# Mutual Hazard Networks: Markov chain models of cancer progression


Dissertation zur Erlangung des Doktorgrades
der Naturwissenschaften (Dr. rer. nat.)
der Fakultät für Biologie und Vorklinischen Medizin
der Universität Regensburg

vorgelegt von

**Rudolf Schill**

aus Akmola, Kasachstan

im Jahr 2022

Das Promotionsgesuch wurde eingereicht am: 05.09.2022
Die Arbeit wurde angeleitet von: Prof. Dr. Rainer Spang

Unterschrift:

Rudolf Schill, Regensburg, den 05.09.2022

# Acknowledgements

This work was carried out in the Department of Statistical Bioinformatics of the Institute for Functional Genomics at the University of Regensburg. I thank all my past and present colleagues for the pleasant work atmosphere and fruitful scientific and non-scientific conversations, in particular my office room mates Marian Schön, Franziska Görtler and Stefan Hansch, and my fellow students Michael Huttner and Tobias Schmidt. I also want to thank Sharon Petersen, Christian Kohler and Claudio Lottaz for their steadfast help with any administrative, technical or academic matters.

This thesis is part of an interdisciplinary project with many contributors without whom this would not have been possible. I'd like to thank my colleagues, students and friends who accompanied me on this journey, especially Kevin Rupp, Maren Klever, Peter Georg, Andreas Lösch, Jonas Süskind, and Stefan Vocht.

I'd like to thank Tilo Wettig from our Physics Department and Lars Grasedyck from the RWTH Aachen for facilitating and participating in this thoroughly enjoyable collaboration. I am especially grateful to my advisor, Rainer Spang, for more than a decade of guidance, inspiration and freedom that made my studies and my work here a formative experience.

Finally, I'd like to thank my parents and my sister for their constant love and support.

## Publications

Parts of thesis have been published in [67]. This includes chapters 2, 3 and 4 and parts of the introduction and abstract. I contributed as the first author and developer of the presented algorithms.

Chapter 5 of this thesis has been made public as a preprint [65]. I contributed as shared first author and mentor during the development of the algorithms.

# Abstract

Cancer progresses by accumulating genomic events, such as mutations and copy number alterations, whose chronological order is key to understanding the disease but difficult to observe. Instead, cancer progression models use co-occurrence patterns in cross-sectional data to infer dependencies between events and thereby uncover their most likely order of occurrence.

Here we introduce Mutual Hazard Networks, a new class of models that improve upon the state of the art by allowing stochastic dependencies between events, inhibiting dependencies and dependencies that form cycles. MHNs model events by their spontaneous rate of occurrence and by multiplicative effects they exert on the rates of successive events. We further propose an approach for modeling and predicting pivotal events such as the diagnosis of the tumor, temporary inflammation of the tumor, seeding of a metastasis or death of the patient.

To this end we formulate an MHN as a large, continuous-time Markov chain whose transition rate matrix is given as a sum of tensor products. We develop efficient algorithms for computing its transient, stationary and time-marginal probability distributions as well as their derivatives with respect to model parameters in order to perform inference.

First results indicate that MHNs consistently outperform state-of-the-art models on publicly available data in terms of cross-validated model fit. In particular, MHN inferred from a glioblastoma dataset from The Cancer Genome Atlas that IDH1 mutations are early events that promote subsequent mutations in TP53, a finding that is independently supported by consecutive biopsies.

Moreover, we demonstrate the general usefulness of our method beyond oncology by applying it to a stochastic SIR model of epidemic spread. We use our algorithm for computing the derivative of a transient distribution to estimate the monthly infection and recovery rates during the first COVID-19 wave in Austria in a full Bayesian analysis.

# Zusammenfassung

Krebs schreitet fort, indem er genomische Ereignisse wie Mutationen und Kopienzahländerungen anhäuft. Ihre chronologische Reihenfolge ist der Schlüssel zum Verständnis der Krankheit, aber schwer zu beobachten. Stattdessen verwenden Krebsprogressionsmodelle Muster von gemeinsam aufgetretenen Ereignissen in Querschnittsdaten, um auf Abhängigkeiten zwischen Ereignissen zu schließen und dadurch ihre wahrscheinlichste Reihenfolge aufzudecken.

In dieser Arbeit führen wir Mutual Hazard Networks ein, eine neue Klasse von Modellen, die den Stand der Technik verbessern, indem sie stochastische Abhängigkeiten zwischen Ereignissen zulassen, als auch inhibierende und zyklische Abhängigkeiten. MHNs modellieren Ereignisse durch ihre spontane Auftrittsrate und durch multiplikative Effekte, die sie auf die Raten nachfolgender Ereignisse ausüben. Wir schlagen ferner einen Ansatz zur Modellierung und Vorhersage von entscheidenden Ereignissen vor, wie der Diagnose des Tumors, der vorübergehenden Entzündung des Tumors, der Disseminierung einer Metastase oder dem Tod des Patienten.

Dazu formulieren wir ein MHN als große, zeitkontinuierliche Markov-Kette, deren Übergangsratenmatrix als Summe von Tensorprodukten gegeben ist. Wir entwickeln effiziente Algorithmen zur Berechnung ihrer transienten, stationären und zeitmarginalen Wahrscheinlichkeitsverteilungen sowie ihrer Ableitungen nach Modellparametern, um Inferenz durchzuführen.

Erste Ergebnisse deuten darauf hin, dass MHNs durchgängig State-of-the-Art-Modelle auf öffentlich zugänglichen Daten in Bezug auf die kreuzvalidierte Modellanpassung übertreffen. Insbesondere leitete MHN aus einem Glioblastom-Datensatz von The Cancer Genome Atlas ab, dass IDH1-Mutationen frühe Ereignisse sind, die nachfolgende Mutationen in TP53 fördern, ein Befund, der unabhängig durch konsekutive Biopsien gestützt wird.

Darüber hinaus demonstrieren wir die allgemeine Nützlichkeit unserer Methode über die Onkologie hinaus, indem wir sie auf ein stochastisches SIR-Modell der Ausbreitung von Epidemien anwenden. Wir verwenden unseren Algorithmus zur Berechnung der Ableitung einer transienten Verteilung, um die monatlichen Infektions- und Genesungsraten während der ersten COVID-19-Welle in Österreich in einer vollständigen Bayes'schen Analyse zu schätzen.

# Contents

# Introduction

Tumors turn malignant in a Darwinian evolutionary process by accumulating genetic mutations, copy number alterations, changes in DNA methylation, gene expression and protein concentration. Such progression events arise in individual tumor cells, but their effect on the reproductive fitness of this cell depends on earlier events [57], which makes some chronological sequences of alterations more likely than others. These sequences and their driving dependencies are a priori unknown and inferring them from data is the goal of cancer progression models. Such models can then be used to predict the future course of progression in new patients and guide therapeutic interventions.

While progression is a dynamic process, available genotype data are cross-sectional and combine static snapshots from different tumors at different stages of development. Nevertheless, assuming that the tumor genomes are observations from the same stochastic process, cancer progression models can infer dependencies between events from their co-occurance patterns. The dependencies are then reported as a directed graph, where each node stands for an event whose probability depends in some way on the events connected to it by incoming edges (Fig. 0.1). For example, one family of models (reviewed by [42]) approximate tumor progression by deterministic dependencies: An event has a non-zero probability if and only if all its parent events have occurred. These models were inspired by Fearon and Vogelstein [29], who inferred that colorectal cancers progress along a chain of mutations in the genes *APC* → *K-RAS* → *TP53*. [24] formalized and extended this concept to Oncogenetic Trees, where a single event may be necessary for multiple successor events in parallel. [8] further generalized these to Conjunctive Bayesian Networks (CBN), where events may require multiple precursors, thus replacing trees by directed acyclic graphs.

In this thesis, we relax three assumptions of this model family:

1. Dependencies need not be deterministic. An event A can make an event B more likely without being absolutely necessary for it. In particular, events can occur with non-zero probability at all times and all event patterns are possible.

2. A dependency graph need not be acyclic. Clearly an event A cannot be necessary for B if B is also necessary for A, but it is certainly possible that A makes B more likely when it occurs first, and vice versa.

3. Besides enabling dependencies, there are also inhibiting dependencies.

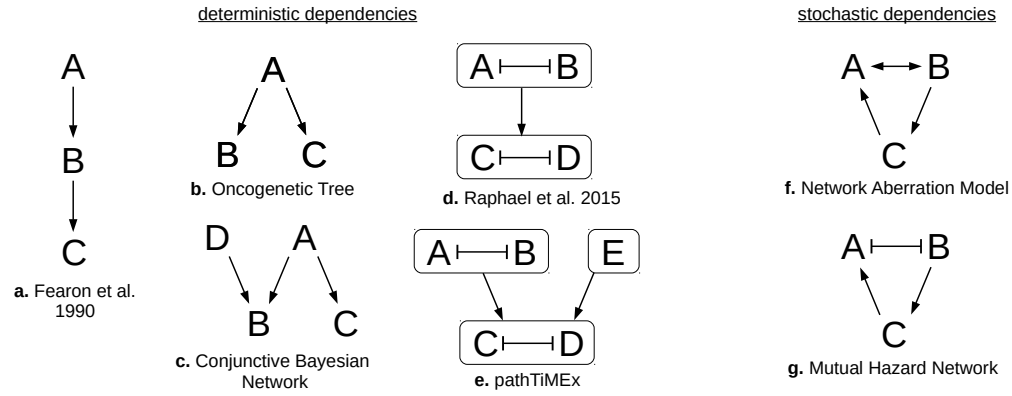



**Figure 0.1.:** Overview of several types of cancer progression models. For models with deterministic dependencies (a-e) $A \rightarrow B$ denotes that A is necessary for B, and $A \dashv B$ denotes that A prevents B. For models with stochastic dependencies (f-g) $A \rightarrow B$ denotes that A makes B more likely, and $A \dashv B$ denotes that A makes B less likely. In (d-e) the arrows between groups of events denote that at least one of the events in the parent group is necessary for the events in the child group.

Although there is, to the best our knowledge, no method that addresses all three issues, there are methods that address one or two of them. Stochastic dependencies (1) have been previously proposed in [28, 53, 59] for acyclic models. Moreover, stochasticity at the point of observation has been addressed by [34] who allow for mislabeled events, and in [9, 55] who treat tumor data as a mixture from multiple stochastic processes. Network Aberration Models (NAM) by [44] have stochastic dependencies (1) and allow cycles in their dependency graph (2).

Inhibition (3) is at the center of mutual exclusivity, which is a frequently observed phenomenon in cancer [80]. Two events are considered mutually exclusive if they co-occur less frequently than expected by chance. There are at least two mechanisms that can cause this data pattern: One is synthetic lethality, where cells carrying two mutations A and B are no longer vital. Alternatively, the events disrupt the same molecular pathway such that whichever event occurs first conveys most of the selective advantage and decreases selective pressure for the others. Both mechanisms can be described by a double edge A $\vdash\dashv$ B (A inhibits B, and B inhibits A)

The currently prevalent approach to including mutual exlusivity in a progression model was introduced by [34] who first grouped events into pathways and in a second step learned acyclic models on the resolution of pathways. Pathways can either be derived from biological knowledge, learned from data by testing groups of events for mutual exclusivity [49, 69, 20] or by a combination of both [19, 47], see [72] for a review. [60] pointed out that inferring pathways separately from their dependencies can lead to

inconsistencies in the presence of noise. They presented the first algorithm that simultaneously groups events into pathways and arranges the pathways in a linear chain. PathTiMEx [23] generalizes this from linear chains to acyclic progression networks (CBN).

Building on both CBNs and NAMs, we propose Mutual Hazard Networks (MHN). MHNs do not group events into pathways but directly model the mechanisms behind mutual exlusivity. MHNs have cyclic dendency networks, in particular allowing for bidirectional and inhibiting edges. MHNs characterize events by a baseline rate and by multiplicative effects they exert on the rates of successive events. These effects can be greater or less than one, i.e., promoting or inhibiting. Moreover we extend MHNs by special events that play an important role in tumor progression, such as diagnosis of the tumor itself, death of the patient, seeding of metastasis or reversible events.

We propose a mathematical framework for learning the parameters of such an MHN via maximum likelihood which is tractable for datasets with up to 25 events per individual patient and potentially hundreds of events overall. To this end we formulate an MHN as a continuous-time Markov chain with a transition rate matrix that is given as a sum of tensor products. We provide efficient algorithms for performing important operations with this matrix, such as the derivative of the matrix exponential or the derivative of the resolvent with respect to the parameters.

Finally, we show that our framework is useful beyond the field of computational oncology. We apply our novel algorithms to parameter inference of stochastic models of epidemic spread as well as stochastic predator-prey models, which have been commonly considered intractable.

# 1. Biological preliminaries

*In this chapter we provide biological background knowledge on tumor progression. We describe it as an evolutionary process that is driven by progression events. We provide examples of progression events, categorize these by their physiological effects, and explain possible dependencies between them. We then describe how progression event data of tumors are acquired and processed before they can be used by progression models.*

## 1.1. Tumor progression as evolutionary process

Tumor progression can be understood as a step-wise, evolutionary process [10, 57, 77] where an initially healthy tissue acquires abnormal traits and behaviors that render it increasingly harmful to its host. It ranges from benign tumors that grow locally to abnormal size but are otherwise harmless, unless pressing on vital organs, to malignant tumors (cancer) that invade nearby tissues or seed metastases, leading to harm or death.

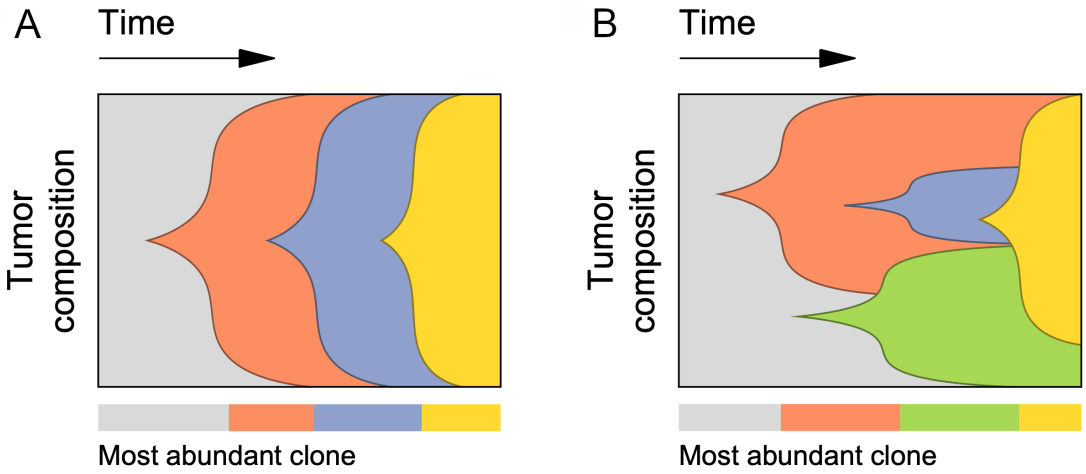



**Figure 1.1.:** Schematic representation of tumor progression as an evolutionary process (A) with "strong selection, weak mutation" (SSWM) assumption and (B) without SSWM assumption. Figure adapted from [25].

