \cdot \delta'(t_n - t_{n-1} - \delta_{n-1})$ by the mean value theorem. We note that this second case only occurs until the first case happens one time. We can hence merge all the ε_k of the first case and set $d = \max(2, \delta'(t_n - t_{n-1} - \delta_{n-1}))$. □

We combine the previous two lemmas to show continuity of channels:

Theorem 9.5. *Let c be an involution channel and let $T \geq 0$. Then the mapping $s \mapsto c(s)$ is continuous with respect to the distance $d(s_1, s_2) = \|s_1 - s_2\|_T$.*

Proof. Let s be a signal. We show that, if $\|s - s_n\|_T \to 0$, then $\|c(s) - c(s_n)\|_T \to 0$. Because

$$|s - s_n| = (\max(s, s_n) - s) + (s - \min(s, s_n)) \ , \tag{9.3}$$

the condition $\|s - s_n\|_T \to 0$ is equivalent to the conjunction of $\|s - \max(s, s_n)\|_T \to 0$ and $\|s - \min(s, s_n)\|_T \to 0$. Furthermore, because $\max(c(s), c(s_n)) \leq c(\max(s, s_n))$ and $\min(c(s), c(s_n)) \geq c(\min(s, s_n))$ by Lemma 9.1, we have

$$\begin{aligned} |c(s) - c(s_n)| \leq\ & c(\max(s, s_n)) - c(s) \\ & + c(s) - c(\min(s, s_n)) \ , \end{aligned} \tag{9.4}$$

which shows that we can suppose without loss of generality $s_n \geq s$ for all n.

Let $(t_m, 0), (t_{m+1}, 1)$ be a negative pulse in s. Since there are only finitely many negative pulses before time T, it suffices to show $\mu_T(c(s_n) - c(s)) \to 0$ in the case that $s_n - s$ is zero outside of $[t_m, t_{m+1}]$, i.e., that the only additions of s_n with respect to s lie in the given negative pulse.

Let $\mu_T(s_n - s) \leq \varepsilon$. It follows from Lemma 9.4 that the increase in measure incurred directly from the new pulses is $O(\varepsilon)$. Furthermore, by Lemma 9.3, the measure incurred by later transitions t_k with $k > m$ are biggest when merging all new pulses at the end of the negative pulse. Because the delays of these transitions depend continuously on ε and $\mu_T(c(s_n) - c(s))$ depends continuously on these delays, we have $\mu_T(c(s_n) - c(s)) \to 0$ as $\varepsilon \to 0$. □

9.1.2 Unsolvability in Forward Circuits

We call a circuit a *forward circuit* if its graph is acyclic. Forward circuits are exactly those circuits that do not contain feed-back loops. From the fact that the composition of continuous functions is continuous, we obtain from Theorem 9.5:

Theorem 9.6. *No forward circuit with involution channels solves bounded SPF.*

Proof. Suppose that there exists a forward circuit that solves bounded SPF with stabilization time bound K. Denote by s_Δ its output signal when feeding it a Δ-pulse at time 0 as the input. Because s_Δ in forward circuits is a finite composition of continuous functions by Theorem 9.5, the measure $\mu_T(s_\Delta)$ depends continuously on Δ.

By the nontriviality condition (F3) of the SPF problem, there exists some Δ_0 such that s_{Δ_0} is not zero. Set $T = 2\Delta_0 + K$.

Let $\varepsilon > 0$ be smaller than $\mu_T(s_{\Delta_0})$. We show a contradiction by finding a Δ such that s_Δ either contains a pulse of length less than ε (contradiction to the no short pulses condition (F4)) or contains a transition after time $\Delta + K$ (contradicting the bounded stabilization time condition (F5)).

Since $\mu_T(s_\Delta) \to 0$ as $\Delta \to 0$ by the no generation condition (F2) of SPF, there exists a $\Delta_1 < \Delta_0$ such that $\mu_T(s_{\Delta_1}) = \varepsilon$ by the intermediate value property of continuity. By the bounded stabilization time condition (F5), there are no transitions in s_{Δ_1} after time $\Delta_1 + K$. Hence s_{Δ_1} is 0 after this time because otherwise it is 1 for the remaining duration $T - (\Delta_1 + K) > \Delta_0 > \varepsilon$, which would mean that $\mu_T(s_{\Delta_1}) > \varepsilon$. There thus exists a pulse in s_{Δ_1} before time $\Delta_1 + K$. But any such pulse is of length at most ε because $\mu_{\Delta_1+K}(s_{\Delta_1}) \leq \mu_T(s_{\Delta_1}) = \varepsilon$. This is a contradiction to the no short pulses condition (F4). □

9.1.3 Simulation with Unrolled Circuits

We next show how to simulate (part of) an execution of a circuit C by a forward circuit C' generated from C by unrolling of feedback channels. Intuitively, the deeper the unrolling, the longer the time C' behaves as C.

Definition 9.7. Let C be a circuit with input i. For v being a gate or input in C and $k \geq 0$, the *k-unrolled circuit* $C_k(v)$ is constructed inductively as follows: If $v = i$, or v is a gate with no predecessor in C, then $C_k(v)$ is the circuit that comprises only of vertex v and whose output is v. We slightly misuse the circuit definition here by allowing circuits with a single vertex. Otherwise, v is a gate with predecessors and we distinguish two cases:

If $k = 0$, $C_k(v)$ comprises of: gate $v^{(\alpha)}$, with α being a unique identifier, and for each predecessor σ of v in C: if $\sigma = i$, add i and an edge from i to $v^{(\alpha)}$; if σ is a channel, add channel $\sigma^{(\beta)}$ and gate $\tilde{x}^{(\gamma)}$, with β and γ being unique identifiers and x being the channel's initial value. Furthermore, add edges from $\tilde{x}^{(\gamma)}$ to $\sigma^{(\beta)}$ and from $\sigma^{(\beta)}$ to $v^{(\alpha)}$. The Boolean function assigned to $v^{(\alpha)}$ is the same as for v and the ordering of the predecessors of $v^{(\alpha)}$ reflects the ordering of the predecessors of v. The Boolean function assigned to $\tilde{x}^{(\alpha)}$ is constant x. The channel functions of $\sigma^{(\beta)}$ and σ are equal.

If $k > 0$, $C_k(v)$ is the circuit that comprises of gate $v^{(\alpha)}$, with unique identifier α, and for each predecessor σ of v in circuit C: If σ is a channel, let w be its predecessor in C. Add and connect the output of circuit $C_{k-1}(w)$ to a channel $\sigma^{(\beta)}$ and the channel to $v^{(\alpha)}$. If $\sigma = i$, add i and connect it to $v^{(\alpha)}$. Again, the Boolean functions, orderings and channel functions are assigned in accordance with those in C.

In all cases, we call a vertex $\sigma^{(\alpha)}$ *corresponding to* σ.

Let o be the single output of circuit C. To each vertex σ in $C_k(o)$, we assign a value $z(\sigma)$ from $\mathbb{N}_0 \cup \{\infty\}$ as follows: $z(\tilde{0}^{(\alpha)}) = z(\tilde{1}^{(\alpha)}) = 0$, $z(i) = z(\sigma) = \infty$ if σ has no predecessor in C, $z(\sigma) = 1 + z(w)$ for a channel σ with predecessor w, and $z(\sigma) = \min\{z(\sigma') \mid \sigma' \text{ is a predecessor of } \sigma\}$ for a gate σ. Figure 9.1 shows an example of a circuit and an unrolled circuit with z values assigned to inputs and gates.

We further adapt the constructive algorithm in Section 7.4 to assign to each generated transition a *causal depth $d(e)$ of transition e*. All initial transitions and input transitions have causal depth 0; all transitions initially added at time 0 have causal depth 1. Algorithm step (i) is extended such that each transition e at time t that was marked fixed in a set E_v, with v being

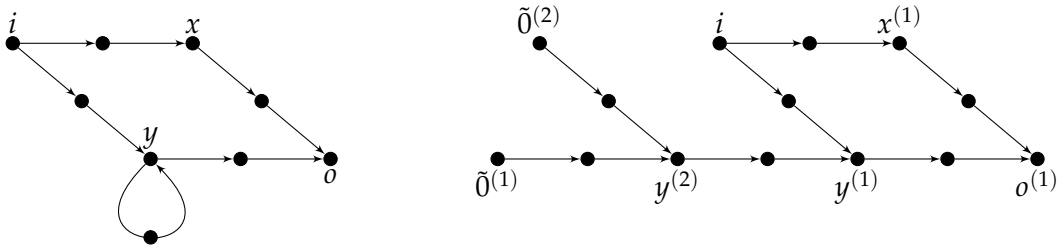


Figure 9.1: Circuit C (left) and $C_2(o)$ (right) under the assumption that both incoming channels to gate y have initial value 0; it is $z(\tilde{0}^{(1)}) = z(\tilde{0}^{(2)}) = 0$, $z(i) = z(x^{(1)}) = \infty$, $z(y^{(2)}) = 1$, $z(y^{(1)}) = 2$, and $z(o^{(1)}) = 3$

a gate, gets assigned $d(e)$ equal to the maximum over all $d(e')$, where e' is a fixed transition at time $t' \leq t$ in E_σ with σ being a predecessor of v. Algorithm step (iii) is extended such that for each transition e in a set E_v, with v being a gate or an input, that generates a transition e' in some $E_{(v,w)}$, $d(e') = d(e) + 1$. We observe:

Lemma 9.8. *For all $k \geq 1$, (a) the constructive algorithm never assigns a causal depth larger than k to a transition marked fixed in iteration k, and (b) at the end of iteration k the sequence of the causal depths of the transitions in E_σ is non decreasing, for all vertices σ.*

We are now in the position to prove the main result of a circuit simulated by an unrolled circuit.

Theorem 9.9. *Let C be a circuit with involution channels with output port o that solves bounded SPF. Let $C_k(o)$ be an unrolling of C, σ a vertex in C and σ' a vertex in $C_k(o)$ corresponding to σ. For all input signals i, if a transition e within E_σ is marked fixed by the execution constructing algorithm run on circuit C (respectively $C_k(o)$) with input signal i and $d(e) \leq z(\sigma')$ then e is added and marked fixed in $E_{\sigma'}$ by the algorithm run on circuit $C_k(o)$ (respectively C) with input signal i.*

Proof. We will show the statement by induction on $d(e) \geq 0$ for the case where e is a transition in C's execution. The proof for the case where e is a transition in $C_k(o)$'s execution is analogous.