Such traits arise at first randomly in individual cells due to the inherently error-prone replication of the genome during cell division. Mutations or copy-number alterations of genes lead to under- or overactivity of the respectively encoded proteins and thereby to changes in the cell's physiology. Usually these changes are without consequence or harmful to the cell and die out again. However, when they provide a selective advantage to the cell they can allow its clones to proliferate and take over the tumor, which we call a *progression event*. In this thesis we assume for simplicity that each new clone takes over the tumor in its entirety, if at all, and each progression event can reach fixation before the next event. This assumption is known as "strong selection, weak mutation" (Figure 1.1) and allows us to model a tumor at any given time as a homogeneous population of cells with the same genome [25].

Although progression events arise randomly, whether they can reach fixation depends on previous events and the nature of their physiological effects. These can be categorized according to the *six hallmarks of cancer* [43] by Hanahan and Weinberg:

1. **Self-sufficiency in growth signals**. Normal cells grow and multiply in response to external signals such as steroid hormones or secreted proteins. Cancer cells can produce and supply these signals to themselves or trigger a growth response even in their absence.

2. **Insensitivity to anti-growth signals**. Normal cells are kept from excessive growth by signals from their surrounding tissue. These can be soluble growth inhibitors or inhibitors embedded in the extracellular matrix or on the surface of nearby cells. Cancer cells are able to ignore such signals, for example through mutations in the gene RB-1.

3. **Evading programmed cell death**. Normal cells can undergo orderly self-destruction (apoptosis) in response to stressors such as infection by a virus, lack of oxygen, or DNA damage from UV radiation. Cancer cells can evade this mechanism, for example through mutations in the gene TP53.

4. **Limitless replicative potential**. Normal cells other than stem cells can divide for only about 60 - 70 generations before they go into senescence and stop growing. This limit is imposed by the shortening of telomeres at the end of chromosomes during each division. Cancer cells circumvent this limit for example by upregulation of telomerase, a ribonucleoprotein that extends telomeres.

5. **Sustained angiogenesis**. The formation of new blood vessels takes place in normal tissues during embryonal development and wound healing, but is otherwise tightly regulated. Cancers hijack this mechanism in order to supply themselves with vital nutrients and oxygen.

6. **Tissue invasion and metastasis**. Individual cancer cells in a malignant primary tumor can break away from their tissue of origin, invade nearby tissues or

travel through blood vessels to distant organs. There they form colonies (metastases) which are the most common cause of human cancer deaths, rather than proliferation of the primary tumor.

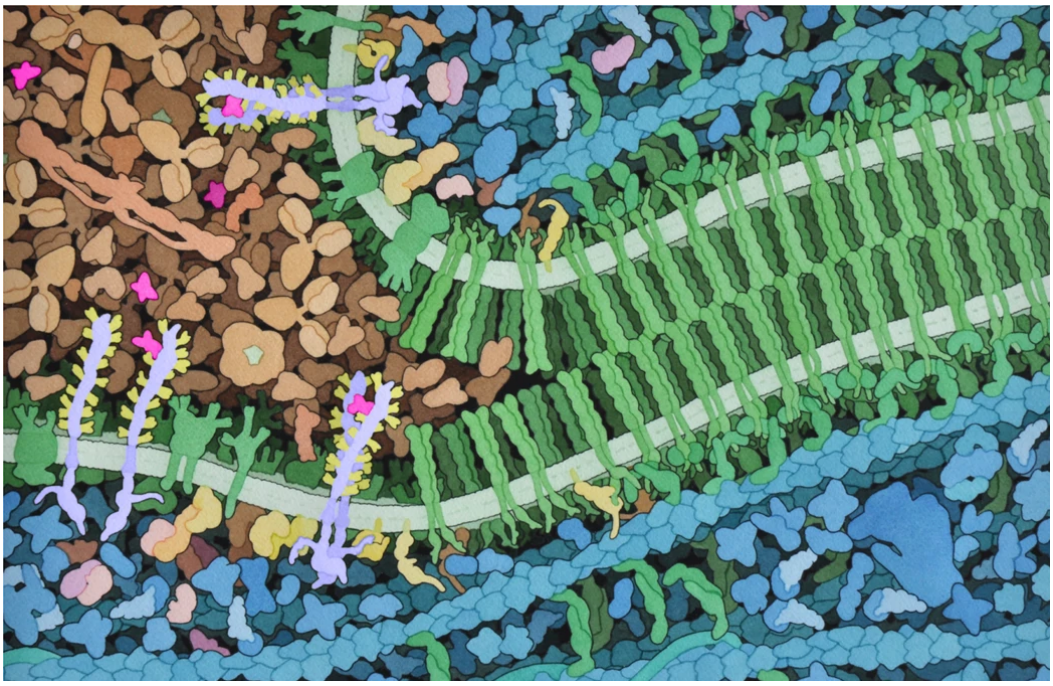
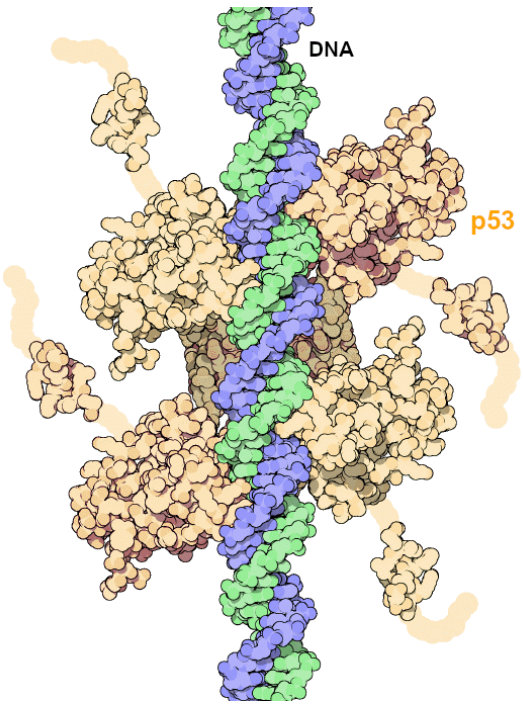


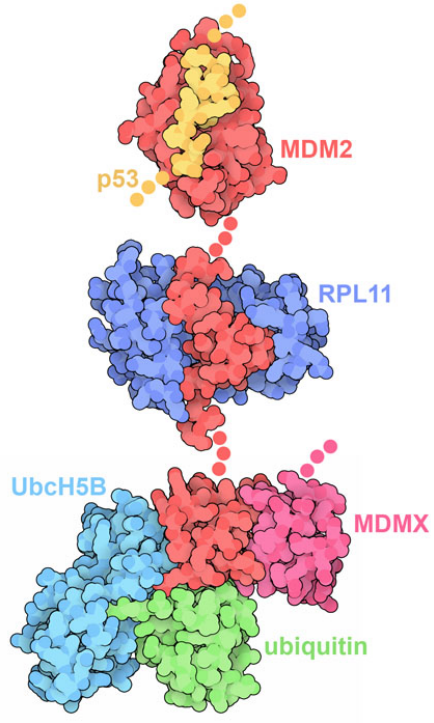
**Figure 1.2.:** Artistic conception of angiogenesis. Tumors can overproduce VegF signaling proteins (magenta) which travel through the blood plasma (tan) to a juncture between cells. VegF binds together two copies of VegFR (lavender/yellow) on the cell surface, triggering an intracellular signaling cascade which makes the cadherin proteins (green) between the two cells (blue) separate, making room for new blood vessels. Figure adapted from [37, 40].

As an example of enabling dependencies between progression events, consider an event that falls under categories (1) or (2). It can cause a primary tumor to grow up to a size of about a cubic millimeter, after which it will be starved for oxygen [39]. This state of hypoxia leads normal cells into apoptosis, thus favoring cells that can prevail due to subsequent progression events in category (3). Moreover, it provides a strong selective advantage to the formation of new blood vessels (5) which can supply the tumor with nutrients and oxygen for further growth (Figure 1.2). Access to blood vessels is in turn a prerequisite for the ability to metastasize (6).

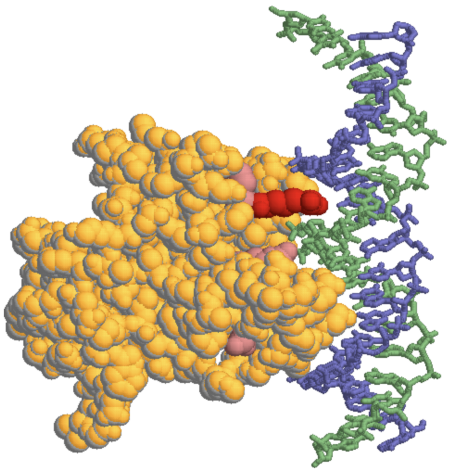
As an example of inhibiting dependencies between progression events, consider the protein p53 (Figure 1.3a) which is encoded by the gene TP53 and known as the "guardian of the genome" [37]. It accumulates in response to stressors such as DNA damage or hypoxia and then binds to specific other genes in the genome. These begin




**(a)** The protein p53 (gold) can bind to the DNA (green/purple) of certain genes that have a specific binding site for p53. By recruiting transcription enzymes p53 activates these genes which are responsible for DNA repair, growth arrest or cell death. Adapted from [37].




**(b)** The protein MDM2 (red) can bind to p53 (gold), deactivate it and transport it out of the nucleus. In the cytoplasm it marks p53 for destruction by combining it with ubiquitin (green). Image from [38].



**(c)** One of the four arms of p53 (gold) is shown. The most common p53 mutation in cancer changes the amino acid arginine 248 (red) which fits in the minor groove of the DNA, thereby disrupting the ability of p53 to bind. Adapted from [37].

 **Figure 1.3.:** Artistic visualization of parts of the p53 signaling pathway.

production of proteins that halt cell division until the damage is repaired, or initiate apoptosis if the damage is too large. This rather severe response is tightly regulated and kept in check by the protein MDM2 (Figure 1.3b) which binds to p53, deactivates it and carries it out of the cell nucleus for destruction ("guarding the guardian") [38].

Tumors have at least two different ways to disrupt this balanced process and thereby evade apoptosis (3). Either directly through a deletion or mutation of TP53 (Figure 1.3c), or they can render p53 ineffective by amplification of the gene for MDM2. Since both progression events have a similar physiological consequence, whichever event comes first will decrease selective pressure for the other. In a tumor progression model this can be represented as a mutually inhibiting dependency.

Our aim is to infer and quantify such relationships from data.

## 1.2. Cross-sectional bulk data of progression events

While tumor progression is a dynamic process, most of the available tumor genotype data are *cross-sectional*. That is, rather than observing any given patient over multiple consecutive time points, we have to rely on observations from many different patients at a single time point each. This is because the biopsy of a tumor is an invasive procedure whose risk must be medically justified. Only in rare cases are tumors removed repeatedly, such as for certain brain tumors where each subsequent growth can be life threatening [76].

Moreover, in this thesis we limit ourselves to datasets of tumors that have been *bulk sequenced*. That is, rather than observing the genome of any particular cell, we observe the genome of the most abundant clone (see Figure 1.1) or rather a mix of all subclone genomes in the tumor. While single-cell sequencing has become increasingly practical in recent years, it is still error-prone and too expensive for large datasets [70]. This is also the primary justification for the SSWM assumption, in addition to simpler modeling.

In high-throughput bulk sequencing the DNA is first extracted from the tumor sample and fragmented into small reads which are then amplified and sequenced individually (see Figure 1.5). The small reads are overlapped and computationally aligned to a generic human reference genome, which allows us to detect several types of progression events in the tumor genome (Figure 1.4).

Deviations from the expected coverage indicate copy-number alterations, i.e., deletions or amplifications of parts of the genome. After thresholding and statistical pre-processing [5], these can be interpreted as deletions or amplifications of particular genes or group of genes. We consider these events as binary, present or not present. Similarly, we can detect single nucleotide variants (SNPs) which we also consider as

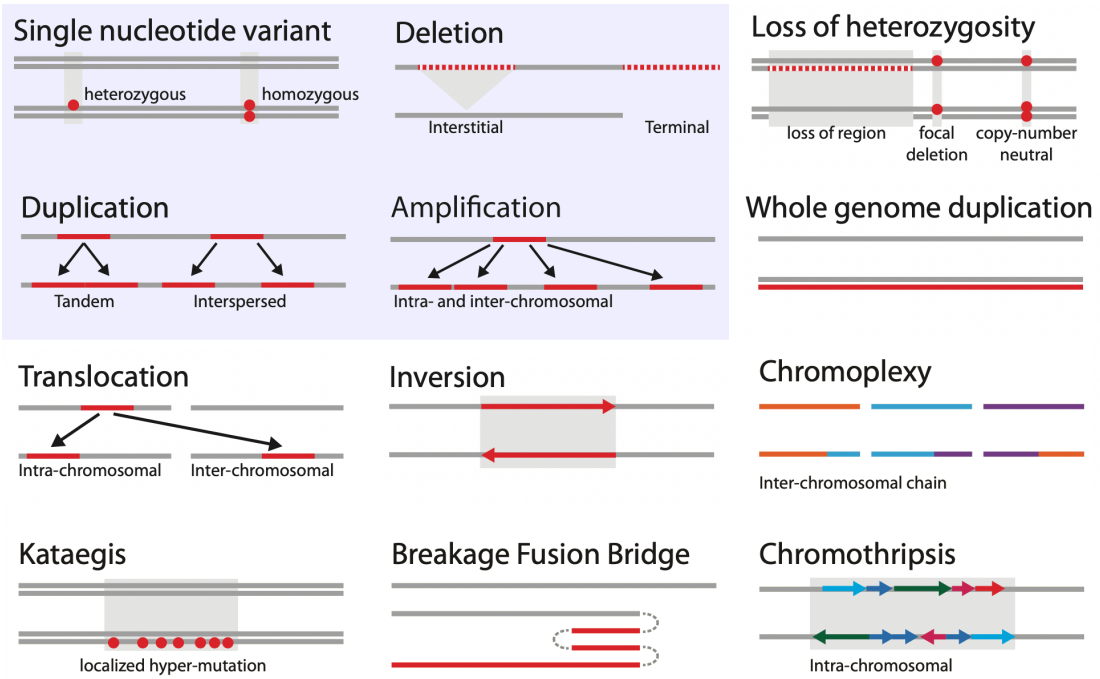



**Figure 1.4.:** Common progression events in cancer genomes. Grey lines indicate the healthy genome on top and cancer genome below. Double lines are used when differentiating heterozygous and homozygous changes is useful. Dots represent single nucleotide changes, whereas lines and arrows represent structural changes. In this thesis we focus on simple copy-number alterations and mutations, highlighted in the upper left in blue. Figure adapted from [10].

binary mutation events, present or not present, ignoring the type of mutation and its exact location in the gene as well as zygosity.

An important database of publicly available datasets is The Cancer Genome Atlas (TCGA) [18], a project initiated in 2006 by the National Cancer Institute and the National Human Genome Research Institute which characterized over 20000 primary tumors from 33 cancer types. It also includes useful clinical information, such as age, sex or smoking status of the patient, as well as follow up survival time in some cases.

In addition to data from TCGA, we also use much older datasets from the Progenetix molecular-cytogenetic database [6]. These were used in the initial publications on Conjunctive Bayesian Networks (CBNs) and hence allow for a fair comparison between our approaches in terms of modeling performance. They were generated using the technique of *comparative genomic hybridization (CGH)* and consist of copy-number alterations on the low resolution of whole chromosome arms, rather than on the level of genes.

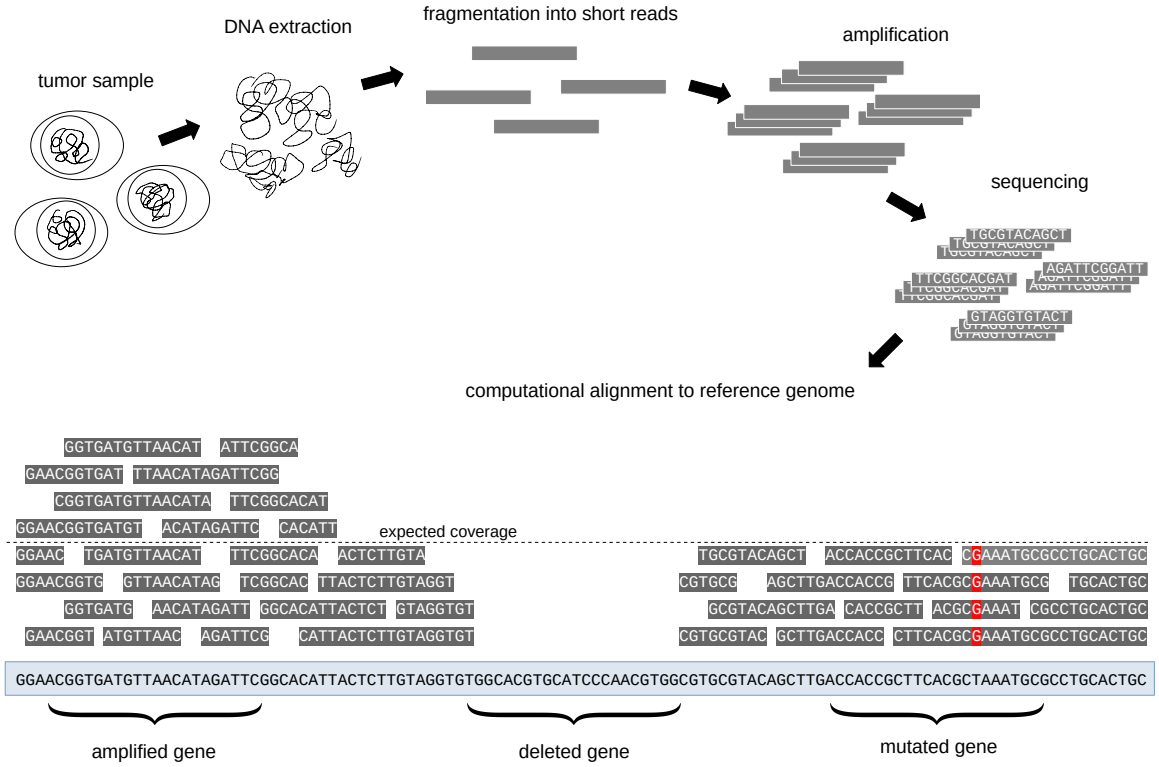



**Figure 1.5.:** Schematic explanation of the bulk sequencing procedure that generates progression event data from tumor samples.

# 2. Mutual Hazard Networks

*In this chapter we introduce the central subject of this thesis, a new class of tumor progression models which we call Mutual Hazard Networks. We formally define a Mutual Hazard Network as a continuous-time Markov chain on a large, combinatorial state space which has a certain parameterization that can be understood as several interlocking Cox proportional hazards models. We show that this parameterization implies that its transition rate matrix can be written as a sum of tensor products, and how one can exploit this structure in order to efficiently perform parameter inference via maximum likelihood.*

## 2.1. Definition

We model tumor progression as a continuous time Markov process $\{X(t), t \geq 0\}$ on all $2^n$ combinations of a predefined set of $n$ events. Its state space is $S = \{0,1\}^n$, where $X(t)_i = 1$ means that event $i$ has occurred in the tumor by age $t$, while $X(t)_i = 0$ means that it has not.

We assume that every progression trajectory starts at a normal genome $X(0) = (0,\dots,0)^T$, accumulates irreversible events one at a time, and ends at a fully aberrant genome $X(\infty) = (1,\dots,1)^T$. Observed tumor genomes correspond to states at unknown intermediate ages $0 < t < \infty$ and typically hold both 0 and 1 entries.

Let $Q \in \mathbb{R}^{2^n \times 2^n}$ be the transition rate matrix of this process with respect to a basis of $S$ in lexicographic order (Figure 2.1, top). An entry

$$Q_{\mathbf{y},\mathbf{x}} = \lim_{\Delta t \to 0} \frac{\Pr\left(X(t+\Delta t) = \mathbf{y} \mid X(t) = \mathbf{x}\right)}{\Delta t}, \quad \mathbf{y} \neq \mathbf{x} \tag{2.1}$$

is the rate from state $\mathbf{x} \in S$ to state $\mathbf{y} \in S$, and diagonal elements are defined as $Q_{\mathbf{x},\mathbf{x}} = -\sum_{\mathbf{y} \neq \mathbf{x}} Q_{\mathbf{y},\mathbf{x}}$ so that columns sum to zero. $Q$ is lower triangular and has non-zero entries only for transitions between pairs of states $\mathbf{x} = (\dots, x_{i-1}, 0, x_{i+1}, \dots)^T$ and $\mathbf{y} = \mathbf{x}_{+i} := (\dots, x_{i-1}, 1, x_{i+1}, \dots)^T$ that differ in a single entry $i$.

Our aim is to learn for each event $i$ how its rate $Q_{\mathbf{x}_{+i},\mathbf{x}}$ depends on already present events in $\mathbf{x}$ as a function $f_i : \{0,1\}^n \to \mathbb{R}$. A common choice in time-to-event analysis is the proportional hazards model [22] which assumes that binary predictors have

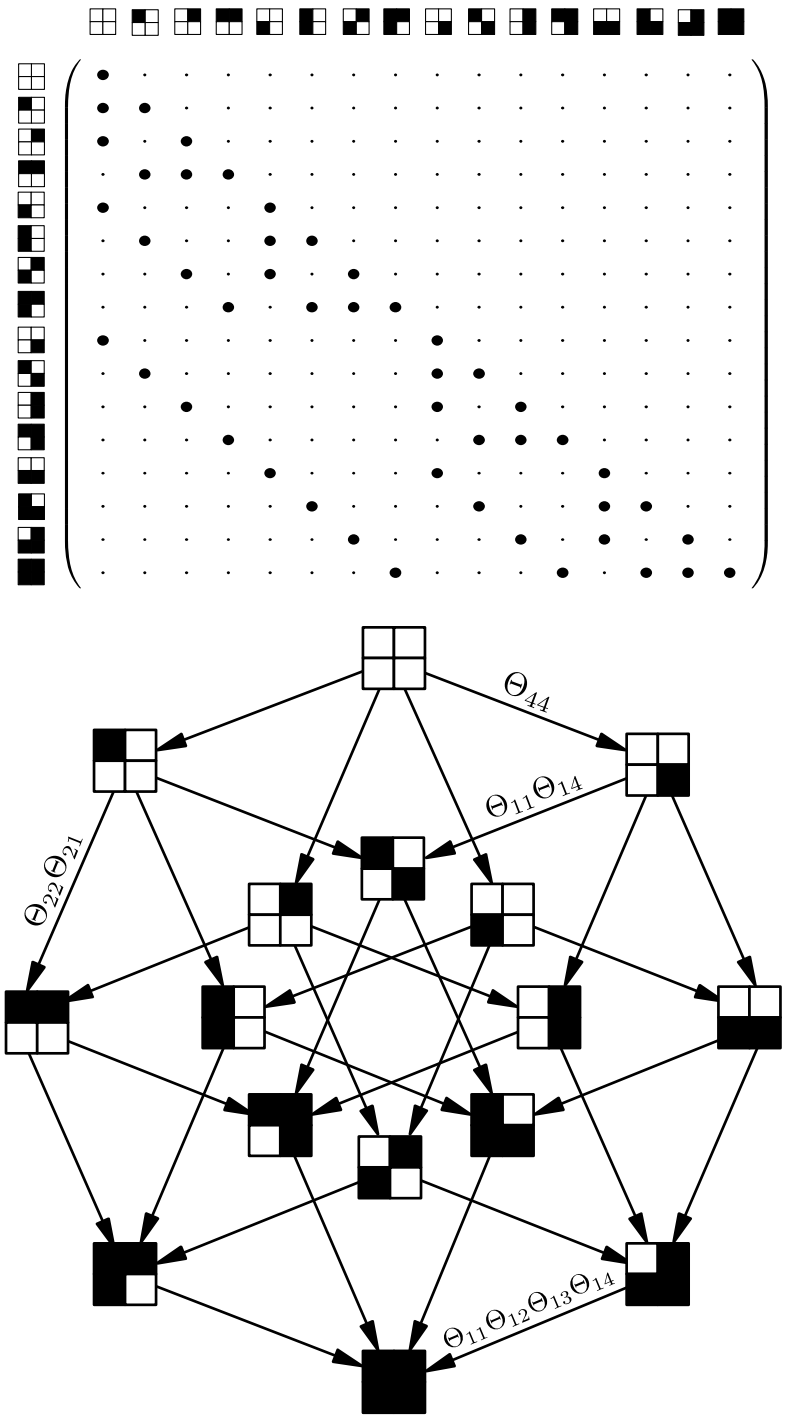



**Figure 2.1.:** (Top) Transition rate matrix $Q$ for the Markov process $X$ with $n = 4$, where $\cdot$ is a zero entry and $\bullet$ is a non-zero entry. The states are depicted as squares with four compartments as shown below. A white compartment denotes 0 and a black compartment denotes 1. The matrix is lower triangular because events are irreversible, and sparse because events accumulate one at a time. (Bottom) Parameterization $Q_\Theta$ of the Markov process by a Mutual Hazard Network.

independent and multiplicative effects on the rate of the event. We therefore specify the Markov process by a system of $n$ functions

$$Q_{\mathbf{x}_{+i},\mathbf{x}} = f_i(\mathbf{x}) = \exp\left(\theta_{ii} + \sum_{j=1}^{n} \theta_{ij} x_j\right) = \Theta_{ii} \prod_{x_j=1} \Theta_{ij} \tag{2.2}$$

and collect their parameters in a matrix $(\Theta_{ij}) := (e^{\theta_{ij}}) \in \mathbb{R}^{n\times n}$. We call $\Theta$ a *Mutual Hazard Network (MHN)*, where the baseline hazard $\Theta_{ii}$ is the rate of the event $i$ before any other events are present and the hazard ratio $\Theta_{ij}$ is the multiplicative effect of event $j$ on the rate of event $i$ (Figure 2.1, bottom). Note that while the baseline hazard in [22] is generally a function of time, here it must be constant so that our model constitutes a Markov process.

## 2.2. Parameter inference

A dataset $\mathcal{D}$ of tumors defines an empirical probability distribution on $S$. It can be represented by a vector $\mathbf{p}_\mathcal{D}$ of size $2^n$, where an entry $(\mathbf{p}_\mathcal{D})_\mathbf{x}$ is the relative frequency of observed tumors with state $\mathbf{x}$ in $\mathcal{D}$.

At $t = 0$ tumors are free of any events, so the Markov process $X$ starts with the initial distribution $\mathbf{p}_\varnothing := (100\%, 0\%, ..., 0\%)^T$, which then evolves according to the parameterized rate matrix $Q_\Theta$. If all tumors had been observed at a common age $t$, $\mathbf{p}_\mathcal{D}$ could be modelled as a sample from the transient distribution

$$e^{tQ_\Theta}\, \mathbf{p}_\varnothing\,. \tag{2.3}$$

Since the tumor age is usually unknown, we follow [33] and consider $t$ to be an exponential random variable with mean 1. Marginalizing over $t$ yields

$$\mathbf{p}_\Theta = \int_0^\infty dt\, e^{-t} e^{tQ_\Theta}\, \mathbf{p}_\varnothing = [\underbrace{I - Q_\Theta}_{=:R_\Theta}]^{-1} \mathbf{p}_\varnothing\,, \tag{2.4}$$

and the marginal log-likelihood score of $\Theta$ given $\mathcal{D}$ is

$$\mathcal{S}_\mathcal{D}(\Theta) = \mathbf{p}_\mathcal{D}^T \log \mathbf{p}_\Theta = \mathbf{p}_\mathcal{D}^T \log(R_\Theta^{-1} \mathbf{p}_\varnothing)\,, \tag{2.5}$$

where the logarithm of a vector is taken component-wise.

When optimizing $\mathcal{S}_\mathcal{D}$ with respect to $\Theta$ we are especially interested in networks that can be easily visualized and interpreted, i.e., where many events do not interact and

off-diagonal entries $\Theta_{ij}$ are exactly 1. To this end, we penalize the score with a sparsity-promoting regularization term,

$$\mathcal{S}_{\mathcal{D}}(\Theta) - \lambda \sum_{i \neq j} |\log \Theta_{ij}|\,, \tag{2.6}$$

where $\lambda$ is a tuning parameter. We will optimize this expression using the Orthant-Wise Limited-Memory Quasi-Newton algorithm (OWL-QN) [4]. This general-purpose optimizer takes care of the non-differentiability introduced by the regularization term, while only requiring a closed form for the derivatives $\partial \mathcal{S}_{\mathcal{D}} / \partial \Theta_{ij}$ with respect to each parameter.

From the chain rule of matrix calculus we have

$$\begin{aligned}
\frac{\partial \mathcal{S}_{\mathcal{D}}}{\partial \theta_{ij}} &= \frac{\partial \mathcal{S}_{\mathcal{D}}}{\partial R_\Theta^{-1}} \cdot \frac{\partial R_\Theta^{-1}}{\partial \theta_{ij}} \\
&= \frac{\mathbf{p}_{\mathcal{D}}}{\mathbf{p}_\Theta} \mathbf{p}_\varnothing^T \cdot \left( -R_\Theta^{-1} \frac{\partial R_\Theta}{\partial \theta_{ij}} R_\Theta^{-1} \right) \\
&= -\left( \frac{\mathbf{p}_{\mathcal{D}}}{\mathbf{p}_\Theta} \right)^T R_\Theta^{-1} \frac{\partial R_\Theta}{\partial \theta_{ij}} R_\Theta^{-1} \mathbf{p}_\varnothing\,,
\end{aligned} \tag{2.7}$$

where $\cdot$ is the Frobenius product and the ratio $\mathbf{p}_{\mathcal{D}} / \mathbf{p}_\theta$ is computed component-wise. Note that we optimize with respect to the logarithmic parameters $\theta_{ij}$ in order to ensure the positivity constraint on $\Theta_{ij}$.