Induction base: By construction of an unrolling, E_σ and $E_{\sigma'}$ have the same initial transitions if σ is a gate or channel, and the same transitions if $\sigma = \sigma'$ is the input. Since for all these transition e, $d(e) = 0 \leq z(\sigma')$, the statement holds for $d(e) = 0$.

Induction step: Assume that the lemma holds for all transitions e with $d(e) \leq k$. We show that it also holds for transitions e' with $d(e') = k+1$ if $k+1 \leq z(\sigma')$. If e' is a transition initially added by the constructive algorithm at time 0, $d(e') = 1$. From the definition of the unrolling we immediately obtain that e' is also added to $E_{\sigma'}$ if $z(\sigma') \geq 1$. Otherwise, e' must have been added within an iteration of the constructive algorithm. Assume by contradiction that e' is the first transition (in the order transitions are generated by the constructive algorithm) with causal depth $k+1$ added to a list in C but not added to the respective list in $C_k(o)$. We distinguish two cases for σ:

If $\sigma = (v, w)$ is a channel in C: Transition e' may only have been added to $E_{(v,w)}$ by the channel algorithm with input transition list E_v and a current transition e'' with $d(e'') = k$. The time of transition e' depends on the time of e'' and on the last output transition only. From the

fact that e' is the first with depth $k+1$ not added to $E_{\sigma'}$, and the induction hypothesis applied to both E_v and $E_{(v,w)}$, we deduce that transition e' is also added to $E_{\sigma'}$; a contradiction to the assumption that it is the first one not added.

Since σ cannot be the input port, because all its transitions have causal depth 0, the case of σ being a gate in C remains: However, then e' is generated due to a transition e'' on a predecessor w of σ in C. Further, $d(e'') \leq k+1$ and $z(w') \geq k+1$ must hold for vertex w' corresponding to w. Since, e' is the first transition with causal depth less or equal $k+1$ not added to both circuits' lists, e'' was added to both $E_{w'}$ and $E_{w'}$, and thus e' is added to $E_{\sigma'}$; a contradiction that it is the first one not added. This completes the induction step. □

9.1.4 The Unsolvability Result

Let the *aligned bounded SPF problem* be the SPF problem with the following modifications: We first require that if the input signal is a pulse, then the pulse must start at time 0 and be of length at most 1. We call such signals, *valid* input signals. Further, we require that if output signal o makes a transition to 1, it must do so before time $K+1$, and o must remain 1 from thereon until time $K+2$, from whereon it is 0 until time $K+3$ followed by a pulse of length 1 at time $K+3$. If the input is constant 0, we require that the output is a pulse of length 1 at time $K+3$. From every circuit that solves the (original) bounded SPF problem, we can easily build a circuit that solves the aligned SPF problem by adding capturing circuitry like Figure 8.5. In the following, we show that no circuit solves the aligned version of bounded SPF and thus, by the above reduction, the original bounded SPF problem.

Let C be a circuit that solves the aligned bounded SPF problem. Then, for all input signals, the output signal o of C always contains a transition $(K+4, 0)$, regardless of the input. Let $D_o^C(i)$ be the causal depth of this transition in circuit C when the input signal is i.

Lemma 9.10. *Let C be a circuit that solves the aligned bounded SPF problem. Then there exists an input signal i such that $D_o^C(i) > (K+5)/\delta_{\min} + 2$.*

Proof. Let $N = (K+5)/\delta_{\min} + 2$, and assume by contradiction that $D_o^C(i) \leq N$ for all valid input signals i. Consider the N-unrolled circuit $C_N(o)$. From Theorem 9.9, we obtain that transition $(K+4, 0)$, with causal depth in C at most N occurs at output o of C if and only if it occurs at $C_N(o)$'s output o' corresponding to o. From Lemma 9.8 (b), we obtain that the same holds for all transitions at output o with times less than $K+4$; i.e., C's and $C_N(o)$'s output signals restricted to time $[-\infty, K+4]$ are the same for all valid input signals i. One can easily extend the forward circuit $C_N(o)$ such that it remains a forward circuit and solves aligned bounded SPF, by suppressing all transitions at the output that occur after time $K+4$. Since Theorem 9.6 also holds for the aligned bounded SPF problem, no such forward circuit exists; a contradiction to the initial assumption. □

From Lemma 9.10 and 9.8, we obtain that, for input i, the constructive algorithm does not mark fixed the output transition $(K+4, 0)$ before iteration $(K+5)/\delta_{\min} + 2$. However, from Lemma 7.9 and the fact that input signal i contains at most 2 transitions besides the initial transition at $-\infty$, we conclude that all iterations $\ell \geq (K+5)/\delta_{\min} + 2$ have $t_\ell \geq K+5$. From Lemma 7.8, we conclude that all transitions still existent at the end of these iterations must have times at least $K+5$; a contradiction to the fact that the output transition occurs at time $K+4$. We thus obtain:

Theorem 9.11. *No circuit with involution channels solves bounded SPF.*

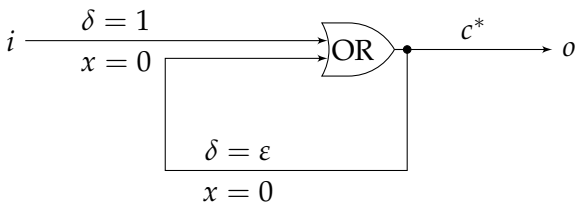


Figure 9.2: Circuit C_{ff}

9.2 Solvability of Bounded Short Pulse Filtration with One Non-Constant Delay Bounded Single History Channel

In this section we prove that bounded SPF is solvable as soon as there is a single non-constant delay bounded single-history channel available. More specifically, we show that, given a bounded single-history channel with non-constant delay, there exists a circuit that uses only constant delay channels apart from the given non-constant channel that solves bounded SPF. Different circuits, and hence proofs, are used in for different types of channels.

The right-sided limit of δ at $-\delta_\infty$ is denoted by $\delta_{\text{inf}} = \lim_{t \to 0^+} \delta(-\delta_\infty + t)$; note that $\delta_{\text{inf}} = -\infty$ is allowed here.

In the rest of this section, let c^* be a bounded single-history channel that is not a constant delay channel as defined in Section 7.3. This is equivalent to saying that its delay function δ is non-constant for $T > -\delta_\infty$, because $T_n > -\delta_\infty$ in every step of the channel algorithm:

Lemma 9.12. *A bounded single-history channel with delay function δ is a constant delay channel if and only if δ is constant in the open interval $(-\delta_\infty, \infty)$.*

Note that $\delta_{\text{inf}} < \delta_\infty$ in case of a non-constant delay channel. From the fact that $-\delta_\infty < T_n \leq \infty$ in every step of the channel algorithm, we also obtain:

Lemma 9.13. *All events in the event list of a bounded single-history channel's input signal are delayed by times within $[\delta_{\text{inf}}, \delta_\infty]$.*

9.2.1 Forgetful Channels

In this subsection, assume that c^* is forgetful. Consider circuit C_{ff} depicted in Figure 9.2, which contains channel c^* as well as two constant delay channels. For the moment assume that the initial value of c^* is 0. We will show at the end of this subsection that bounded SPF is also solvable with c^* if its initial value is 1.

It remains to describe how to choose delay parameter $\varepsilon > 0$. We will show in the following that for each non-constant delay forgetful bounded single-history channel c there exists a $\gamma(c) > 0$ such that $c(s)$ is the zero signal whenever s is a pulse of length less than $\gamma(c)$. More generally we will show that, if signal s does not contain pulses of length greater or equal to $\gamma(c)$, then $c(s)$ is the zero signal. We then choose $0 < \varepsilon < \gamma(c^*)$ for the delay parameter ε in circuit C_{ff}.

If the input signal of circuit C_{ff} is a pulse of length at least ε, then the signal s_{OR} at the OR gate is eventually stable 1 because of the ε-delay feedback loop, and hence the circuit's output signal is eventually stable 1. If the circuit's input signal is a pulse of length $\Delta < \varepsilon$, then s_{OR}

only contains pulses of length $\Delta < \gamma(c^*)$, from which it follows that the circuit's output signal is zero.

Let δ be the delay function of a bounded single-history channel c. We define:

$$\gamma(c) = \inf\{\Delta > 0 \mid \Delta - \delta_\infty + \delta(\Delta - \delta_\infty) > 0\} \tag{9.5}$$

We will prove $\gamma(c^*) > 0$ in Lemma 9.15. Before characterizing the non-constant delay channels as those c with $\gamma(c) > 0$, we need a preliminary lemma on pulse-filtration properties of non-constant delay channels.

Lemma 9.14. *Let c be a non-constant delay bounded single-history channel with initial value 0. If s is a pulse of length less than $\gamma(c)$, then $c(s)$ is zero.*

Proof. The event list of s consists of two events $(S, 1)$ and $(T, 0)$, possibly preceded by an additional event $(0, 0)$, depending on whether $S = 0$ or $S > 0$. Because the initial value of c is 0, we may assume without loss of generality that the sequence consists of only these two events.

After iteration $n = 0$ of the channel algorithm, the output list is equal to $\big((-\infty, 0), (S + \delta_\infty, 1)\big)$. Hence, in iteration $n = 1$,

$$T_1 = T - S - \delta_\infty < \gamma(c) - \delta_\infty \ , \tag{9.6}$$

i.e., $T_1 + \delta_\infty < \gamma(c)$. By definition of $\gamma(c)$, this implies

$$(T_1 + \delta_\infty) - \delta_\infty + \delta((T_1 + \delta_\infty) - \delta_\infty) \leq 0 \ , \tag{9.7}$$

and thus $T_1 + \delta(T_1) \leq 0$. Thus, the event $(S + \delta_\infty, 1)$ gets removed from the output list and the output signal is the constant zero signal. □

Lemma 9.15. *Let c be a bounded single-history channel with initial value 0. The following statements are equivalent:*

1. *c is not a constant delay channel.*
2. *There exist a pulse s such that $c(s)$ is the zero signal.*
3. *$\gamma(c) > 0$*

Proof. Let δ be the delay function of c. If s is a pulse of length Δ, then $c(s)$ is zero if and only if

$$\Delta - \delta_\infty + \delta(\Delta - \delta_\infty) \leq 0 \ . \tag{9.8}$$

This implies $\gamma(c) \geq \Delta$ and hence establishes the equivalence of (2) and (3). If we can show that c is not a constant delay channel if and only if

$$\exists \varepsilon > 0 \ : \ \delta(-\delta_\infty + \varepsilon) \leq \delta_\infty - \varepsilon \ , \tag{9.9}$$

then we can choose $\Delta = \varepsilon$, concluding the proof.