In general OWL-QN will converge to one local optimum out of possibly several. In this thesis we always report the optimum reached from an independence model as starting point, where the baseline hazard $\Theta_{ii}$ of each event was set to its empirical odds in the data and all hazard ratios were set to exactly 1. See [74] for a comprehensive analysis of identifiability of MHNs.

To compute the score in equation (2.5) and its gradient in equation (2.7) we must solve the exponentially sized linear systems $[I - Q_\Theta]^{-1} \mathbf{p}_\varnothing$ and $(\mathbf{p}_{\mathcal{D}} / \mathbf{p}_\Theta)^T [I - Q_\Theta]^{-1}$. To this end, we employ the (left) Kronecker product which is defined for matrices $A \in \mathbb{R}^{k \times l}$ and $B \in \mathbb{R}^{p \times q}$ as the block matrix

$$A \otimes B = \begin{bmatrix} b_{11}A & \cdots & b_{1l}A \\ \vdots & \ddots & \vdots \\ b_{k1}A & \cdots & b_{kl}A \end{bmatrix} \in \mathbb{R}^{kp \times lq}\,. \tag{2.8}$$

We follow the literature on structured analysis of large Markov chains [13, 3] and write the transition rate matrix $Q_\Theta$ as a sum of $n$ such Kronecker products,

$$Q_\Theta = \sum_{i=1}^{n} \left[ \bigotimes_{j<i} \begin{pmatrix} 1 & 0 \\ 0 & \Theta_{ij} \end{pmatrix} \otimes \begin{pmatrix} -\Theta_{ii} & 0 \\ \Theta_{ii} & 0 \end{pmatrix} \otimes \bigotimes_{j>i} \begin{pmatrix} 1 & 0 \\ 0 & \Theta_{ij} \end{pmatrix} \right]. \tag{2.9}$$

Here, the $i$-th term in the sum is a sparse $2^n \times 2^n$ matrix consisting of all transitions that introduce event $i$ to the genome. It corresponds to a single subdiagonal of $Q_\Theta$, together with a negative copy on the diagonal to ensure that columns sum to zero (see Figure 2.2). The benefit of this compact representation is that matrix-vector products can be computed in $\mathcal{O}(n2^{n-1})$ rather than $\mathcal{O}(2^{2n})$ without holding the matrix explicitly in memory [16]. We split $R_\Theta = I - Q_\Theta$ into a diagonal and strictly lower triangular part,

$$R_\Theta = D + L = D(I + D^{-1}L)\,, \tag{2.10}$$

and use the nilpotency of $D^{-1}L$ to compute

$$\begin{aligned} R_\Theta^{-1}\mathbf{p}_\varnothing &= (I + D^{-1}L)^{-1}D^{-1}\mathbf{p}_\varnothing \\ &= \left(\sum_{k=0}^{n-1}(-D^{-1}L)^k\right)D^{-1}\mathbf{p}_\varnothing. \end{aligned} \tag{2.11}$$

Finally, in order to compute the gradient $\partial\mathcal{S}/\partial\theta_{ij}$ we exploit the Kronecker structure of the derivative

$$\frac{\partial Q}{\partial\theta_{ij}} = \bigotimes_{k<i}\begin{pmatrix}1-\delta_{kj} & 0\\ 0 & e^{\theta_{ik}}\end{pmatrix} \otimes \begin{pmatrix}-e^{\theta_{ii}} & 0\\ e^{\theta_{ii}} & 0\end{pmatrix} \otimes \bigotimes_{k>i}\begin{pmatrix}1-\delta_{kj} & 0\\ 0 & e^{\theta_{ik}}\end{pmatrix}. \tag{2.12}$$

Using the fact that all matrices $\partial Q/\partial\theta_{ij}$ for a fixed $i$ share the same entries and differ only in a mask of zeros we can share most of the computational effort for computing the derivatives with respect to all parameters (see appendix C). Finally, by applying a state space restriction (see appendix A) we can further reduce the computational cost to exponential in the number of events that have occured per individual patient.

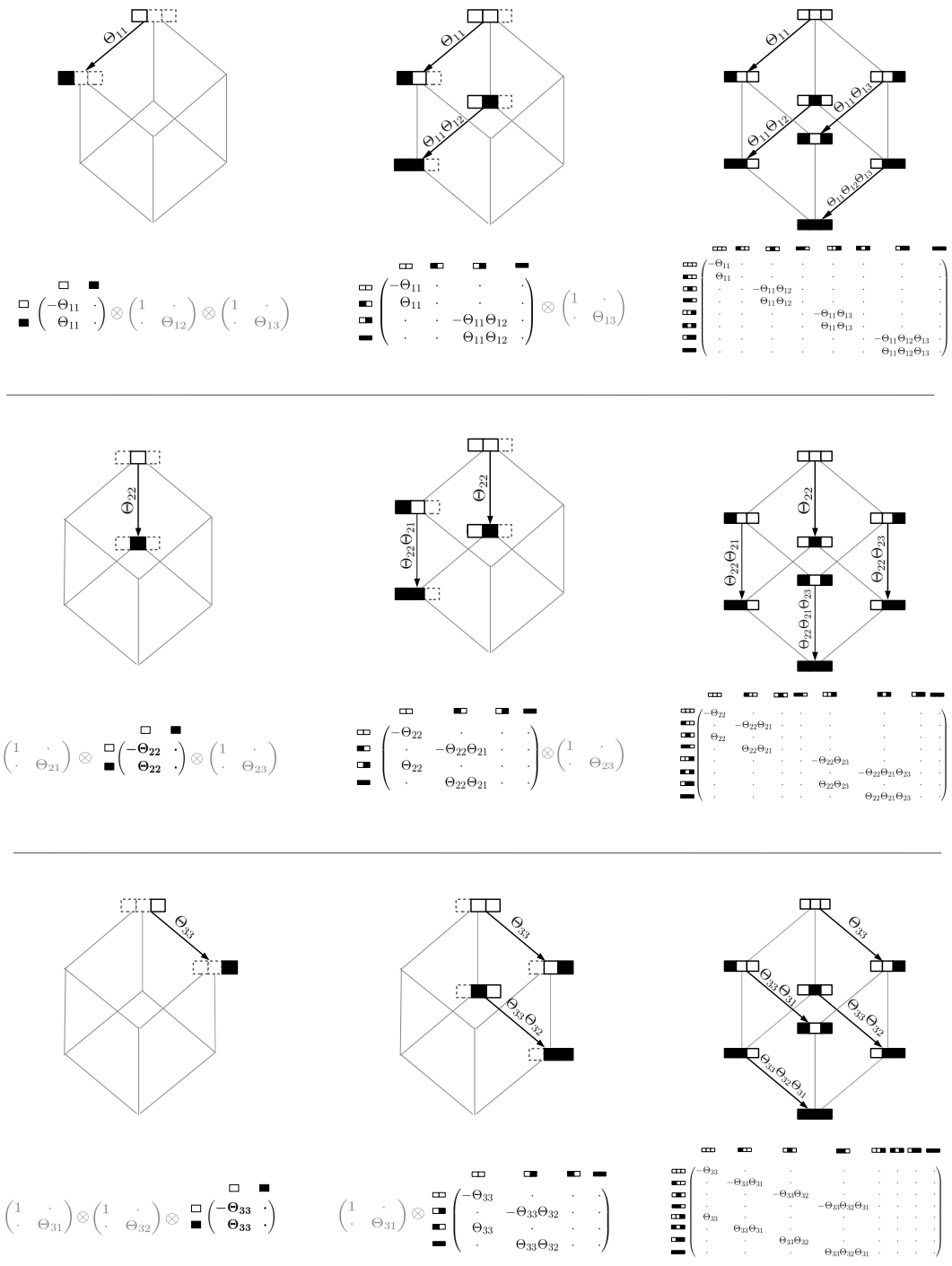


**Figure 2.2.:** Illustration of $Q_\Theta$ represented as a sum of Kronecker products for $n = 3$ in eq. (2.9). The $i$-th row corresponds to the $i$-th term in the sum and contains all transitions that introduce event $i$ to the genome. A row is read from left to right and shows how the Kronecker product successively describes all possible transition rates that can arise due to multiplicative interactions with other events. The first highlighted Kronecker factor describes the two possible states of event $i$ and a transition with base rate $\Theta_{ii}$. Each subsequent Kronecker factor that is multiplied from the left or from the right appends the two states of the corresponding event $j$ to all previously modelled states. This doubles the number of modelled states, where one half lacks the event $j$ and retains their previous transition rates, while the other half has $j$ present, which modulates their transition rates by the factor $\Theta_{ij}$.

# 3. Simulation experiments

We tested in simulation experiments how well an MHN of a given size can learn a probability distribution on $S$ when trained on a given amount of data. We ran 100 simulations for each of several sample sizes $|\mathcal{D}| \in \{50, 100, 250, 500\}$ and number of events $n \in \{10, 15\}$, and 10 simulations for $n = 20$.

In each simulation run, we chose a ground truth model $\Theta$ with $n$ possible events. A random half of its off-diagonal entries were set to 1 (no dependency) and the remaining entries were drawn from a standard log-normal distribution. We then generated a dataset of size $|\mathcal{D}|$ from this model and trained on it another model $\hat{\Theta}$ by optimizing expression (2.6). We chose a common regularization parameter for all 100 (resp. 10) simulation runs, which we found to be roughly $\lambda = 1/|\mathcal{D}|$ through validation on separate datasets of each sample size. We then assessed the reconstructed model $\hat{\Theta}$ by the Kullback-Leibler (KL) divergence from its probability distribution to the distribution of the true model $\Theta$,

$$D_{\mathrm{KL}}(\mathbf{p}_\Theta \| \mathbf{p}_{\hat{\Theta}}) = \mathbf{p}_\Theta^T \log \mathbf{p}_\Theta - \mathbf{p}_\Theta^T \log \mathbf{p}_{\hat{\Theta}} \tag{3.1}$$

The median KL divergence, as well as its variance over the simulation runs, improved with larger training datasets and reached almost zero (Fig. 3.2).

Next, we simulated datasets of size $|\mathcal{D}| = 500$ from random MHNs and CBNs as ground truth models with $n = 8$ events. We added noise by flipping each event independently with probability $\epsilon$, trained MHNs and CBNs on both datasets and evaluated how well the estimated models fit the distribution of the ground truth models. (Fig. 3.3) shows the average KL divergence over 5 simulation runs for each noise level $\epsilon \in \{1\%, 5\%, 10\%, 15\%, 20\}$. For CBNs as ground truth we found that CBNs outperform MHNs when noise is below 10%, while MHNs perform better than CBNs at higher levels of noise. For MHNs as ground truth we found that MHNs perform better than CBNs at all levels of noise.

Lastly, we tested the performance of our implementation. MHN was written in R, and its performance-critical parts were implemented in C (using the R package `inline`) to avoid unnecessary memory-copy operations. We made explicit calls to BLAS routines and compiled R to use the Intel MKL library for vectorized and threaded matrix and vector operations. Fig. 3.1 shows the runtime of a single gradient step for random and dense $\Theta$ on a Dell OptiPlex 9020 workstation with 8GB RAM and an Intel® Core™

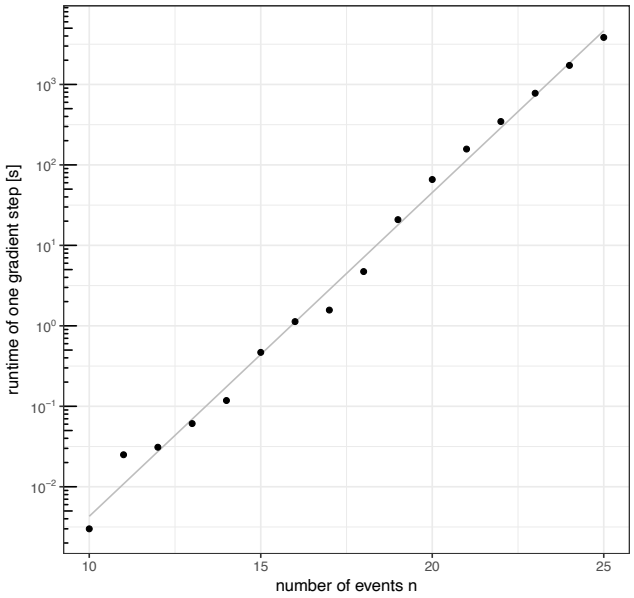



**Figure 3.1.:** Runtime of a single gradient step for random and dense $\Theta$.

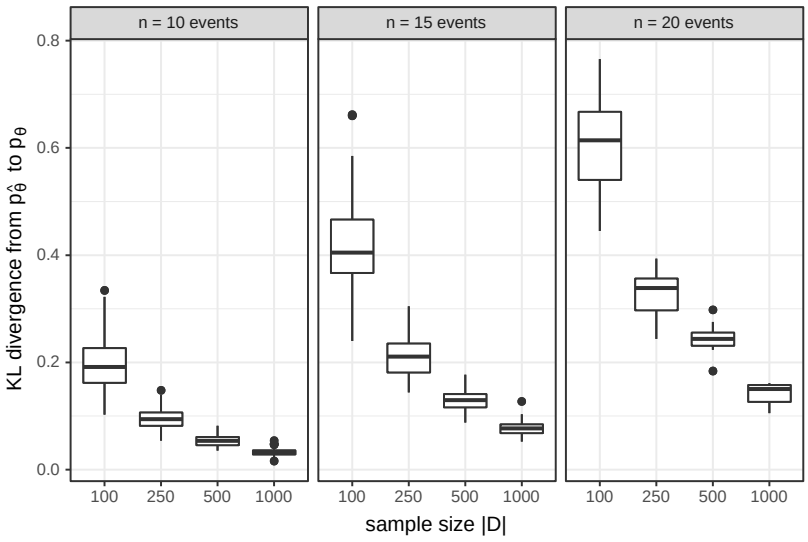



**Figure 3.2.:** Average KL divergence of estimated MHNs to ground truth MHNs over 100 simulations for the shown sample sizes and number of events.

i5-4590 CPU. The runtime was about 1 minute for $n = 20$ and scaled exponentially with $n$ as expected.

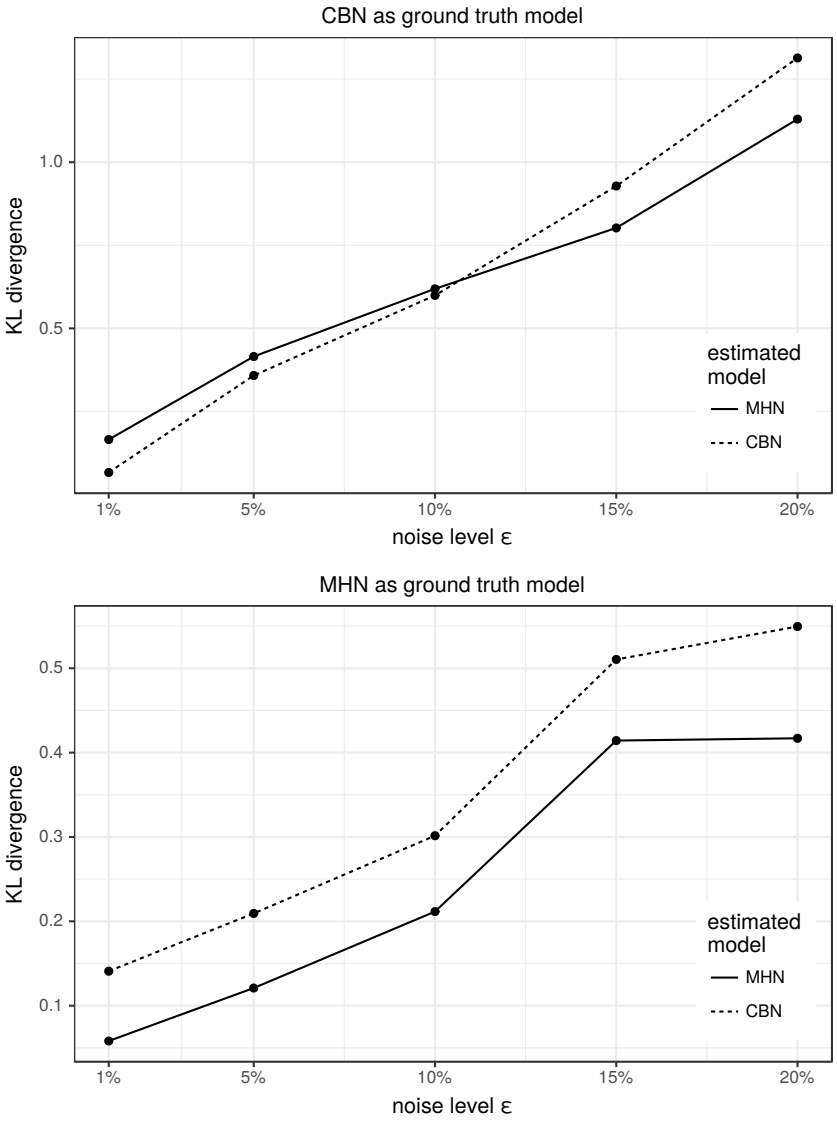



**Figure 3.3.:** Average KL divergence of estimated MHNs/CBNs to ground truth MHNs/CBNs over 5 simulations for each of the shown noise levels.

# 4. Analysis of inferred MHNs

*In this chapter we train Mutual Hazard Networks on biological data. We compare the resulting models on three CGH datasets to Conjunctive Bayesian Networks, and on one bulk sequencing dataset to pathTiMEx. We discuss differences in terms of model fit and biological plausibility.*

## 4.1. Breast cancer, Colorectal cancer, Renal cell carcinoma

We tested our method and first compared it to Conjunctive Bayesian Networks (CBN) on three cancer datasets that were previously used by [33]. They were obtained from the Progenetix molecular-cytogenetic database [6] and consist of 817 breast cancers, 570 colorectal cancers, and 251 renal cell carcinomas. The cancers are characterized by 10, 11, and 12 recurrent copy number alterations, respectively, which were detected by comparative genomic hybridization (CGH).

We trained MHNs on all three datasets and compared them to the CBNs given in [33] which provide log-likelihood scores in-sample. The in-sample scores of MHNs are not directly comparable because MHNs have more degrees of freedom than CBNs. Therefore we additionally provide the average log-likelihood scores of MHNs in 5-fold cross-validation and the Akaike Information Criterion [1] (AIC) for both models. MHN compared favourably on all three datasets (Table 4.1).

In the plots below (A) shows the raw data, where rows are copy number alterations and are sorted by frequency, while columns are event tumors whose 0/1-patterns are sorted lexicographically. (B) shows the CBN estimated from this data, where edges denote that all parent alterations must have occured before the child alteration can. Afterwards this alteration happens with the rate annotated in the corresponding node. (C) shows the MHN estimated from this data, where alterations initially happen with rate annotated in the corresponding node. Once an alteration has occured, it multiplies the rate of other events by the factor annotated on the edges.

Note that the MHNs show mutual dependencies as well as inhibiting edges, features that a CBN does not have. Our observation that MHNs compare favorably in terms

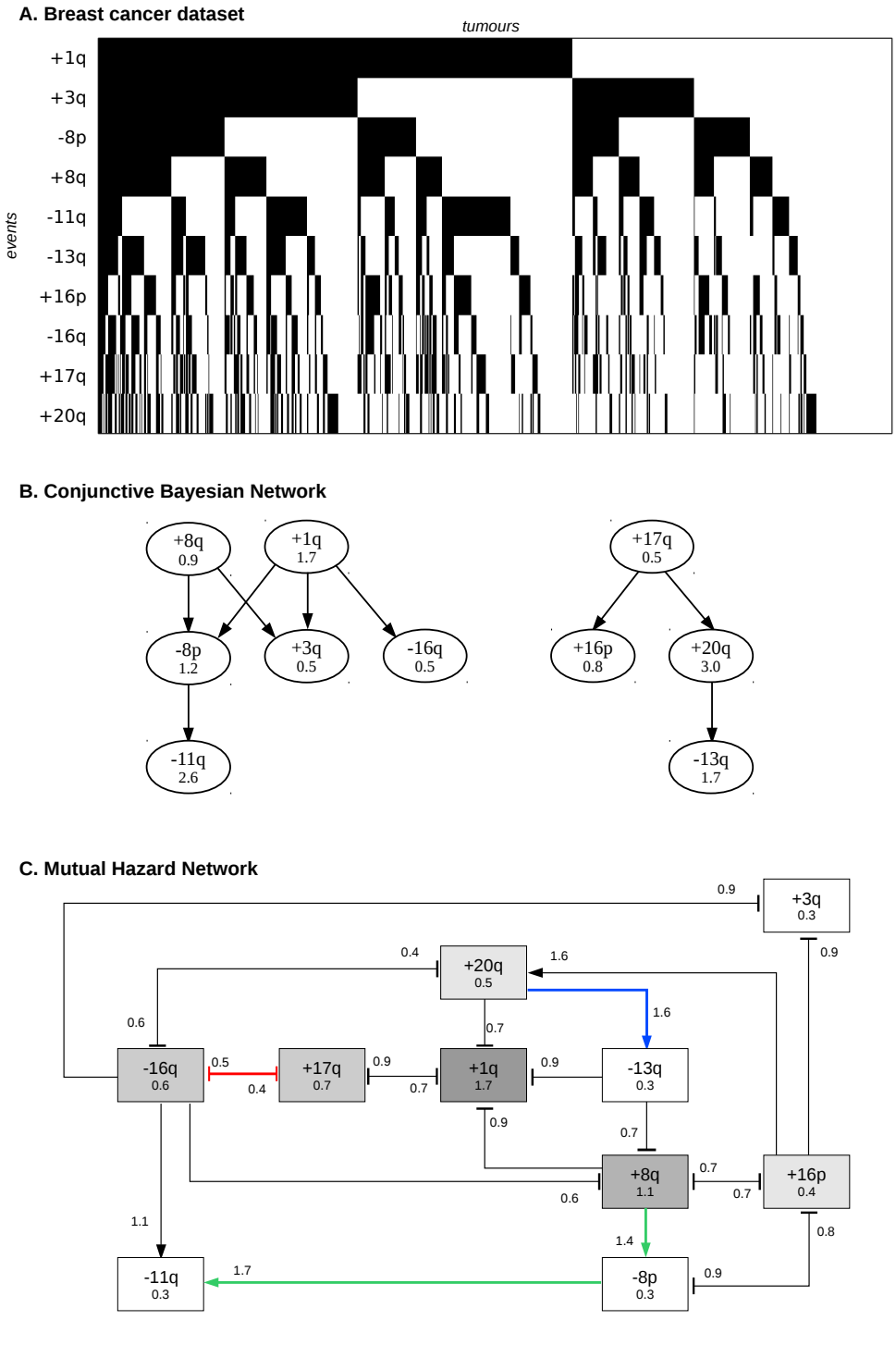



**Figure 4.1.:** Breast cancer models

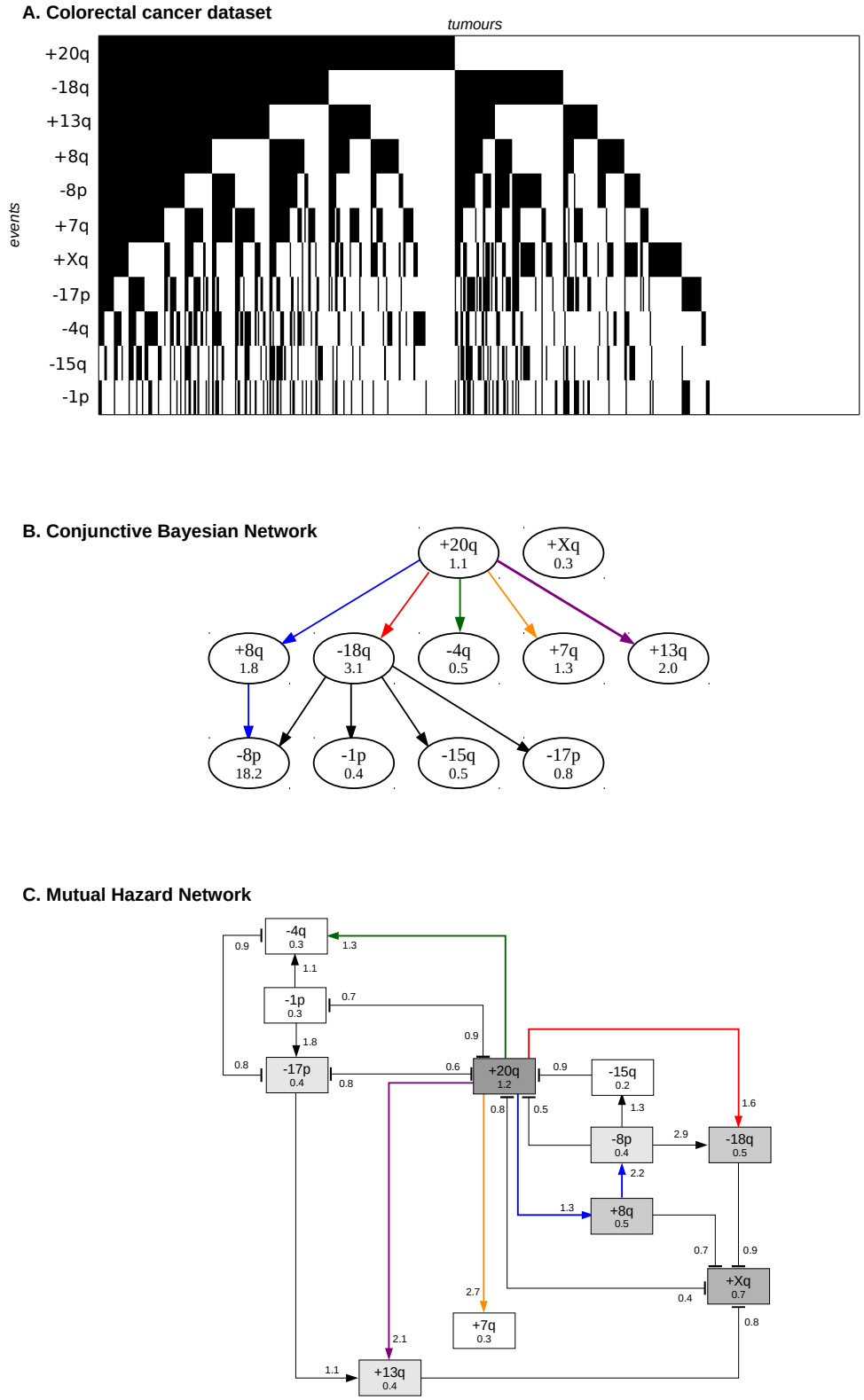



**Figure 4.2.:** Colorectal cancer models

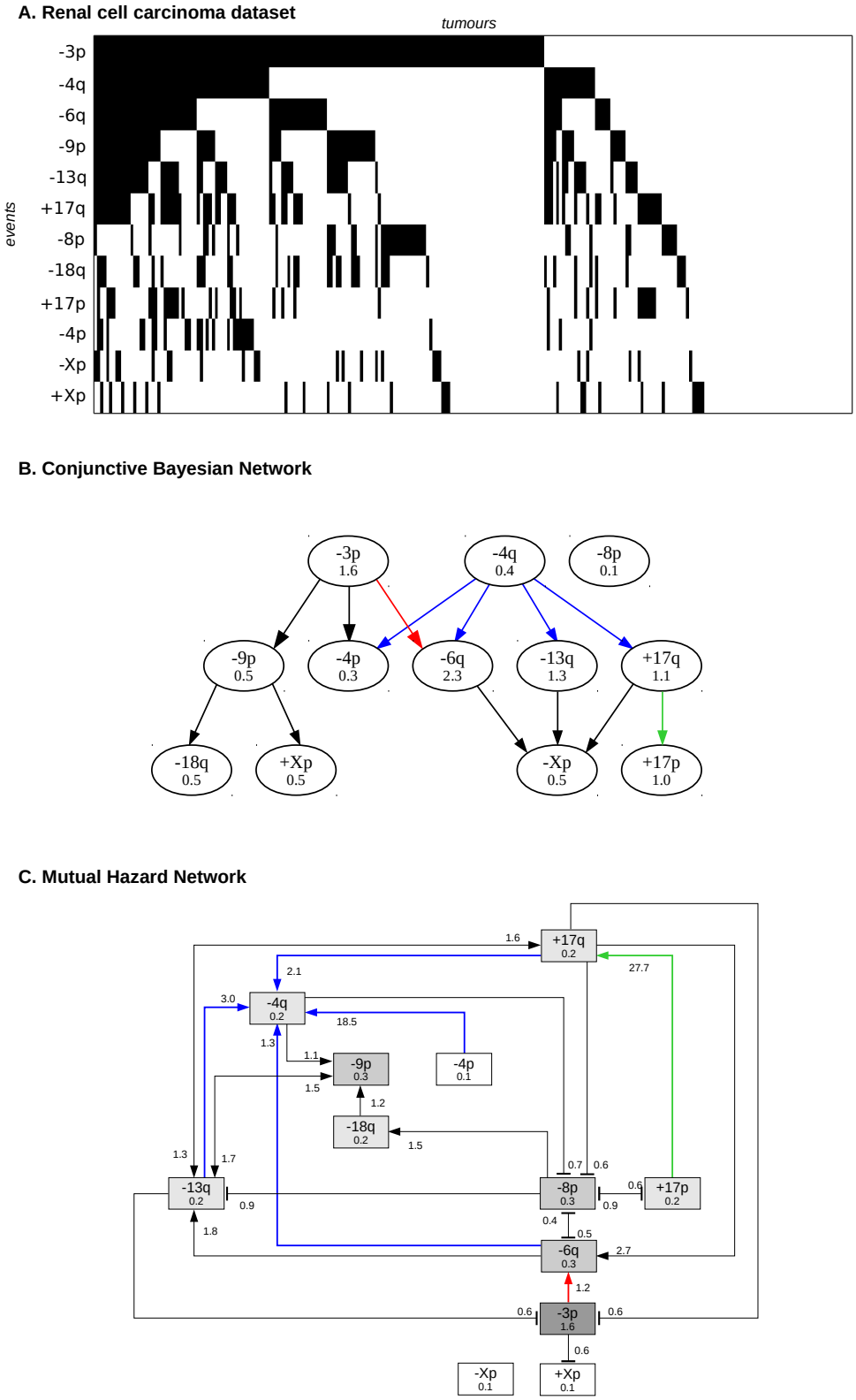



**Figure 4.3.:** Renal cell carcinoma models

of both cv-log-likelihood and AIC can be interpreted as evidence that such mutual dependencies between progression events exist in all three types of cancer. The models show striking overlap as well as marked differences.