The sufficiency of (9.9) for c not being a constant delay channel is immediate. To prove the necessity of (9.9), assume that c is not a constant delay channel. Then there exist $\beta, \beta' > 0$

such that $\delta(\beta - \delta_\infty) < \delta(\beta' - \delta_\infty)$ and since δ is nondecreasing, $\delta(\beta - \delta_\infty) < \delta_\infty$. Thus, there exists a $z > 0$, such that,

$$\delta(\beta - \delta_\infty) \leq \delta_\infty - z \ . \tag{9.10}$$

There are two cases for z: If $\beta \leq z$, we obtain from (9.10) that $\delta(\beta - \delta_\infty) \leq \delta_\infty - \beta$. Choosing $\varepsilon = \beta$ shows that (9.9) holds. Otherwise, i.e., if $\beta > z$, we obtain from (9.10) and the fact that δ is nondecreasing

$$\delta(z - \delta_\infty) \leq \delta(\beta - \delta_\infty) \leq \delta_\infty - z \ . \tag{9.11}$$

Choosing $\varepsilon = z$ shows that (9.9) holds. □

Note that, while Lemmas 9.14 and 9.15 hold for both forgetful and non-forgetful channels, the following lemma does not hold for arbitrary non-forgetful channels.

Lemma 9.16. *Let c be a non-constant delay forgetful bounded single-history channel with initial value 0. Let s be a signal that does not contain pulses of length greater or equal to $\gamma(c)$ and that is not eventually equal to 1. Then $c(s)$ is the zero signal.*

Proof. The lemma is proved by inductively repeating the proof of Lemma 9.14 for all pulses contained in s. □

Lemma 9.17. *Circuit C_{ff} solves bounded SPF.*

Proof. We first note that, given an input signal, there is a unique execution for circuit C_{ff} according to Lemma 8.2, because the sole non-constant channel c^* is not part of a feedback loop.

The well-formedness property (F1) of SPF is hence fulfilled. The no generation property (F2) is also obvious.

If the input signal is a pulse of length at least ε, then $s_{\text{OR}}(t) = 1$ for all $t \geq S + 1$, and hence $s_o(t) = 1$ for all $t \geq S + 1 + \delta^*(\infty)$. In particular, this shows the nontriviality property (F3).

If the input signal is a pulse of length less than ε, then $s_{\text{OR}}(t)$ only contains pulses of lengths less than ε, hence less than $\gamma(c^*)$ by the choice of ε. By Lemma 9.16, the output signal is zero in this case. This, together with the above, shows (F4) and (F5). □

It remains to show that assuming c^* to have initial value 0 is is not restricting: If its initial value is 1 we modify circuit C_{ff} by adding an inverter before and after channel c^*. A proof analogous to Lemma 9.17's yields:

Theorem 9.18. *Let c^* be a non-constant delay forgetful bounded single-history channel. Then there exists a circuit solving bounded SPF whose channels are either constant delay channels or c^*.*

9.2.2 Non-Forgetful Channels

Theorem 9.19 reveals that a single non-constant delay *non-forgetful* bounded single-history channel c^* (with initial value 0) also allows to solve bounded SPF:

Theorem 9.19. *Let c^* be a non-constant delay non-forgetful single history channel with initial value 0. Then there exists a circuit solving SPF whose channels are all either constant delay channels or c^*.*

Let δ be the delay function of c^*. Recall from Lemma 9.12 that $\delta_{\inf} < \delta_\infty$, since δ is non-decreasing and not constant. We distinguish three cases for function δ with respect to its behavior at $-\delta_{\inf}$.

1. There exists a $t > -\delta_{\inf}$ such that $\delta(t) < \delta_\infty$.
2. $\delta(t) = \delta_\infty$ for all $t > -\delta_{\inf}$, and
   - 2.1 δ is continuous at $-\delta_{\inf}$, i.e., at $-\delta_{\inf}$ its left limit $\lim_{t\to 0^-} \delta(-\delta_{\inf} + t)$ equals its right limit δ_∞.
   - 2.2 δ is non-continuous at $-\delta_{\inf}$, i.e., $\delta^- = \lim_{t\to 0^-} \delta(-\delta_{\inf} + t) < \delta_\infty$.

For Cases 1 and 2.1, we show that circuit C_{NF} depicted in Figure 9.5 solves bounded SPF. All its clocks $CLK_{A/C/F}$ produce a signal with period $A + B + C + D$, where parameters A to D are chosen later on in accordance with δ. Let $\tau_k = k(A + B + C + D)$ denote the beginning of the k-th *round*, for $k \geq 0$. Clock CLK_C is designed such that its output signal is 0 during $[\tau_k, \tau_k + A + B) \cup [\tau_k + A + B + C, \tau_{k+1})$ and 1 during $[\tau_k + A + B, \tau_k + A + B + C)$. Such a clock can easily be built from constant delay channels and inverters only. Clock CLK_A's output signal is 1 during $[\tau_k, \tau_k + A)$ and 0 during $[\tau_k + A, \tau_{k+1})$. The output signal of CLK_F is 0 during $[\tau_k, \tau_k + E) \cup [\tau_k + E + F, \tau_{k+1})$ and 1 during $[\tau_k + E, \tau_k + E + F)$. Again, E and F are chosen later on in accordance with δ.

Abbreviating $t_k = \tau_k + 2$, we observe that circuit C_{NF} generates a signal s_{OR} at the input of channel c^*, which is the OR of two subsignals that consist of four *phases* within time $[t_k, t_{k+1})$, $k \geq 0$ (i.e., per round): Phase A (of round k) denotes the interval of times $[t_k, t_k + A)$, phase B the interval $[t_k + A, t_k + A + B)$, phase C the interval $[t_k + A + B, t_k + A + B + C)$ and phase D the interval $[t_k + A + B + C, t_k + A + B + C + D)$. The value of s_{OR} is 1 during phase A, and 0 during phases B and D. During phase C it is either 0 or contains a pulse, depending on signal i. Analogously, we define output phase F (of round k) as the interval of times $[t_k + E, t_k + E + F)$. Note that phase E and F of round k follow phase D of round k, and overlap with phase A of round $k + 1$.

Informally, for Cases 1 and 2.1, circuit C_{NF} solves bounded SPF according to the following reasoning: Property (F1) trivially holds for circuit C_{NF}. Clearly, if the circuit's input signal is 0, then the channel's input signal s_{OR} is 0 during phase C of all rounds $k \geq 0$. Subsequently, we will prove that if this is the case, then the channel's output signal $c^*(s_{OR})$ during phase F is 0 for all rounds $k \geq 0$. Since phase F is the only phase where o could possibly produce a non-0 output due to the AND gate, both (F2) and (F4) follow. Property (F3) is implied by the fact that there exists an input signal i such that s_{OR} contains a pulse during phase C of some round $k \geq 0$. We will prove below that if this is the case, then the channel's output signal is 1 during phase F of round $k + 1$. Essentially, this follows from a reduced delay of the rising transition at the end of phase D, caused by not forgetting the (canceled) pulse in phase C. From this and the fact that all delays are bounded, (F5) follows.

Case 1. In this case, we choose

(i) $C > 0$, $D > 0$ and $0 < \Delta < \delta_\infty$ such that $\delta(C + D - \delta_{\inf}) \leq \delta_\infty - \Delta$. Such values for C, D and Δ exist, because of the assumption of Case 1.

(ii) $\varepsilon > 0$, $\varepsilon' > 0$ and $C > 0$ small enough such that $\delta_\infty - \varepsilon' \geq \delta_{\inf} + \varepsilon + C$ and $\varepsilon' < \Delta/4$.

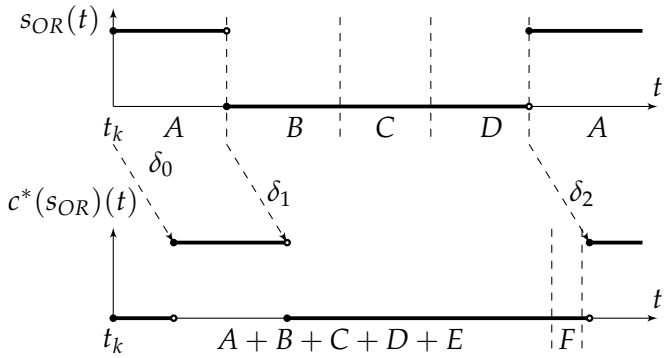


Figure 9.3: Case 1: Input and Output of channel c^* in circuit C_{NF} if phase C does not contain a pulse

(iii) $C > 0$ and $\varepsilon' > 0$ small enough such that $\delta(C + \varepsilon' - \delta_\infty) \leq \delta_{\text{inf}} + \varepsilon$.

(iv) $A = B > \max(\varepsilon', \Delta, \delta_\infty - \delta_{\text{inf}})$ and large enough such that $\delta(A - \delta_\infty) \geq \delta_\infty - \varepsilon'$.

(v) $E = \delta_\infty - \Delta$ and $F = \Delta/2$.

It is easy to check that Assumptions (i)–(v) are compatible with each other.

Figures 9.3 and 9.4 depict signal s_{OR} in absence and presence of a pulse. We will first show that the channel's output signal $c^*(s_{OR})$ has value 0 during output phase F of round 0:

The signal is depicted in Figure 9.3: Signal s_{OR}'s transition to value 1 at time t_0 is delayed by c^* by $\delta_0 = \delta_\infty > 0$. Its next transition back to value 0 at time $t_0 + A$ is delayed by, say, δ_1. Because of Lemma 9.13, $\delta_1 \geq \delta_{\text{inf}}$. From this and Assumption (iv) on A,

$$A + \delta_1 > (\delta_\infty - \delta_{\text{inf}}) + \delta_{\text{inf}} = \delta_0 \ . \tag{9.12}$$

It follows that output $c^*(s_{OR})$'s transition to 0 does not cancel $c^*(s_{OR})$'s transition to 1 from before. All of s_{OR}'s following transitions occur at times at least $t_0 + A + B$, and by (iv), at times greater than $t_0 + \delta_\infty - \delta_{\text{inf}}$. Since all these transitions are delayed by at least δ_{inf} time, none of them can cancel $c^*(s_{OR})$'s transition to 1 at time $t_0 + \delta_\infty$ either. Since channel c^* has initial value 0, it follows that its output has value 0 during $[0, t_0 + \delta_\infty)$. Since

$$t_0 + \delta_\infty > t_0 + \delta_\infty - \Delta/2 = t_0 + E + F \ , \tag{9.13}$$

the channel's output indeed has value 0 during output phase F of round 0.

We next show, for $k \geq 0$, that if signal s_{OR} does not contain a pulse within phase C of round k, signal $c^*(s_{OR})$ has value 0 during output phase F of round $k+1$:

Assume the input signal s_{OR} of channel c^* does not contain a pulse within phase C of round k. The signal is depicted in Figure 9.3.