**Table 4.1.:** MHNs compare favourably to CBNs on three datasets in terms of average log-likelihood scores in 5-fold cross-validation and in terms of the AIC which penalizes the number of parameters in a model. While MHNs have $n^2$ continuous parameters, CBNs have $n$ continuous parameters and a discrete graph structure that is hard to quantify in terms of degrees of freedom, hence we ignore the latter and bound the AIC of CBNs from below.

| | cross-validated | in-sample | | AIC | |
|---|---|---|---|---|---|
| dataset | MHN | CBN | MHN | CBN | MHN |
| Breast cancer | -5.63 | -5.73 | -5.54 | ≥9373 | 9152 |
| Colorectal cancer | -5.64 | -5.79 | -5.41 | ≥6612 | 6288 |
| Renal cell carcinoma | -5.02 | -5.13 | -4.81 | ≥2587 | 2559 |

**Breast cancer:** Both models have the edges +8q → -8p → -11q in common (green). The tumor initiating role of +8q may arise from the oncogene MYC on chromosome arm 8q. The two models also agree in the edge +20q → -13q (blue), where the oncogene AURKA and the tumor suppressor BRCA2 are located.

However, MHN identifies mutual exclusivity between +17q and -16q (red) which the CBN cannot. Interestingly, gains at 17q, the locus of the oncogene ERBB2, are associated with a poor prognosis [14] while losses at 16q are associated with a good prognosis [62]. Moreover, in the CBN the event +1q predisposes cancers to a subsequent 16q loss. The MHN model agrees that the two events are related but interprets their interplay differently: Here +17q inhibits both +1q and -16q and their association can be explained away by the absence of +17q. Hence the MHN does not see a driver event in +1q which facilitates a subsequent 16q loss and thus a favorable course of progression.

**Colorectal cancer:** Both models agree that +20p is an initiating event and promotes -18q (red), -4q (green), +7q (orange), and +13q (purple), likely due to genetic instability caused by the oncogene AURKA [11]. They also agree in the edges +20q → +8q → -8p. CBN further identifies a subsequent major event in the loss of 18q which is the locus of the tumor suppressor SMAD4 and appears to trigger -1p, -15q, -17p and -8p. In contrast, MHN identifies -8p as an intermediate rather than terminal event which promotes both -18q and -15q, thereby explaining away their positive association. Compared to the CBN this interpretation avoids assigning the exceptionally

large rate of 18.2 to the event -8p, which would mean that -8p occurs immediately after its parent.

**Renal cell carcinoma:** Both models identify -3p as an initiating event which is the locus of VHL, a tumor suppressor which regulates the hypoxia response pathway [21] and plays a known initiating role in RCC [36]. The models also agree that -3p promotes -6p. While both models find that -4q is related to -4p, -6q, -13q and +17q, these edges point away from -4q in the CBN and point towards -4q in the MHN. Similarly, CBN finds that +17q promotes +17p, while MHN finds that +17p promotes +17q.

## 4.2. Glioblastoma

Next, we compared MHN to pathTiMEx on a glioblastoma dataset from The Cancer Genome Atlas [18] which was previously used in [23]. The data consist of $|\mathcal{D}| = 261$ tumors characterized by 486 point mutations (M), amplifications (A), or deletions (D). We focus on $n = 20$ of these events which were pre-selected by pathTiMEx using the TiMEx algorithm [20].

We trained MHN as above for 100 iterations, which achieved a log-likelihood score of -7.70 in-sample and a score of -7.97 in 5-fold cross-validation. While pathTiMEx does not yield a directly comparable log-likelihood score, it quantifies discrepancies between model and data by considering the data to be corrupted by noise, each event in a tumor being independently flipped with probability $\varepsilon$. PathTiMEx estimated this noise parameter as $\hat{\varepsilon} = 20\%$, from which we gauge an upper bound on its log-likelihood score as follows: even a hypothetical model that learns the data distribution $\mathbf{p}_{\mathcal{D}}$ perfectly but assumes a level of noise

$$\mathbf{p}_{\hat{\varepsilon}} = \bigotimes_{i=1}^{n} \begin{pmatrix} 1-\hat{\varepsilon} & \hat{\varepsilon} \\ \hat{\varepsilon} & 1-\hat{\varepsilon} \end{pmatrix} \mathbf{p}_{\mathcal{D}} \tag{4.1}$$

achieves only a score of $\mathbf{p}_{\mathcal{D}}^{T} \log \mathbf{p}_{\hat{\varepsilon}} = -8.50$ in-sample, which is less than the cross-validated score of MHN.

The results are shown in Fig. 4.4 as follows:

A. Raw dataset, where rows show events sorted by frequency and columns show tumors sorted lexicographically. The purple stripes highlight tumors which have IDH1(M) but lack TP53(M).

B. PathTiMEx model inferred in [23]. It simultaneously divides the dataset into pathways, i.e., into mutually exclusive groups of events and learns a CBN of these pathways. The CBN considers a pathway altered if at least one of its

constituent events has occured. A pathway alteration fixates at the rate given in the upper right-hand corner once all its parent pathways in the CBN have been altered.

C. Highlighted discrepancies between the data and the pathTiMEx model due to its assumption of interchangeable events. Although *CDKN2A(D)* and *CDK4(A)* were grouped into the same pathway, *CDKN2A(D)* is negatively associated with *MDM2(A)* in the data while *CDK4(A)* is positively associated with it.

D. Mutual Hazard Network, where nodes show the base rates $\Theta_{ii}$ and edges show the multiplicative interactions $\Theta_{ij}$. Similarities to pathTiMEx are highlighted in colour and roughly correspond to the signaling pathways Rb, p53, and PI(3)K (red, blue, and green).

MHN largely agreed with pathTiMEx on the inhibitions implied by the three most mutally exclusive groups of events, which broadly correspond to the signaling pathways Rb, p53, and PI(3)K (red, blue, and green in Fig. 4.4) and are well known to be affected in glioblastoma [52].

The RB1 signaling pathway (red) regulates cell cycle progression and involves the genes *CDKN2A*, *CDK4* and *RB1*. *CDKN2A* codes for the tumor suppressor protein $p16^{INK4a}$ which binds to CDK4 and prevents it from phosphorylizing RB1, thereby blocking cell cycle transition from G1 to S-phase. This function can be disrupted by deletion of *CDKN2A* or *RB1*, or by amplification of *CDK4*. MHN and pathTiMEx both report a corresponding inhibition between the events *CDKN2A(D)* and *CDK4(A)*, while MHN additionally reports inhibition between *CDKN2A(D)* and *RB1(D)*.

The p53 signaling pathway (blue) induces apoptosis in response to stress signals and involves the genes *TP53*, *MDM2*, *MDM4* and *CDKN2A*. *TP53* codes for the tumor suppressor protein p53 which is antagonized by MDM2 and MDM4 in a non-redundant manner [71]. PathTiMEx identifies the events *TP53(M)*, *MDM2(A)*, *MDM4(A)* as mutually exclusive ways to evade apoptosis, while MHN reports inhibition only between *TP53(M)* and each of *MDM2(A)* and *MDM4(A)* separately. This may reflect non-diminishing returns due to their complementary roles in the pathway.

The gene *CDKN2A* is, in addition to its role in the RB1 pathway, also involved in the p53 pathway by coding for the protein $p14^{ARF}$ in an alternate reading frame. $p14^{ARF}$ physiologically inhibits MDM2, which suggests that a deletion of *CDKN2A* may be functionally similar to an amplification of *MDM2*. While MHN reports a corresponding inhibition between *CDKN2A(D)* and *MDM2(A)*, pathTiMEx cannot because this would lead to an overlap of pathways.

To the contrary, pathTiMEx implies that *CDKN2A(D)* promotes *MDM2(A)* despite their anti-correlation in the data (Fig. 6C). We argue that this is an artifact driven by the assumption that all events in a group are interchangeable, and by the need to

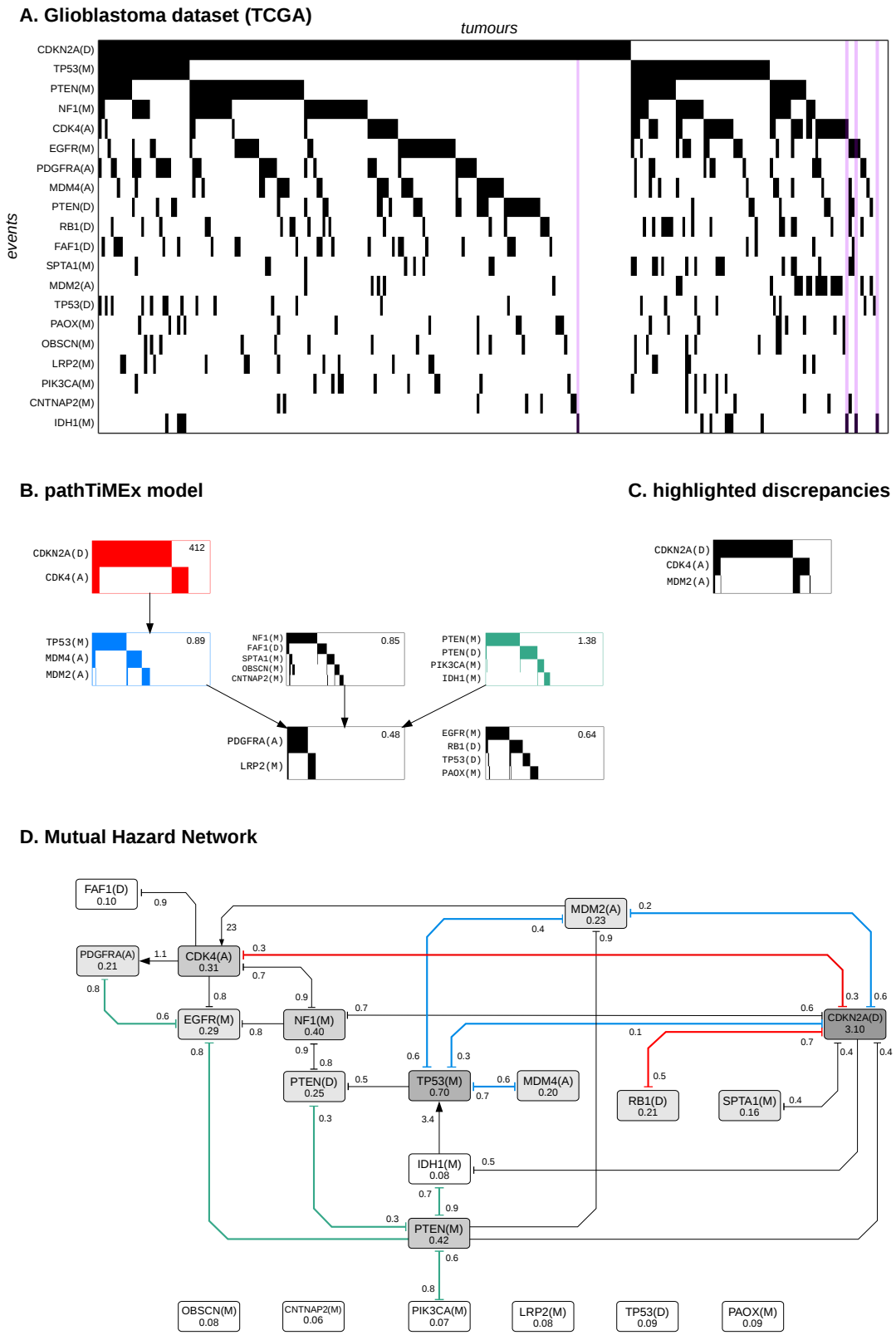



**Figure 4.4.:** Glioblastoma models

group *CDKN2A(D)* with *CDK4(A)* which is in turn highly correlated with *MDM2(A)*.

The PI(3)K pathway (green) regulates cell proliferation and involves the genes *PTEN*, *PIK3CA*, *EGFR*, *PDGFRA*. While IDH1 is not a canonical member of the PI(3)K pathway, MHN reports an inhibition between IDH1(M) and PTEN(M) and path-TiMEx groups *IDH1(M)* together with *PTEN(M)*, *PTEN(D)* and *PIK3CA(M)*. Notably, MHN inferred that the rare event *IDH1(M)* promotes the more common event *TP53(M)*. This is further illustrated in Fig. 4.5 which shows the most likely chronological order of events for all 261 tumors. Each of their 193 distinct states is represented by a path that starts at the root node and terminates at either a leaf node or an internal node with a black outline. As can be seen in the lower left, all tumors that contain *IDH1(M)* are located on a common branch and thus share an early mutation history initiated by *IDH1(M)*. This interpretation is in line with the fact that *IDH1(M)* is considered a defining attribute of the Proneural subtype of gliobastoma which is clinically distinct and also associated with *TP53(M)* [73]. It is further supported by independent data from consecutive biopsies of gliomas where *IDH1(M)* in fact preceded *TP53(M)* [76].

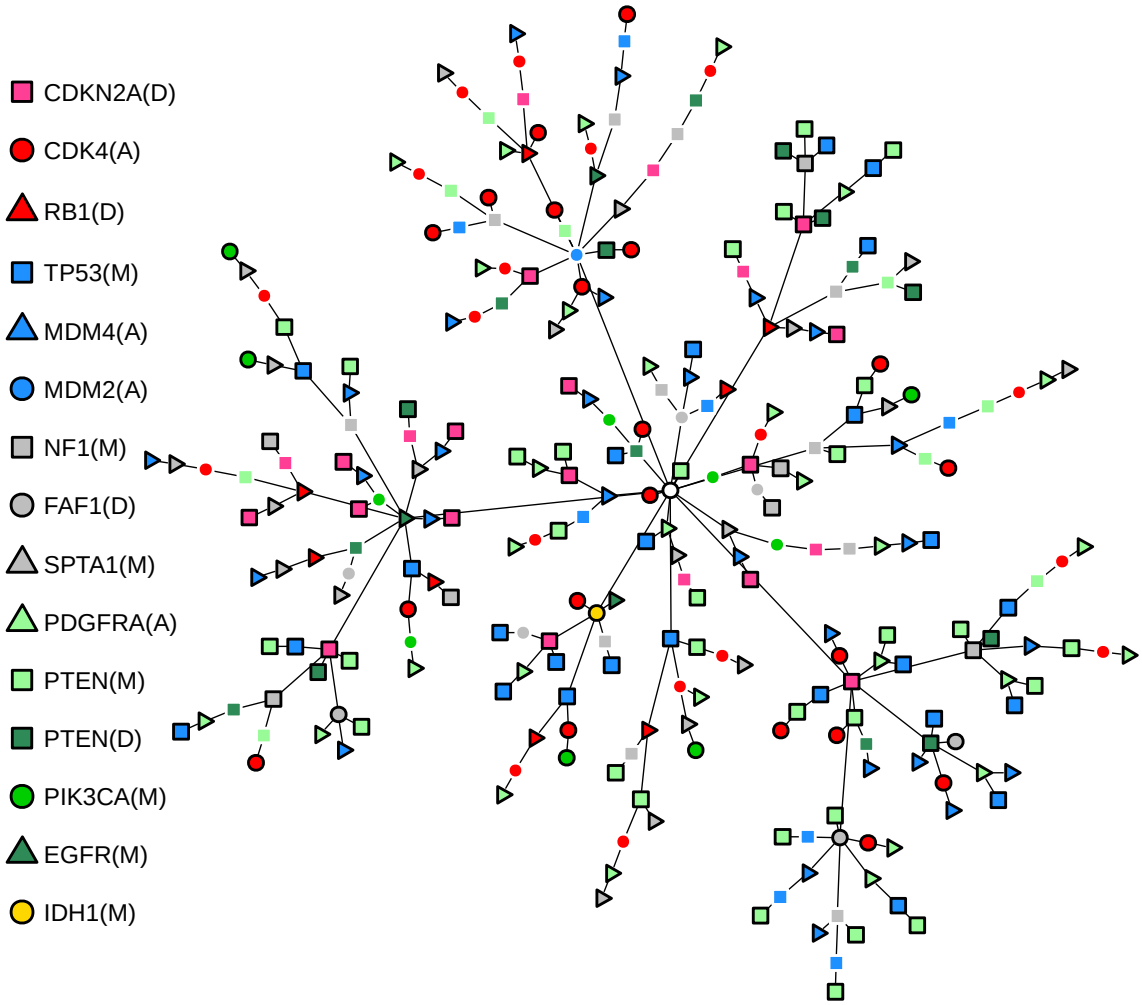



**Figure 4.5.:** Maximum likelihood paths through the state space $S$ from the starting state to each observed tumor state. They were computed from the time-discretized transition rate matrix $I + Q_{\hat{\Theta}}/\gamma$, where $\gamma$ is the greatest absolute diagonal entry of $Q_{\hat{\Theta}}$.

# 5. Computing the transient distribution and its derivative

*In this chapter we introduce the differentiated uniformization method, an algorithm for computing the transient distribution and its derivative of a continuous-time Markov chain on a large, combinatorial state space. We consider Markov chains whose transition rate matrix can be written as a sum of tensor products and exploit this structure to efficiently perform the matrix exponential and its derivative. Initially we developed this algorithm as a crucial building block for extensions of Mutual Hazard Networks (see chapter 6). However, due to its importance we decided to publish the algorithm itself in a self-contained paper with focus on a simpler application, which formed this chapter. Since most of this work was carried out during the COVID-19 pandemic, our choice naturally fell on stochastic modeling of epidemic spread.*

## 5.1. Application to epidemic spread

Predicting the time evolution of complex dynamical systems has a wide range of applications in medicine and public health. One of them is the SIR model of epidemic spread, which describes a population by the numbers of susceptible ($S$), infected ($I$) and recovered ($R$) people. Until recently the SIR model has been approximated deterministically [46] and was considered computationally intractable in its stochastic formulation [51]. The stochastic SIR model is a continuous-time Markov chain (CTMC) in which infections happen randomly with a rate proportional to $S$ and proportional to $I$ [2]. The state of the system at a given time is thus fully specified by the combination of $S$ and $I$. Since infections happen randomly one must keep track of a huge number of probabilities, one for every possible state. For example, the Austrian population of 9 million people can go through roughly 9 million × 9 million = 81 trillion possible states during the course of an epidemic.

More generally, we consider CTMCs that describe the evolution of a transient probability distribution $\mathbf{p}(t)$ over a huge discrete state space according to the Kolmogorov forward equation

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = Q\mathbf{p}(t) \quad \text{with solution} \quad \mathbf{p}(t) = \exp(tQ)\mathbf{p}(0). \tag{5.1}$$

Here $Q$ is the transition-rate matrix and $\exp(tQ)$ is the matrix exponential. For an SIR model of the Austrian population $Q$ has 81 trillion $\times$ 81 trillion entries, and naively computing the matrix exponential requires on the order of 81 trillion $\times$ 81 trillion $\times$ 81 trillion operations [54], which is practically impossible.

Even more dauntingly, when $Q$ depends on an unknown parameter $\theta$, such as the infection or recovery rate in the SIR model, we must first infer $\theta$ from data by maximizing its likelihood or by sampling from its posterior in a full Bayesian analysis. This typically requires the derivative of the matrix exponential $\partial \exp(tQ)\mathbf{p}(0)/\partial\theta$ in order to compute $\partial\mathbf{p}(t)/\partial\theta$. However, Ho et al. [45] have recently provided an algorithm that solves the Kolmogorov equation in the Laplace domain and evaluates the inverse Laplace transform numerically, thus avoiding the matrix exponential. Their algorithm is applicable to systems where each discrete variable increases monotonically. This includes the SIR model,[1] for which their algorithm scales quadratically in the population size.

Here, we provide an alternative algorithm that directly computes $\exp(tQ)$ and $\partial \exp(tQ)/\partial\theta$. For the SIR model it scales cubically in the population size but is still practical. Importantly, our approach is applicable to a broader class of CTMCs with large state spaces that arise from interacting discrete variables, without requiring monotonicity. For example, in tumor progression models the states are combinations of possible mutations ([7], [66]), in stochastic neural networks the states are activation patterns of neurons [79], in predator-prey dynamics they are joint population sizes of interacting species [58], or in chemical reaction networks they are joint counts of chemical species [78].

For many of these models $Q$ can be written as a sum of tensor products [12]. We provide such a representation for the stochastic SIR model. To the best of our knowledge, this representation is novel. We use it for matrix-vector products that do not require explicit storage of $Q$ [15] and make computation of the matrix exponential tractable via the uniformization method [41]. A similar approach by Sherlock [68] exploits the sparsity of $Q$. We extend the uniformization method and provide an analogous algorithm that also computes the derivative of the matrix exponential. Finally, we use Hamiltonian Monte Carlo sampling to provide a full Bayesian analysis of the first wave of the COVID-19 pandemic for the Austrian population, shedding new light on the uncertainties associated with the estimation of infection and recovery rates.

[1] By changing variables from susceptibles and infected to infections and recoveries.

## 5.2. Differentiated Uniformization

A discrete-state, continuous-time Markov chain (CTMC) describes probability distributions $\mathbf{p}(t) \in \mathbb{R}^{|X|}$ over a state space $X$, where an entry $\mathbf{p}(t)_x$ denotes the probability that the CTMC is in state $x \in X$ at time $t \in [0, \infty)$. Its change over time is governed by the Kolmogorov forward equation

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = Q\mathbf{p}(t) \tag{5.2}$$

with transition-rate matrix $Q \in \mathbb{R}^{|X|\times|X|}$, where an off-diagonal entry $Q_{y,x}$ is the transition rate from state $x \in X$ to state $y \in X$ and diagonal entries are set such that columns sum to zero.

The solution to eq. (5.2) is given by the matrix exponential

$$\mathbf{p}(t) = \exp(tQ)\mathbf{p}(0) = \sum_{n=0}^{\infty} \frac{t^n}{n!} Q^n \mathbf{p}(0), \tag{5.3}$$

which could be approximated in principle by terminating after a finite number of terms. However, catastrophic cancellations occur [54] due to the fact that $Q$ has negative entries and negative eigenvalues.[2] The uniformization method [41] addresses this problem by introducing a strictly nonnegative matrix

$$P := \frac{1}{\gamma} Q + \mathrm{I} \quad \text{for some } \gamma \geq \max_x |Q_{x,x}| \tag{5.4}$$

such that

$$\begin{aligned} \mathbf{p}(t) = \exp(tQ)\mathbf{p}(0) &= \exp(\gamma t(-\mathrm{I} + P))\mathbf{p}(0) \\ &= \exp(-\gamma t\,\mathrm{I})\exp(\gamma t P)\mathbf{p}(0) \\ &= \sum_{n=0}^{\infty} e^{-\gamma t} \frac{(\gamma t)^n}{n!} P^n \mathbf{p}(0) \end{aligned} \tag{5.5}$$

does not suffer from cancellations. $P$ can be viewed as the transition probability matrix of a discrete-time Markov chain where the number of transitions is a Poisson-distributed random variable with mean $\gamma t$.

Using the recursions

$$P^n = PP^{n-1}, \tag{5.6}$$

$$\frac{(\gamma t)^n}{n!} = \frac{\gamma t}{n} \frac{(\gamma t)^{n-1}}{(n-1)!}, \tag{5.7}$$

$\mathbf{p}(t)$ can be computed according to eq. (5.5) by algorithm 1 [41].

**Algorithm 1:** Uniformization

**input :** $\mathbf{p}(0), t, P, \gamma, \varepsilon$
**output:** $\mathbf{p}(t)$

$n \leftarrow 0$
$w \leftarrow 1$
$\mathbf{p}(t) \leftarrow \mathbf{0}$
$\mathbf{q} \leftarrow \mathbf{p}(0)$
**repeat**
  $\mathbf{p}(t) \leftarrow \mathbf{p}(t) + e^{-\gamma t} w \mathbf{q}$
  $n \leftarrow n + 1$
  $\mathbf{q} \leftarrow P\mathbf{q}$
  $w \leftarrow \frac{\gamma t}{n} w$
**until** $1 - |\mathbf{p}(t)|_1 < \varepsilon$;
**return** $\mathbf{p}(t)$

**Algorithm 2:** Differentiated Uniformization

**input :** $\mathbf{p}(0), t, P, P', \gamma, \gamma', \varepsilon$
**output:** $\mathbf{p}(t), \mathbf{p}(t)'$

$n \leftarrow 0$
$w \leftarrow 1$
$\mathbf{p}(t) \leftarrow \mathbf{0}$
$\mathbf{p}(t)' \leftarrow \mathbf{0}$
$\mathbf{q} \leftarrow \mathbf{p}(0)$
$\mathbf{q}' \leftarrow \mathbf{0}$
**repeat**
  $\mathbf{p}(t) \leftarrow \mathbf{p}(t) + e^{-\gamma t} w \mathbf{q}$
  $\mathbf{p}(t)' \leftarrow \mathbf{p}(t)' + e^{-\gamma t} w \left( \mathbf{q}' + \gamma' \left( \frac{n}{\gamma} - t \right) \mathbf{q} \right)$
  $n \leftarrow n + 1$
  $\mathbf{q}' \leftarrow P'\mathbf{q} + P\mathbf{q}'$
  $\mathbf{q} \leftarrow P\mathbf{q}$
  $w \leftarrow \frac{\gamma t}{n} w$
**until** $1 - |\mathbf{p}(t)|_1 < \varepsilon$;
**return** $\mathbf{p}(t), \mathbf{p}(t)'$

Note that $P^n\mathbf{p}(0)$ sums to 1 and hence eq. (5.5) sums to less than 1 when terminated after a finite number of terms. The algorithm stops once this probability mass defect

$$1 - \sum_{n=0}^{m} e^{-\gamma t}\frac{(\gamma t)^n}{n!} \tag{5.8}$$

is smaller than a preset tolerance $\varepsilon$. The required number $m$ of iterations is in $\mathcal{O}(\gamma)$ [61] and can be determined, e.g., using the numerically robust method by Sherlock [68].

In this thesis we are interested in statistical models where $Q$ depends on a parameter $\theta$ that we want to estimate from data by maximizing its likelihood or by sampling from its posterior. To this end, we propose a novel algorithm for computing the derivative

$$\begin{aligned}
\frac{\partial \mathbf{p}(t)}{\partial \theta} &= \frac{\partial \exp(tQ)\mathbf{p}(0)}{\partial \theta} \\
&= \frac{\partial}{\partial \theta}\left(\sum_{n=0}^{\infty} e^{-\gamma t}\frac{(\gamma t)^n}{n!}P^n\mathbf{p}(0)\right) \\
&= \sum_{n=0}^{\infty} e^{-\gamma t}\frac{(t\gamma)^n}{n!}\frac{\partial P^n}{\partial \theta}\mathbf{p}(0) + e^{-\gamma t}\frac{\partial \gamma}{\partial \theta}\left(-\frac{t^{n+1}\gamma^n}{n!} + \frac{t^n\gamma^{n-1}}{(n-1)!}\right)P^n\mathbf{p}(0) \\
&= \sum_{n=0}^{\infty} e^{-\gamma t}\frac{(t\gamma)^n}{n!}\left(\frac{\partial P^n}{\partial \theta}\mathbf{p}(0) + \frac{\partial \gamma}{\partial \theta}\left(\frac{n}{\gamma} - t\right)P^n\mathbf{p}(0)\right)
\end{aligned} \tag{5.9}$$

building on the uniformization method. We use the recursions (5.6), (5.7) and additionally

$$\begin{aligned}
\frac{\partial P^n}{\partial \theta} &= \frac{\partial P}{\partial \theta}P^{n-1} + P\frac{\partial P}{\partial \theta}P^{n-2} + \ldots + P^{n-2}\frac{\partial P}{\partial \theta}P + P^{n-1}\frac{\partial P}{\partial \theta} \\
&= \frac{\partial P}{\partial \theta}P^{n-1} + P\left(\frac{\partial P}{\partial \theta}P^{n-2} + \ldots + P^{n-3}\frac{\partial P}{\partial \theta}P + P^{n-2}\frac{\partial P}{\partial \theta}\right) \\
&= \frac{\partial P}{\partial \theta}P^{n-1} + P\left(\frac{\partial P^{n-1}}{\partial \theta}\right)
\end{aligned} \tag{5.10}$$

to compute $\mathbf{p}(t)' := \partial\mathbf{p}(t)/\partial\theta$ according to eq. (5.9) by algorithm 2.