Signal s_{OR}'s transition to value 1 at time t_k is delayed by c^* by $\delta_0 \leq \delta_\infty$.

There is no transition of s_{OR} before s_{OR}'s transition back to value 0 at time $t_k + A$. Let δ_1 be its delay. Because of (iv), and δ being non-decreasing, $A + \delta_1 > (\delta_\infty - \delta_{\text{inf}}) + \delta_{\text{inf}}$. Thus, and because transitions are delayed by at least δ_{inf}, none of the transitions from time $t_k + A$ on may cancel $c^*(s_{OR})$'s transition to 1 at time $t_k + \delta_0$.

The transition of s_{OR} to value 1 at time $t_{k+1} = t_k + A + B + C + D$ is delayed by δ_2, where

$$\delta_2 = \delta(B + C + D - \delta_1) \geq \delta(B - \delta_\infty) \geq \delta_\infty - \varepsilon' \ , \tag{9.14}$$

because of Assumption (iv). Together with (ii) this yields

$$\delta_2 > \delta_\infty - \Delta/4 \ . \tag{9.15}$$

It will thus not occur at output $c^*(s_{OR})$ before time $t_{k+1} + \delta_\infty - \Delta/4$, and thus, by (v), not before the end of output phase F of round $k+1$ at time $t_{k+1} + \delta_\infty - \Delta/2$.

Furthermore, from (9.14) and (iv),

$$B + C + D + \delta_2 > \delta_\infty \geq \delta_1 \ , \tag{9.16}$$

because (iv) in particular implies $B > \varepsilon'$. It follows that output $c^*(s_{OR})$'s transition to 1 does not cancel $c^*(s_{OR})$'s transition to 0 at time $t_k + A + \delta_1$. All s_{OR}'s subsequent transitions occur at earliest at time $t_{k+1} + A > t_{k+1} + \delta_\infty - \delta_{\inf}$, by (iv) and the fact that they are delayed by at least $\delta_{\inf}$, hence cannot cancel $c^*(s_{OR})$'s transition to 1 at time $t_{k+1} + \delta_2$. Thus, $c^*(s_{OR})$ has value 0 during $[t_k + A + \delta_1, t_{k+1} + \delta_2)$. Together with (9.15), this implies that $c^*(s_{OR})$'s value is 0 during phase F of round $k+1$.

We now show, for $k \geq 0$, that if signal s_{OR} contains a pulse within phase C of round k, signal $c^*(s_{OR})$ has value 1 during output phase F of round $k+1$:

Assume the input signal s_{OR} of channel c^* contains a pulse within phase C of round k. The signal is depicted in Figure 9.4.

Signal s_{OR}'s transition to value 1 at time t_k is delayed by $\delta_0 \leq \delta_\infty$. By the same arguments as in the proof before, it is not canceled by any following transition.

Signal s_{OR}'s transition to 0 at time $t_k + A$ is delayed by δ_1. Since no further transition of s_{OR} occurs before time $t_k + A + B$, and since $B > \delta_\infty - \delta_{\inf}$, it follows that s_{OR}'s transition to 0 is not canceled by any following transition. The transition of s_{OR} to 1 at time $t_k + A + u$ is delayed by δ_2, where $\delta_2 = \delta(u - \delta_1) \geq \delta(B - \delta_\infty)$, since $u \geq B$, $\delta_1 \leq \delta_\infty$ and δ is non-decreasing. Thus, by (iv),

$$\delta_2 \geq \delta_\infty - \varepsilon' \ . \tag{9.17}$$

The transition of s_{OR} back to value 0 at time $t_k + A + u + x$ is delayed by δ_3, where

$$\delta_3 = \delta(x - \delta_2) \leq \delta(C + \varepsilon' - \delta_\infty) \ , \tag{9.18}$$

since $x \leq C$, δ is non-decreasing, and by (9.17). By (iii),

$$\delta_3 \leq \delta_{\inf} + \varepsilon \ . \tag{9.19}$$

The pulse occurring during phase C is filtered out at the output $c^*(s_{OR})$ of channel c^*, since $\delta_2 \geq x + \delta_3$: The latter follows from (9.17), (ii) and (9.19), as $\delta_2 \geq \delta_\infty - \varepsilon' \geq \delta_{\inf} + \varepsilon + C \geq \delta_3$.

The transition of s_{OR} to value 1 at time $t_{k+1} = t_k + A + u + x + y$ is delayed by δ_4, where $\delta_4 = \delta(y - \delta_3) \leq \delta(C + D - \delta_{\inf})$, since δ is non-decreasing and $y \leq C + D$, $\delta(t) \geq \delta_{\inf}$ for all $t > -\delta_\infty$ such that $\delta_3 = \delta(x - \delta_2) \geq \delta_{\inf}$. By Assumption (i), we may thus deduce $\delta_4 \leq \delta_\infty - \Delta$. Since no further transition of s_{OR} occurs before time $t_{k+1} + A$, and $A > \delta_\infty - \delta_{\inf}$ by Assumption (iv), $c^*(s_{OR})$'s transition at time $t_{k+1} + \delta_4$ is not canceled by any later transition. Since $A > \delta_\infty - \delta_{\inf} > E + F - \delta_{\inf}$, by Assumptions (iv) and (v), and the fact that a transition is delayed by at least time $\delta_{\inf}$, no other transition of $c^*(s_{OR})$ occurs during $(t_{k+1} + \delta_4, t_{k+1} + E + F]$. It follows that $c^*(s_{OR})$'s value is 1 during phase F of round $k+1$.

Case 2.1. In this case, we choose

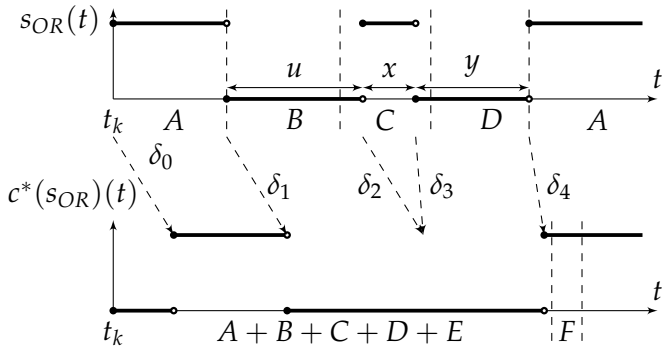


Figure 9.4: Case 1: Input and Output of channel c^* in circuit C_{NF} if phase C contains a pulse

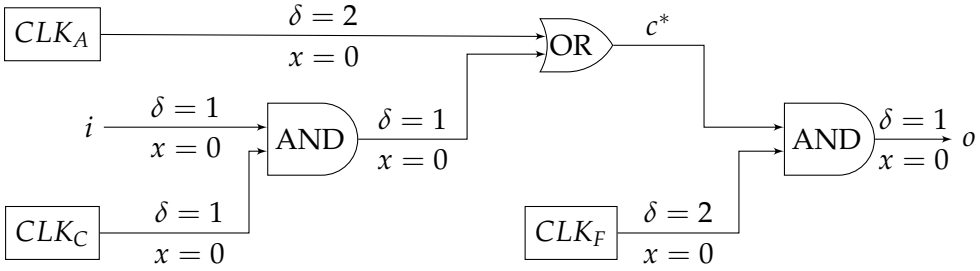


Figure 9.5: Circuit C_{NF} used in Cases 1 and 2.1

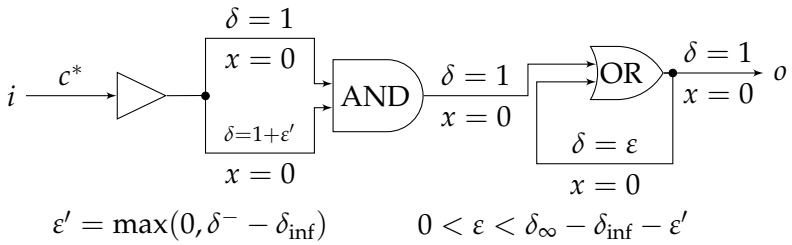


Figure 9.6: Circuit C_{NC} used in Case 2.2

(i) $A = D > \max(0, \delta_\infty - \delta_{\text{inf}})$ and large enough such that $\delta(A - \delta_\infty) = \delta_\infty$. Such an A must exist, because of the assumption of Case 2.1.

(ii) $B, C, \varepsilon > 0$ small enough such that $B + C + \varepsilon + \delta_{\text{inf}} \leq \delta_\infty$.

(iii) $0 < \varepsilon' < B + C$

(iv) $\varepsilon > 0$ small enough such that $\delta(-\delta_{\text{inf}} - \varepsilon) \geq \delta_\infty - \varepsilon'$. Such a value exists, since δ is continuous at $-\delta_{\text{inf}}$ by the assumption of Case 2.1.

(v) $B + C > 0$ small enough such that $\delta(B + C - \delta_\infty) \leq \delta_{\text{inf}} + \varepsilon$.

(vi) $E = A + \delta_\infty$ and $F = B + C - \varepsilon'$.

Again, it is easy to verify that Assumptions (i)-(vi) are compatible with each other.

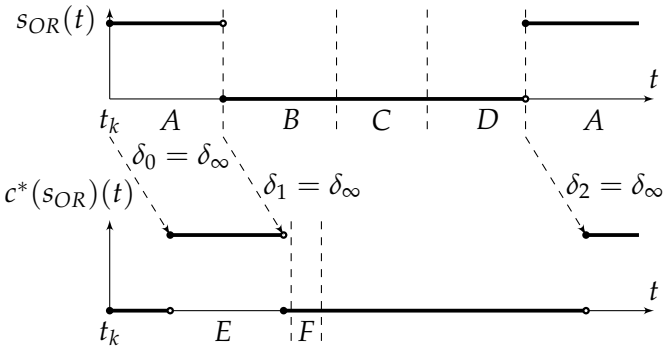


Figure 9.7: Case 2.1: Input and Output of channel c^* in circuit C_{NF} if phase C does not contain a pulse

Figures 9.7 and 9.8 depict signal s_{OR} in absence and presence of a pulse.

We next show by induction on $k \geq 0$ that signal s_{OR}'s transition at time t_k is delayed by δ_∞, and that the channel's output $c^*(s_{OR})$ has value 0 during phase F of round k in the absence of a pulse within phase C of round k, and value 1 in the presence of a pulse.

Assume the input signal s_{OR} of channel c^* contains no pulse within phase C of round k. The signal is depicted in Figure 9.7.

Signal s_{OR}'s transition to value 1 at time t_k is delayed by some δ_0. Clearly, if $k = 0$ (i.e., in round 0), $\delta_0 = \delta_\infty$. As induction hypothesis assume in the following that signal s_{OR}'s transition at time t_k is delayed by δ_∞. We will show that this implies that signal s_{OR}'s transition at time t_{k+1} is delayed by δ_∞.

Obviously, the next transition of s_{OR} back to value 0 at time $t_k + A$ is delayed by δ_1, where

$$\delta_1 = \delta(A - \delta_0) = \delta(A - \delta_\infty) = \delta_\infty \ , \tag{9.20}$$

by the choice of A according to Assumption (i). Further, by Assumption (i), $A > \delta_\infty - \delta_{\text{inf}}$, implying that no transition of s_{OR} after time t_k can cancel the transition of $c^*(s_{OR})$ to 1 at time $t_k + \delta_0$.

The transition of s_{OR} to value 1 at time $t_{k+1} = t_k + A + B + C + D$ is delayed by δ_2, where

$$\delta_2 = \delta(B + C + D - \delta_1) = \delta(B + C + D - \delta_\infty) = \delta_\infty \ , \tag{9.21}$$

because of Assumption (i). Thus, the initial transition of round $k + 1$ at time t_{k+1} will be delayed by δ_∞, which completes the inductive step. Since $D > \delta_\infty - \delta_{\text{inf}} > 0$, by Assumption (i), it follows that $c^*(s_{OR})$'s transition to 0 at time $t_k + A + \delta_1$ is not canceled by any transition. By analogous arguments, the transition to 1 at time $t_{k+1} + \delta_2$ is not canceled by any transition. Our choice of E and F in (vi) thus implies that the channel output's value is 0 during phase F of round k, see Figure 9.7.

Now assume that there is a pulse within phase C of round k. The channel's input and output signals are depicted in Figure 9.8.

Signal s_{OR}'s initial transition to value 1 at time t_k clearly is delayed by $\delta_0 = \delta_\infty$ if $k = 0$. As induction hypothesis assume in the following that s_{OR}'s transition at time t_k is delayed by δ_∞. We will show that this implies that s_{OR}'s transition at time t_{k+1} is delayed by δ_∞.

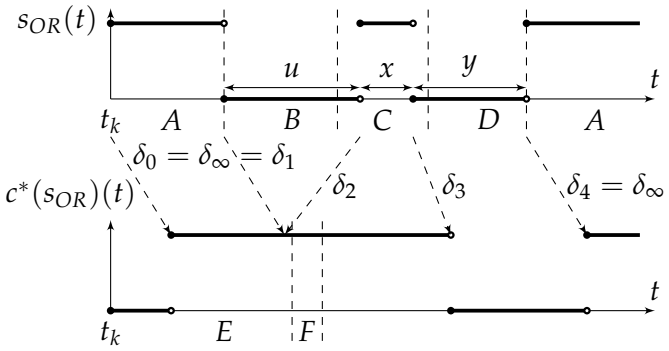


Figure 9.8: Case 2.1: Input and Output of channel c^* in circuit C_{NF} if phase C contains a pulse

By the same reasoning as in the proof before, $c^*(s_{OR})$'s transition to 1 at time $t_k + \delta_0$ is not canceled by any following transition. Further, s_{OR}'s transition back to value 0 at time $t_k + A$ is delayed by $\delta_1 = \delta_\infty$.

The transition of s_{OR} to value 1 at time $t_k + A + u$ is delayed by δ_2, where $\delta_2 = \delta(u - \delta_1) \leq \delta(B + C - \delta_\infty) \leq \delta_{\inf} + \varepsilon$, by Assumption (v). From (ii), we further obtain $u + \delta_2 \leq B + C + \delta_{\inf} + \varepsilon \leq \delta_\infty$. It follows that this output transition cancels the last output transition to 0.

The transition of s_{OR} back to value 0 at time $t_k + A + u + x$ is delayed by δ_3, where $\delta_3 = \delta(x - \delta_2) \geq \delta(-\delta_{\inf} - \varepsilon) \geq \delta_\infty - \varepsilon'$, holds because of Assumption (iv).

The transition of s_{OR} to value 1 at time t_{k+1} is delayed by δ_4, where $\delta_4 = \delta(y - \delta_3) \geq \delta(D - \delta_\infty) = \delta_\infty$, by Assumption (i), which completes the inductive step.

Moreover, since $D > \delta_\infty - \delta_{\inf} > 0$, it follows that $c^*(s_{OR})$'s transition to 0 at time $t_k + A + u + x + \delta_3$ is not canceled by any transition. By similar arguments, $c^*(s_{OR})$'s transition to 1 at time $t_{k+1} + \delta_4$ is not canceled by any following transition.

Assumption (vi) hence implies that $c^*(s_{OR})$'s value is 1 during phase F of round $k + 1$, see Figure 9.8.

Case 2.2. For this case, circuit C_{NC} depicted in Figure 9.6 solves bounded SPF. The algorithm and its proof rest on the following idea: We first show in Lemma 9.20 that every channel c^* whose δ is in accordance with Case 2.2 does not produce pulses of length within the non-zero interval $[\max(0, \delta^- - \delta_{\inf}), \delta_\infty - \delta_{\inf})$. The remaining part of circuit C_{NC} thus just has to filter out all pulses with duration less than $\max(0, \delta^- - \delta_{\inf})$ (ensured by the AND gate) and continuously hold all pulses of length $\delta_\infty - \delta_{\inf}$ (done by the OR gate).

Lemma 9.20. *Let c^* be a non-constant delay non-forgetful channel chosen in accordance to Case 2.2. If the channel's input signal is a pulse, then its output signal is either* 0 *or a pulse whose length is not within the non-zero interval* $[\max(0, \delta^- - \delta_{\inf}), \delta_\infty - \delta_{\inf}]$.

Proof. Assume that $\delta(-\delta_{\inf}) = \delta_\infty$; the proof for the case $\delta(-\delta_{\inf}) = \delta^- < \delta_\infty$ is almost the same. Without loss of generality, assume that the input pulse starts at time 0 and let $x > 0$ be its length. Clearly, the transition of the output signal to 1 is scheduled at time δ_∞, the transition back to 0 is scheduled at time $x + \delta(x - \delta_\infty)$. We distinguish two cases for the input pulse length x:

In case $x < \delta_\infty - \delta_{\inf}$, we have $\delta(x - \delta_\infty) \leq \delta^-$ and the following two sub-cases: If additionally $x \leq \delta_\infty - \delta^-$, then $x + \delta(x - \delta_\infty) \leq x + \delta^- \leq \delta_\infty$, so the output events cancel. If $\delta_\infty - \delta_{\inf} > x > \delta_\infty - \delta^-$, the length of the output pulse is $x + \delta(x - \delta_\infty) - \delta_\infty < \delta^- - \delta_{\inf}$.

This confirms the lower boundary of the "forbidden pulse length interval" given in our lemma. In case of $x \geq \delta_\infty - \delta_{\text{inf}}$, on the other hand, $\delta(x - \delta_\infty) = \delta_\infty$ a pulse with length $x + \delta(x - \delta_\infty) - \delta_\infty \geq \delta_\infty - \delta_{\text{inf}}$ is generated at the output of c^*, which also confirms the upper boundary of the interval. □

If we choose the circuit parameters in Figure 9.6 according to $\varepsilon' = \max(0, \delta^- - \delta_{\text{inf}})$ and $0 < \varepsilon < \delta_\infty - \delta_{\text{inf}} - \varepsilon'$, it is not difficult to show that the resulting circuit C_{NC} solves bounded SPF in Case 2.2: Properties (F1) and (F2) trivially hold for circuit C_{NC}. To prove (F3), consider that if the input signal i is a pulse of length $2\delta_\infty$, the output signal $s_{c^*(i)}$ of c^* is a pulse of length at least δ_∞. Thus, the output of the AND gate s_{AND} is a pulse of length at least $\delta_\infty - \varepsilon' > \varepsilon$, resulting in the circuit's output o making a transition to 1 and remaining 1 from there on.

Property (F4) directly follows from Lemma 9.20: If $s_{c^*(i)}$ is a pulse whose length is smaller than $\max(0, \delta^- - \delta_{\text{inf}}) = \varepsilon'$, then it is completely filtered out; s_{AND} and hence o are hence permanently 0. Otherwise, by Lemma 9.20, $s_{c^*(i)}$ must be a pulse of length at least $\delta_\infty - \delta_{\text{inf}}$. Thus, s_{AND} is a pulse of length at least $\delta_\infty - \delta_{\text{inf}} - \varepsilon' > \varepsilon$, which is sufficiently long to be permanently captured in the storage looped formed by the OR gate. The circuit's output o hence makes a transition to 1 and remains 1 from there on.

Finally, (F5) is due to bounded channel delays.

Chapter 10

Conclusion

This thesis dealt with prediction of transient behavior in certain distributed systems and digital circuits. We contributed to the state of the art in different ways: For max-plus linear systems, we gave analytic upper bounds on the finite transient from which on the system is periodic. For the linear algorithm for asymptotic consensus, we gave upper bounds and worst-case lower bound examples on the rate of convergence to the common limit value. For glitch propagation in digital circuits, we showed that all previously existing binary models were insufficient for correctly predicting the occurrence of short pulses, and we defined a new binary model that does not suffer from the same deficiency. Our results hence came in two different styles: The results for linear distributed systems were theoretical, analytical, upper bounds on performance parameters, whereas the results for glitch propagation reasoned about automated circuit simulators as implemented, for example, in VHDL or Verilog tool chains to circumvent resource intensive Spice simulator runs. We discuss the contribution in each of these three domains in more detail, and give an outlook on future research directions they enable, in the following paragraphs.

The first bulk of results were on transients of max-plus systems and matrices. We showed their explicit applicability in transportation systems, synchronized networks, link reversal routing and scheduling, and cyclic scheduling. From the viewpoint of applications, the study of system transients is of much higher immediate importance than the study of matrix transients. We chose to give results, and develop them in parallel, for two reasons: Firstly, the transient of the system matrix is an upper bound on the system transients independent of the initial vector. In fact, it is equal to the maximum system transient when varying the initial vector. Matrix transients are hence, in particular, system transients. Secondly, matrix transients are of interest in their own right as when interpreting max-plus matrices as edge-weighted digraphs. Under this interpretation, matrix transients are generalizations of a well-studied object in non-weighted digraphs, namely the index of convergence, also called the exponent.

Due to the lower bound example by Hartmann and Arguelles, it is clear that there can be no transience bounds only in terms of the dimension of the matrix. Indeed, the transient of their example is unbounded in the second largest cycle mean. The second largest cycle mean, or more precisely the difference between the largest and the second largest, is hence the fundamental parameter controlling the global transients of max-plus matrices and systems. Our study refined this observation in two different directions: We identified a set of graph parameters that also influence the transient. Each of those allow a system designer to construct a system with a small transient in graph-theoretic terms. Moreover, we identified three alterna-

tive cycle means, which we referred to as "schemes". For all of these, there exist—and one can construct—examples for which the scheme is, in fact, equal to the second largest cycle mean. However, we exhibited examples of classes of matrices for which the three schemes were all different from the second largest cycle mean, and different from each other. We showed with our transience bounds using the schemes that each of the three schemes can take the role of the fundamental parameter for transients, just as the second largest cycle mean. In this sense, all three schemes yield lower transience bounds than taking the second largest cycle mean. We showed a strict order between the different schemes, and identified the Cycle Threshold scheme as the smallest one. In particular, it is equal to $-\infty$ the most often, in which case our transience bounds become independent of the specific edge weights. However, the graph-theoretic constructions needed to utilize the Cycle Threshold scheme are more complicated than for the other schemes.

We showed that the second largest cycle mean is not the most significant parameter for the transients of critical nodes. We showed transience bounds for critical nodes that are independent of the specific weights, but only depend on the matrix's digraph and its set of critical cycles (i.e., its critical digraph). Compared with the lower bound for global matrix transients, there can hence be an arbitrarily large gap between the smallest and the largest transient of entries of a max-plus matrix; even if the matrix's dimension is fixed. Moreover, the results are direct generalizations of the known transience bounds in the Boolean case, i.e., the bounds on the index of convergence, or exponent, of digraphs to weighted digraphs. We therefore extended the theorems of Wielandt, Dulmage and Mendelsohn, Denardo, Schwarz, Kim, and Gregory, Kirkland, and Pullman, from non-weighted to weighted digraphs. We developed the remainder of our bounds on transients of max-plus matrices and systems. They bounded the global transient, i.e., the maximum transient among all entries. We were able to strictly improve *all* max-plus transience bounds known to date. Chapter 4 also contained a qualitative and quantitative comparison between matrix and system transients, and the necessary steps to extend Nachtigall's matrix decomposition to a fully-blown transience bound, which gives an alternative proof technique.

Our results on max-plus matrices and systems hence allow for a fine-grained analysis and prediction of a large class of systems with some form of synchronization primitive which render them max-plus linear systems. In one of our discussed examples, however, the parameter of interest was not the the transient, but rather a related parameter; namely the recovery time of train networks. One direction of future work can hence focus on directly bounding parameters related to the transient in various application examples. Another route to extending the work on max-plus systems presented in this thesis is to integrate the parameters into a design toolbox. As we have mentioned, the transient is computable in polynomial time. It is hence possible to re-compute it after every change to the system design to see whether the transient changes. However, visualising the parameters we identified here and presenting them in aggregated form to the system designer can help to guide their decisions and the direction in which they explore the design space.

The next set of results was on asymptotic consensus systems, which model natural phenomena like bird flocking, firefly synchronization, or opinion dynamics, as well as engineered systems like sensor fusion networks, robot control formation, or dynamic load balancing protocols. We gave new results on the rate of convergence of such systems and also some new sufficient conditions for convergence. The contribution here was twofold: On one hand, we refined existing proofs for convergence to extract explicit upper bounds on the rate of convergence. On the other hand, we extended also existing convergence conditions by removing

the assumption of necessary self-confidence of every agent at every time instance. We relaxed this assumption of self-confidence to that of an arbitrary perpetually existing aperiodic sub-digraph of positive confidence, for which the self-confidence assumption is a particular instance. Another instance are non-synchronous systems, for which we were able to prove new convergence results and convergence rate bounds. On the technical side, the first batch of results on asymptotic consensus used spectral techniques to show convergence rate bounds for systems with constant communication, as well as systems with constant Perron vectors. The second batch of results used graph-theoretic arguments, bearing some similarities to arguments for max-plus systems, to show our results for completely dynamic settings. Future work can focus on identifying applications in which self-confidence is not necessarily a given.

The second part of the thesis dealt with glitch propagation in digital circuits. More specifically, we constructed a fast binary-valued model that has the potential to faithfully capture glitch propagation phenomena. We also showed that all other binary valued models existing to date fail to do so. While the behavior of digital circuits can be simulated very accurately with numerical toolboxes such as Spice, the resource intensity of these simulations can be prohibitive; this is true, in particular, when the results of the simulations are needed over and over in an iterative design process. Predicting the occurrence and timing of glitches is of utmost importance since input glitches can drive a circuit into a metastable state, which can render it unusable for a significant amount of time.

Technically, we defined the Short-Pulse Filtration problem, which is also closely related to arbitration and synchronization. We showed that physical circuits cannot solve bounded SPF, while there exist physical circuits that solve unbounded SPF. We then identified a common generalization of all existing binary-valued circuit models, which we called bounded single-history channels. After that, we showed that the classical constant-delay channels cannot model circuits solving unbounded SPF, which shows that constant-delay channels cannot capture the SPF problem correctly since unbounded SPF is indeed solvable in physical circuits. Similarly, we showed that non constant-delay bounded single-history channels can always model circuits that solve bounded SPF, which again shows that these channel models do not capture the SPF problem. On the other hand, we identified the boundedness of all existing channels as the reason for their inability to faithfully model glitch propagation. We therefore defined a binary-valued model based on involution channels, which do have the single-history property, but are not bounded from below. And indeed, we showed that, with this class of channel models, unbounded SPF is solvable while bounded SPF is not, in accordance with physical reality.

While we identified the involution channel model as the only existing binary-valued circuit model that is theoretically able to correctly predict the occurrence of glitches, it is not clear that simulations based on this model are actually sufficiently accurate. In principle, even though we showed that the involution channel model is the only candidate that passes the theoretical test of correctly modeling glitch propagation, it is possible that previous models give more accurate simulation results for some classes of circuits and executions. Future work should hence focus on implementing the involution channel model in a simulation toolbox and compare its results with those of other binary-valued models and Spice simulations or even measurements on real chips.

Bibliography

[1] Marianne Akian, Stéphane Gaubert, and Cormac Walsh. Discrete max-plus spectral theory. In G.L. Litvinov and V.P. Maslov, editors, *Idempotent Mathematics and Mathematical Physics*, pages 53–78. American Mathematical Society, Providence, 2005.

[2] David Aldous. Random walks on finite groups and rapidly mixing Markov chains. In *Séminaire de probabilités de Strasbourg*, volume 17, pages 243–297. Springer, Heidelberg, 1983.

[3] David Aldous and Persi Diaconis. Shuffling cards and stopping times. *American Mathematical Monthly*, 93(5):333–348, 1986.

[4] David Aldous and Persi Diaconis. Strong uniform times and finite random walks. *Advances in Applied Mathematics*, 8(1):69–97, 1987.

[5] Peter J. Ashenden. *The Designers Guide to VHDL*. Morgan Kaufmann, Burlington, third edition, 2008.

[6] Baruch Awerbuch. Complexity of network synchronization. *Journal of the ACM*, 32(4):804–823, 1985.

[7] M. Ballerini, N. Cabibbo, R. Candelier, A. Cavagna, E. Cisbani, I. Giardina, V. Lecomte, A. Orlandi, G. Parisi, A. Procaccini, M. Viale, and V. Zdravkovic. Interaction ruling animal collective behavior depends on topological rather than metric distance: Evidence from a field study. *Proceedings of the National Academy of Sciences*, 105(4):1232–1237, 2008.

[8] Valmir C. Barbosa and Eli Gafni. Concurrency in heavily loaded neighborhood-constrained systems. *ACM Transactions on Programming Languages and Systems*, 11(4):562–584, 1989.

[9] Greg Barnes and Uriel Feige. Short random walks on graphs. *SIAM Journal on Discrete Mathematics*, 9(1):19–28, 1996.

[10] José C. Barros and Brian W. Johnson. Equivalence of the arbiter, the synchronizer, the latch, and the inertial delay. *IEEE Transactions on Computers*, 32(7):603–614, 1983.

[11] M. J. Bellido-Díaz, J. Juan-Chico, A. J. Acosta, M. Valencia, and J. L. Huertas. Logical modelling of delay degradation effect in static CMOS gates. *IEE Proceedings – Circuits, Devices, and Systems*, 147(2):107–117, 2000.

[12] Manuel J. Bellido-Díaz, Jorge Juan-Chico, and Manuel Valencia. *Logic-Timing Simulation and the Degradation Delay Model*. Imperial College Press, London, 2006.

[13] Vincent D. Blondel, Julien M. Hendrickx, Alex Olshevsky, and John N. Tsitsiklis. Convergence in multiagent coordination, consensus, and flocking. In *Proceedings of the 44th IEEE Conference on Decision and Control, and the European Control Conference (CDC-ECC)*, pages 2996–3000. IEEE, New York City, 2005.

[14] Anne Bouillard and Bruno Gaujal. Coupling time of a (max,plus) matrix. In *Proceedings of the Workshop on Max-Plus Algebra at the 1st IFAC Symposium on System Structure and Control*. Elsevier, Amsterdam, 2001.

[15] Anne Bouillard and Éric Thierry. An algorithmic toolbox for network calculus. *Discrete Event Dynamic Systems*, 18(1):3–49, 2008.

[16] Michael S. Branicky. Universal computation and other capabilities of hybrid and continuous dynamical systems. *Theoretical Computer Science*, 138(1):67–100, 1995.

[17] Alfred Brauer. On a problem of partitions. *American Journal of Mathematics*, 64(1):299–312, 1942.

[18] Richard A. Brualdi and Herbert J. Ryser. *Combinatorial Matrix Theory*. Cambridge University Press, Cambridge, 1991.

[19] Janusz A. Brzozowski and Jo C. Ebergen. On the delay-sensitivity of gate networks. *IEEE Transactions on Computers*, 41(11):1349–1360, 1992.

[20] Costas Busch and Srikanta Tirthapura. Analysis of link reversal routing algorithms. *SIAM Journal on Computing*, 35(2):305–326, 2005.

[21] Ming Cao, A. Stephen Morse, and Brian D. O. Anderson. Reaching a consensus in a dynamically changing environment: A graphical approach. *SIAM Journal on Control and Optimization*, 47(2):575–600, 2008.

[22] Ming Cao, A. Stephen Morse, and Brian D. O. Anderson. Reaching a consensus in a dynamically changing environment: Convergence rates, measurement delays, and asynchronous events. *SIAM Journal on Control and Optimization*, 47(2):601–623, 2008.

[23] K. M. Chandy and J. Misra. The drinking philosophers problem. *ACM Transactions on Programming Languages and Systems*, 6(4):632–646, 1984.

[24] Bernadette Charron-Bost. Orientation and connectivity based criteria for asymptotic consensus. Preprint, arXiv:1303.2043v1 [cs.DC], 2013.

[25] Bernadette Charron-Bost, Matthias Függer, Jennifer L. Welch, and Josef Widder. Full reversal routing as a linear dynamical system. In Adrian Kosowski and Masafumi Yamashita, editors, *Proceedings of the 18th International Colloquium on Structural Information and Communication Complexity (SIROCCO)*, volume 6796 of *Lecture Notes in Computer Science*, pages 101–112. Springer, Heidelberg, 2011.

[26] Bernadette Charron-Bost, Matthias Függer, Jennifer L. Welch, and Josef Widder. Partial is full. In Adrian Kosowski and Masafumi Yamashita, editors, *Proceedings of the 18th International Colloquium on Structural Information and Communication Complexity (SIROCCO)*, volume 6796 of *Lecture Notes in Computer Science*, pages 113–124. Springer, Heidelberg, 2011.

[27] Bernadette Charron-Bost and André Schiper. The Heard-Of model: computing in distributed systems with benign faults. *Distributed Computing*, 22(1):49–71, 2009.

[28] Bernard Chazelle. The total s-energy of a multiagent system. *SIAM Journal on Control and Optimization*, 49(4):1680–1706, 2011.

[29] Fan Chung. Laplacians and the Cheeger inequality for directed graphs. *Annals of Combinatorics*, 9:1–19, 2005.

[30] G. Cohen, P. Moller, J.-P. Quadrat, and M. Viot. Algebraic tools for the performance evaluation of discrete event systems. *Proceedings of the IEEE*, 77(1):39–85, 1989.

[31] Guy Cohen, Didier Dubois, Jean-Pierre Quadrat, and Michel Viot. Analyse du comportement périodique de systèmes de production par la théorie des dioïdes. INRIA Research Report 191, INRIA, Le Chesnay, 1983.

[32] Guy Cohen, Didier Dubois, Jean Pierre Quadrat, and Michel Viot. A linear-system-theoretic view of discrete-event processes and its use for performance evaluation in manufacturing. *IEEE Transactions on Automatic Control*, 30(3):210–220, 1985.

[33] Cristian Constantinescu. Trends and challenges in VLSI circuit reliability. *IEEE Micro*, 23(4):14–19, 2003.

[34] Eric V. Denardo. Periods of connected networks and powers of nonnegative matrices. *Mathematics of Operations Research*, 2(1):20–24, 1977.

[35] Persi Diaconis and Daniel Stroock. Geometric bounds for eigenvalues of Markov chains. *Annals of Applied Probability*, 1(1):36–61, 1991.

[36] D. Dolev, M. Függer, C. Lenzen, and U. Schmid. Fault-tolerant algorithms for tick-generation in asynchronous logic: Robust pulse generation. In Xavier Défago, Franck Petit, and Vincent Villain, editors, *Proceeding of the 13th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS)*, volume 6976 of *LNCS*, pages 163–177. Springer, Heidelberg, 2011.

[37] Shlomi Dolev. *Self-Stabilization*. MIT Press, Cambridge, 2000.

[38] Michael Drmota and Robert F. Tichy. *Sequences, Discrepancies, and Applications*. Springer, Heidelberg, 1997.

[39] Didier Dubois and Kathryn E. Stecke. Dynamic analysis of repetitive decision-free discrete event processes: The algebra of timed marked graphs and algorithmic issues. *Annals of Operations Research*, 26(1):150–193, 1990.

[40] A. L. Dulmage and N. S. Mendelsohn. Gaps in the exponent set of primitive matrices. *Illinois Journal of Mathematics*, 8(4):642–656, 1964.

[41] S. Even and S. Rajsbaum. The use of a synchronizer yields the maximum computation rate in distributed networks. *Theory of Computing Systems*, 30(5):447–474, 1997.

[42] James Allen Fill. Eigenvalue bounds on convergence to stationarity for nonreversible Markov chains, with an application to the exclusion process. *Annals of Applied Probability*, 1(1):62–87, 1991.

[43] G. Frobenius. Über Matrizen aus nicht negativen Elementen. *Sitzungsberichte der Königlich Preussischen Akademie der Wissenschaften*, pages 456–477, 1912.

[44] Gottfried Fuchs, Matthias Függer, and Andreas Steininger. On the threat of metastability in an asynchronous fault- tolerant clock generation scheme. In *Proceedings of the 15th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC)*, pages 127–136. IEEE Press, New York City, 2009.

[45] Matthias Függer and Ulrich Schmid. Reconciling fault-tolerant distributed computing and systems-on-chip. *Distributed Computing*, 14(6):323–355, 2012.

[46] Matthew J. Gadledge, Paul H. Eaton, Joseph M. Benedetto, Marty Carts, Vivian Zhu, and Thomas L. Turflinger. Digital device error rate trends in advanced CMOS technologies. *IEEE Transactions on Nuclear Science*, 53(6):3466–3471, 2006.

[47] E. M. Gafni and D. P. Bertsekas. Asymptotic optimality of shortest path routing algorithms. *IEEE Transactions on Information Theory*, 33(1):83–90, 1987.

[48] Michael R. Garey and David S. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. Freeman, San Francisco, 1979.

[49] R. M. P. Goverde, P. H. L. Bovy, and G. J. Olsder. The max-plus algebra approach to transportation problems. In *Proceedings of the 8th World Conference on Transport Research (WCTR)*, pages 377–390. Pergamon, Amsterdam, 1999.

[50] D. A. Gregory, S. J. Kirkland, and N. J. Pullman. A bound on the exponent of a primitive matrix using Boolean rank. *Linear Algebra and Its Applications*, 217:101–116, 1995.

[51] J. Hajnal. Weak ergodicity in non-homogeneous Markov chains. *Proceedings of the Cambridge Philosophical Society*, 54(2):233–246, 1958.

[52] C. Hanen and A. Munier. Cyclic scheduling on parallel processors: an overview. In Philippe Chrétienne, Edward G. Coffmann, Jr., Jan Karel Lenstra, and Zhen Liu, editors, *Scheduling Theory and Its Applications*, pages 193–225. John Wiley & Sons, Chichester, 1995.

[53] Mark Hartmann and Cristina Arguelles. Transience bounds for long walks. *Mathematics of Operations Research*, 24(2):414–439, 1999.

[54] Rainer Hegselmann and Ulrich Krause. Opinion dynamics and bounded confidence models, analysis, and simulation. *Journal of Artificial Societies and Social Simulation*, 5(3), 2002.

[55] Rainer Hegselmann and Ulrich Krause. Truth and cognitive division of labour: First steps towards a computer aided social epistemology. *Journal of Artificial Societies and Social Simulation*, 9(3), 2006.

[56] Bernd Heidergott, Geert Jan Olsder, and Jacob van der Woude. *Max Plus at Work*. Princeton University Press, Princeton, 2006.

[57] Julien M. Hendrickx and Vincent D. Blondel. Convergence of linear and non-linear versions of Vicsek's model. CESAME Research Report 2005.57, Université catholique de Louvain, Louvain-la-Neuve, 2005.

[58] Ali Jadbabaie, Jie Lin, and A. Stephen Morse. Coordination of groups of mobile autonomous stability agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, 48(6):988–1001, 2003.

[59] K. H. Kim. An extension of the Dulmage-Mendelsohn theorem. *Linear Algebra and Its Applications*, 27:187–197, 1979.

[60] David J. Kinniment. *Synchronization and Arbitration in Digital Systems*. Wiley, Chichester, 2007.

[61] H.J. Landau and A.M. Odlyzko. Bounds for eigenvalues of certain stochastic matrices. *Linear Algebra and Its Applications*, 38:5–15, 1981.

[62] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, 2009.

[63] Nancy A. Lynch. *Distributed Algorithms*. Morgan Kaufmann, San Francisco, 1996.

[64] Yossi Malka, Shlomo Moran, and Shmuel Zaks. A lower bound on the period length of a distributed scheduler. *Algorithmica*, 10(5):383–398, 1993.

[65] Yossi Malka and Sergio Rajsbaum. Analysis of distributed algorithms based on recurrence relations. In Sam Toueg, Paul G. Spirakis, and Lefteris Kirousis, editors, *Proceedings of the 5th International Workshop on Distributed Algorithms (WDAG)*, volume 579 of *Lecture Notes in Computer Science*, pages 242–253. Springer, Heidelberg, 1992.

[66] Leonard R. Marino. The effect of asynchronous inputs on sequential network reliability. *IEEE Transactions on Computers*, 26(11):1082–1090, 1977.

[67] Leonard R. Marino. General theory of metastable operation. *IEEE Transactions on Computers*, 30(2):107–115, 1981.

[68] M. Salim Maza and M. Linares Aranda. Analysis of clock distribution networks in the presence of crosstalk and groundbounce. In *Proceedings of the 8th IEEE International Conference on Electronics, Circuits, and Systems (ICECS)*, pages 773–776. IEEE Press, New York City, 2001.

[69] Michael Mendler and Terry Stroup. Newtonian arbiters cannot be proven correct. *Formal Methods in System Design*, 3(3):233–257, 1993.

[70] Milena Mihail. Conductance and convergence of Markov chains: A combinatorial treatment of expanders. In *Proceedings of the 30th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 526–531. IEEE, New York City, 1989.

[71] Luc Moreau. Stability of multiagent systems with time-dependent communication links. *IEEE Transactions on Automatic Control*, 50(2):169–182, 2005.

[72] Karl Nachtigall. Powers of matrices over an extremal algebra with applications to periodic graphs. *Mathematical Methods of Operations Research*, 46:87–102, 1997.

[73] Alex Olshevsky and John N. Tsitsiklis. Convergence speed in distributed consensus and averaging. *SIAM Review*, 53(4):747–772, 2011.

[74] Rohit J. Parikh. On context-free languages. *Journal of the ACM*, 13(4):570–581, 1966.

[75] J. W. Pitman. Uniform rates of convergence for Markov chain transition probabilities. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, 29(3):193–227, 1974.

[76] Craig W. Reynolds. Flocks, herds, and schools: A distributed behavioral model. *ACM SIGGRAPH Computer Graphics*, 21(4):25–34, 1987.

[77] Fred U. Rosenberger, Charles E. Molnar, Thomas J. Chaney, and Ting-Pien Fang. Q-modules: Internally clocked delay-insensitive modules. *IEEE Transactions on Computers*, 37(9):1005–1018, 1988.

[78] Hans Schneider and Michael H. Schneider. Max-balancing weighted directed graphs and matrix scaling. *Mathematics of Operations Research*, 16(1):208–222, 1991.

[79] Johannes Schoissengeier. The discrepancy of $(n\alpha)_{n\geq 1}$. *Mathematische Annalen*, 296(1):529–545, 1993.

[80] Štefan Schwarz. On a sharp estimation in the theory of binary relations on a finite set. *Czechoslovak Mathematical Journal*, 20(4):703–714, 1970.

[81] Štefan Schwarz. On the semigroup of binary relations on a finite set. *Czechoslovak Mathematical Journal*, 20(4):632–679, 1970.

[82] E. Seneta. *Non-negative Matrices and Markov Chains*. Springer, Heidelberg, 1993.

[83] Alistair Sinclair and Mark Jerrum. Approximate counting, uniform generation, and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, 1989.

[84] Gerardo Soto y Koelemeijer. *On the Behaviour of Classes of Min-Max-Plus Systems*. PhD thesis, TU Delft, 2003.

[85] Ivan E. Sutherland. Micropipelines. *Communications of the ACM*, 32(6):720–738, 1989.

[86] H. G. Tanner, A. Jadbabaie, and G. J. Pappas. Flocking in fixed and switching networks. *IEEE Transactions on Automatic Control*, 52(5):863–868, 2007.

[87] Behrouz Touri and Angelia Nedić. Product of random stochastic matrices. Preprint, arXiv:1110.1751v2 [math.PR], 2011.

[88] John N. Tsitsiklis. *Problems in Decentralized Decision Making and Computation*. PhD thesis, Massachusetts Institute of Technology, 1984.

[89] Stephen H. Unger. Asynchronous sequential switching circuits with unrestricted input changes. *IEEE Transaction on Computers*, 20(12):1437–1444, 1971.

[90] J. L. Welch and J. E. Walter. *Link Reversal Algorithms*. Morgan & Claypool, San Rafael, 2012.

[91] Helmut Wielandt. Unzerlegbare, nicht negative Matrizen. *Mathematische Zeitschrift*, 52(1):642–645, 1950.

Author's Publications

[92] Bernadette Charron-Bost, Matthias Függer, and Thomas Nowak. New transience bounds for long walks. Preprint, submitted, arXiv:1209.3342v1 [cs.DM], 2012.

[93] Bernadette Charron-Bost, Matthias Függer, and Thomas Nowak. New transience bounds for long walks in weighted digraphs. In Jaroslav Nešetřil and Marco Pellegrini, editors, *Proceedings of the 7th European Conference on Combinatorics, Graph Theory, and Applications (EuroComb)*, volume 16 of *CRM Series*, pages 623–624. Scuola Normale Superiore, Pisa, 2013.

[94] Bernadette Charron-Bost, Matthias Függer, and Thomas Nowak. Transience bounds for distributed algorithms. In Víctor Braberman and Laurent Fribourg, editors, *Proceeding of the 11th International Conference on Formal Modeling and Analysis of Timed Systems (FORMATS)*, volume 8053 of *Lecture Notes in Computer Science*, pages 77–90. Springer, Heidelberg, 2013.

[95] Bernadette Charron-Bost and Thomas Nowak. General transience bounds in tropical linear algebra via Nachtigall decomposition. In G. L. Litvinov, V. P. Maslov, A. G. Kushner, and S. N. Sergeev, editors, *Proceedings of the Workshop on Tropical and Idempotent Mathematics*, pages 46–52. Institute for Information Transmission Problems of the Russian Academy of Sciences, Moscow, 2012.

[96] Matthias Függer, Alexander Kößler, Thomas Nowak, Ulrich Schmid, and Martin Zeiner. The effect of forgetting on the performance of a synchronizer. In Paola Flocchini, Jie Gao, Evangelos Kranakis, and Friedhelm Meyer auf der Heide, editors, *Proceedings of the 9th International Symposium on Algorithms and Experiments for Sensor Systems, Wireless Networks and Distributed Robotics (Algosensors)*, Lecture Notes in Computer Science, pages 185–200. Springer, Heidelberg, 2014.

[97] Matthias Függer, Alexander Kößler, Thomas Nowak, and Martin Zeiner. Brief announcement: The degrading effect of forgetting on a synchronizer. In Andréa W. Richa and Christian Scheideler, editors, *Proceeding of the 14th International Symposium on Stabilization, Safety, and Security of Distributed Systems (SSS)*, volume 7596 of *Lecture Notes in Computer Science*, pages 90–91. Springer, Heidelberg, 2012.

[98] Matthias Függer, Thomas Nowak, and Ulrich Schmid. Unfaithful glitch propagation in existing binary circuit models. In *Proceedings of the 19th IEEE International Symposium on Asynchronous Circuits and Systems (ASYNC)*, pages 191–199. IEEE Press, New York City, 2013. Full paper available at arXiv:1311.1423v1 [cs.OH].

[99] Glenn Merlet, Thomas Nowak, Hans Schneider, and Sergeĭ Sergeev. Generalizations of bounds on the index of convergence to weighted digraphs. To appear in *Discrete Applied Mathematics*, arXiv:1307.3716v2 [math.CO], 2014.

[100] Glenn Merlet, Thomas Nowak, and Sergeĭ Sergeev. Weak CSR expansions and transience bounds in max-plus algebra. *Linear Algebra and its Applications*, 461:163–199, 2014.

[101] Thomas Nowak and Bernadette Charron-Bost. An overview of transience bounds in max-plus algebra. In G. L. Litvinov and S. N. Sergeev, editors, *Tropical and Idempotent Mathematics and Applications*, volume 616 of *Contemporary Mathematics*, pages 277–289. American Mathematical Society, Providence, 2014.

[102] Thomas Nowak, Matthias Függer, and Alexander Kößler. On the performance of a retransmission-based synchronizer. In Adrian Kosowski and Masafumi Yamashita, editors, *Proceedings of the 18th International Colloquium on Structural Information and Communication Complexity (SIROCCO)*, volume 6796 of *Lecture Notes in Computer Science*, pages 234–245. Springer, Heidelberg, 2011.

[103] Thomas Nowak, Matthias Függer, and Alexander Kößler. On the performance of a retransmission-based synchronizer. *Theoretical Computer Science*, 509:25–39, 2013.

Comportement transitoire d'algorithmes distribués et modèles de circuits

Thomas Nowak

Le thème global de la thèse est le comportement transitoire de certains systèmes répartis. Les résultats peuvent être divisés en trois groupes : transients de matrices et systèmes max-plus, convergence de systèmes de consensus asymptotique et la modélisation de "glitches" dans des circuits numériques.

Pour l'algèbre max-plus, les résultats sont des bornes supérieures sur les transients de matrices et système linéaires max-plus. Elles améliorent strictement les bornes publiées. La thèse inclut une discussion de l'impact des bornes dans des applications. Les preuves utilisent notamment des réductions de chemins. La thèse contient aussi des bornes plus précises pour les transients des indices critiques. Ces bornes sont, en fait, indépendantes des poids spécifiques et ne dépendent que de la structure du graphe de la matrice et son graphe critique. De plus, elles sont des généralisations strictes des bornes booléennes pour des graphes non pondérés ; par exemple les bornes de Wielandt ou de Dulmage et Mendelsohn.

Quant au consensus asymptotique, la thèse améliore des bornes supérieures sur le taux de convergence et établit de nouveaux résultats sur la convergence dans le cas où les agents n'ont pas nécessairement de confiance en soi, c'est-à-dire qu'ils peuvent ignorer leurs propres valeurs. Ces résultats sont notamment pour des réseaux complètement dynamiques. Elle contient aussi un exemple d'un réseau complètement statique dont le taux de convergence est dans le même ordre que celui d'une grande classe de réseaux dynamiques.

La dernière partie de la thèse est sur la propagation de "glitches" (signaux transitoires très courts) dans des circuits numériques. Plus spécifiquement, elle traite des modèles à valeur discrète et temps continu pour des circuits numériques. Ces modèles sont utilisés dans des outils pour la conception de circuits car ils sont beaucoup plus vites que la résolution des équations différentielles. Cependant, comme c'est prouvé dans la thèse, les modèles existants ne prédisent pas correctement l'occurrence de glitches dans le signal sortant d'un circuit. De plus, la thèse contient une proposition d'un nouveau modèle qui ne partage pas les caractéristiques avec les modèles existants qui leur interdisent de prédire correctement l'occurrence de glitches.

Mots clés : systèmes distribués ; systèmes dynamiques ; comportement transitoire

Transient Behavior of Distributed Algorithms and Digital Circuit Models

Thomas Nowak

The overall theme of the thesis is the transient behavior of certain distributed systems. The results can be grouped into three different categories: Transients of max-plus matrices and linear systems, convergence of asymptotic consensus systems, and glitch modeling in digital circuits.

For max-plus algebra, the results are upper bounds on the transient (coupling time) of max-plus matrices and systems. They strictly improve all existing transience bounds. An account of the impact of these bounds in applications is given. The proofs mainly consist of walk reduction and completion procedures. For critical indices, sharper bounds are possible. In fact, they turn out to be independent of the specific weights, and to only depend on the structure of the matrix's digraph and its critical digraph. They are also strict generalizations of the Boolean transience bounds in non-weighted digraphs by the likes of Wielandt or Dulmage and Mendelsohn.

For asymptotic consensus, i.e., a set of agents possessing a real value each and repeatedly updating it by forming weighted averages of its neighbors' values, the thesis strengthens certain upper bounds on the rate of convergence and shows new convergence results for the case of non self-confidence, i.e., agents possibly disregarding their own value. Asymptotic consensus can be described by a non time-homogeneous linear system in classical algebra. The results here are typically in completely dynamic networks. The thesis also presents a worst-case example that shows that exponentially large convergence time is possible even in static networks; meaning that the worst case convergence time in large classes of dynamic networks is actually achieved with a completely static one.

The last part of the thesis is about glitch propagation in digital circuits. More specifically, it is about discrete-value continuous-time models for digital circuits. These models are used in hardware design tool chains because they are much faster than numerically solving the differential equations for timing simulations. However, as is shown in the thesis, none of the existing discrete-value models can correctly predict the occurrence of glitches (short pulses) in the output signal of circuits. Moreover, the thesis proposes a new discrete-value model and proves analytically that it does not share the same characteristics with the existing models that prevented them to correctly predict glitches.

Keywords: distributed systems; dynamical systems; transient behavior