Applying differentiated uniformization for a particular statistical model requires the scalars

$$\gamma \geq \max_x |Q_{x,x}| \quad \text{and} \quad \gamma' := \frac{\partial \gamma}{\partial \theta}, \tag{5.11}$$

[2] Negative entries directly lead to cancellations in eq. (5.3). If eq. (5.3) is transformed to the eigenbasis of $Q$, negative eigenvalues of $Q$ lead to cancellations as well.

where a generic choice for $\gamma$ can be the 2-norm of the diagonal of $Q$ or any $p$-norm with even $p$. It also requires the operators

$$P = \frac{1}{\gamma}Q + \mathrm{I} \quad \text{and} \quad P' := \frac{\partial P}{\partial \theta} = -\frac{1}{\gamma^2}\frac{\partial \gamma}{\partial \theta}Q + \frac{1}{\gamma}\frac{\partial Q}{\partial \theta}. \tag{5.12}$$

Crucially, these operators are only needed for matrix-vector products in lines 11 and 12 of algorithm 2 and do not need to be stored explicitly. This makes our method especially useful for models where $Q$ is large but has a compact representation as a sum of tensor products, which allows one to cheaply compute matrix-vector products [15].

Differentiated uniformization thus opens the door to **parameter inference** for CTMCs on huge discrete state spaces. Let $\{x_1, \ldots, x_K\}$ be observations of the Markov chain at corresponding time points $\{t_1, \ldots, t_K\}$. We represent each data point by an empirical probability distribution $\boldsymbol{\delta}(t_k) \in \mathbb{R}^{|X|}$, where $\boldsymbol{\delta}(t_k)_{x_k} = 1$ and all other entries are zero. The likelihood of $\theta$ for a single observation of state $x_k$ at time $t_k$ with $k > 1$ is

$$\mathbf{p}(t_k)_{x_k}, \text{ where } \mathbf{p}(t_k) = \exp\big((t_k - t_{k-1})Q\big)\boldsymbol{\delta}(t_{k-1}). \tag{5.13}$$

The log-likelihood for the whole data set,

$$\ell(\theta) = \sum_{k=2}^{K} \log(\mathbf{p}(t_k)_{x_k}), \tag{5.14}$$

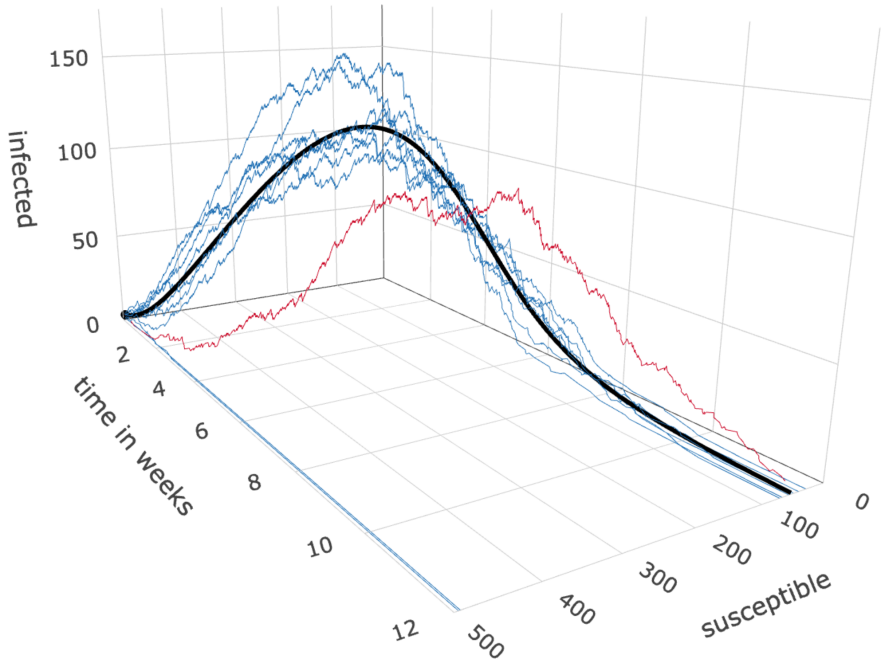
can be maximized using its derivative

$$\frac{\partial \ell(\theta)}{\partial \theta} = \sum_{k=2}^{K} \frac{\mathbf{p}(t_k)'_{x_k}}{\mathbf{p}(t_k)_{x_k}}, \tag{5.15}$$

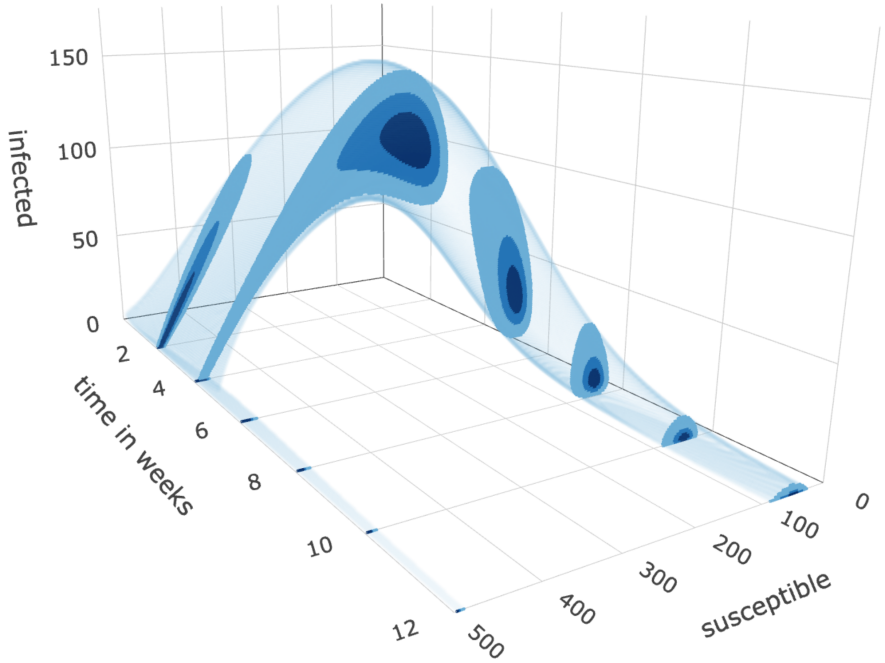
for example by gradient ascent. This derivative can also be used for sampling a posterior distribution of $\theta$ in a full Bayesian model using a Hamiltonian Monte Carlo method [30].

## 5.3. Stochastic SIR models

The most basic models of epidemic spread are SIR models, which describe the numbers of susceptible ($S$), infected/infectious ($I$) and recovered ($R$) people during an epidemic in a closed population of constant size $N$. There is a widely known deterministic and a lesser-known stochastic variant of the SIR model [2]. The latter involves a huge discrete state space and is therefore challenging to use. However, it allows for uncertainty quantification in its dynamics and in inferred parameters.




**(a)** Solution of a deterministic SIR model (black curve) and 10 randomly sampled trajectories of the corresponding stochastic SIR model (blue and red). The trajectory highlighted in red deviates drastically from the deterministic solution.




**(b)** Analytic solution of the Kolmogorov equation for a stochastic SIR model. The time slices show distributions $\mathbf{p}(t)$ where the shades of blue show the smallest (not necessarily contiguous) areas that contain 30%, 60% and 90% of the probability mass at time $t$.

**Figure 5.1.:** Illustration of SIR models with $N = 500$, $\alpha = 1w^{-1}$, $\beta = 2.5w^{-1}$, $I(0) = 3$, $S(0) = 497$.

The **deterministic SIR model** [46] assumes that $S(t), I(t), R(t) \in [0, N]$ are continuous and describes their evolution over time $t \in [0, \infty)$ by the following system of nonlinear ordinary differential equations:

$$\begin{aligned} \frac{\mathrm{d}S(t)}{\mathrm{d}t} &= \overbrace{-\beta \frac{I(t)S(t)}{N}}^{\text{infections}} \quad \overbrace{\qquad}^{\text{recoveries}}, \\ \frac{\mathrm{d}I(t)}{\mathrm{d}t} &= +\beta \frac{I(t)S(t)}{N} \quad - \alpha I(t), \\ \frac{\mathrm{d}R(t)}{\mathrm{d}t} &= \qquad\qquad\quad + \alpha I(t), \end{aligned} \tag{5.16}$$

where $\alpha, \beta \in \mathbb{R}^+$ are parameters. Note that once $S(t)$ and $I(t)$ are given, $R(t) = N - S(t) - I(t)$ is already determined and can be omitted in further analysis.

In words, an infection occurs when a susceptible person comes in sufficiently close contact with an infected person, which happens proportionally to the number of susceptible and to the density of infected people in the population and proportionally to an infection rate $\beta$. This rate $\beta$ encompasses, for example, disease characteristics, people's behavior, public policy and weather. An infected person recovers with rate $\alpha$ and can then no longer become susceptible or infected again. The basic reproduction number $\mathcal{R}_0 := \beta/\alpha$ is the number of people (in a fully susceptible population) that one infected person infects before recovering.

There is no analytical solution to system (5.16), but it can be solved numerically, for example by Euler's method:

$$\begin{aligned} S(t + \Delta t) &= S(t) - \beta \frac{S(t)I(t)}{N} \Delta t, \\ I(t + \Delta t) &= I(t) + \beta \frac{S(t)I(t)}{N} \Delta t - \alpha I(t) \Delta t. \end{aligned} \tag{5.17}$$

The black curve in Figure 5.1a illustrates this solution for given parameters $\alpha = 1w^{-1}$, $\beta = 2.5w^{-1}$ and initial conditions $N = 500$, $I(0) = 3$, $S(0) = 497$.

This model has several limitations. First, an epidemic is in fact a stochastic process and not a deterministic dynamical system. Second, without modeling the stochastics explicitly it is not possible to quantify the uncertainties of inferred parameters, which contributes to the uncertainties in the course of the epidemic.

The **stochastic SIR model** [51, 2] is a continuous-time Markov chain over all possible states of the population. A state is a pair of integers $(S, I) \in \{0, \dots, N\} \times \{0, \dots, N\}$ denoting the number of susceptible and infected people. States with $S + I > N$ are unreachable but still accounted for in the model.[3]

[3] This is necessary for the tensor representation in eq. (5.21).

Let $\mathbf{p}(t) \in \mathbb{R}^{(N+1)^2}$ denote the probability distribution at time $t$ over all states $(S, I)$. That is, $\mathbf{p}(t)_{(S,I)}$ is the probability that at time $t$ there are $S$ susceptible and $I$ infected people. Its time evolution is governed by the Kolmogorov forward equation

$$\frac{\mathrm{d}\mathbf{p}(t)}{\mathrm{d}t} = Q\mathbf{p}(t), \tag{5.18}$$

where the matrix $Q \in \mathbb{R}^{(N+1)^2 \times (N+1)^2}$ contains the transition rates from a state $(S, I)$ to a state $(S + \Delta S, I + \Delta I)$:

$$Q_{(S+\Delta S, I+\Delta I),(S,I)} = \begin{cases} \beta \frac{SI}{N} & \text{if } \Delta S = -1, \Delta I = +1, \\ \alpha I & \text{if } \Delta S = 0, \Delta I = -1, \\ -\beta \frac{SI}{N} - \alpha I & \text{if } \Delta S = 0, \Delta I = 0, S \neq 0, I \neq N, \\ -\alpha I & \text{if } \Delta S = 0, \Delta I = 0, S = 0 \text{ or } I = N, \\ 0 & \text{otherwise.} \end{cases} \tag{5.19}$$

The blue and red curves in Figure 5.1a depict 10 randomly sampled trajectories where transitions happen according to the rates in eq. (5.19), generated by the Gillespie [35] algorithm. The trajectory highlighted in red shows how stochastic fluctuations, especially at the beginning of the epidemic, can drastically alter the shape of the curve compared to its deterministic counterpart. Figure 5.1b shows the analytic solution to eq. (5.18) and further illustrates that the stochasticity is not merely additive noise around the deterministic solution. In particular, the stochastic SIR model allows for a bifurcation where the epidemic dies out in the beginning with nonzero probability.

The parameters $\alpha, \beta \in \mathbb{R}^+$ can be inferred from data using differentiated uniformization. This requires multiple matrix-vector products with $Q$ which is, however, too large to be stored explicitly, even for populations of only thousands of people. Hence, we propose a novel representation of $Q$ that does not require explicit storage. To this end, we introduce band matrices of size $(N + 1) \times (N + 1)$:

$$\begin{aligned} \mathcal{S}^+_{\text{inf}} &= \text{superdiag}(1, \ldots, N), & \mathcal{I}^+_{\text{inf}} &= \text{subdiag}(0, \ldots, N-1), \\ \mathcal{S}^-_{\text{inf}} &= \text{diag}(0, \ldots, N), & \mathcal{I}^-_{\text{inf}} &= \text{diag}(0, \ldots, N-1, 0), \\ \mathcal{S}^+_{\text{rec}} &= \text{diag}(1, 1, \ldots, 1) = \mathrm{I}, & \mathcal{I}^+_{\text{rec}} &= \text{superdiag}(1, \ldots, N), \\ \mathcal{S}^-_{\text{rec}} &= \text{diag}(1, 1, \ldots, 1) = \mathrm{I}, & \mathcal{I}^-_{\text{rec}} &= \text{diag}(0, \ldots, N). \end{aligned} \tag{5.20}$$

This yields a representation of the transition-rate matrix

$$Q = \frac{\beta}{N}(\mathcal{S}^+_{\text{inf}} \otimes \mathcal{I}^+_{\text{inf}}) + \alpha(\mathcal{S}^+_{\text{rec}} \otimes \mathcal{I}^+_{\text{rec}}) - \frac{\beta}{N}(\mathcal{S}^-_{\text{inf}} \otimes \mathcal{I}^-_{\text{inf}}) - \alpha(\mathcal{S}^-_{\text{rec}} \otimes \mathcal{I}^-_{\text{rec}}) \tag{5.21}$$

as a sum of tensor products[4] (see Figure 5.2 for an illustrated explanation). Note that eq. (5.21) is not an approximation but an exact reformulation of eq. (5.19). The benefit

[4] For clarity, we did not factor out $\mathcal{S}^+_{\text{rec}} = \mathcal{S}^-_{\text{rec}} = \mathrm{I}$, which would allow for a representation with only three terms.

of this representation is that its storage complexity is $\mathcal{O}(N)$ rather than $\mathcal{O}(N^4)$ and that performing matrix-vector products has a complexity in only $\mathcal{O}(N^2)$ [15] rather than $\mathcal{O}(N^4)$.

Additionally, differentiated uniformization requires the derivative $\partial Q/\partial\theta$. Here we perform inference with respect to logarithmic parameters $\theta = (\log\alpha, \log\beta)$ in order to ensure the positivity constraint on $\alpha$ and $\beta$:

$$\frac{\partial Q}{\partial \log\alpha} = \alpha(\mathcal{S}^+_{\text{rec}} \otimes \mathcal{I}^+_{\text{rec}}) - \alpha(\mathcal{S}^-_{\text{rec}} \otimes \mathcal{I}^-_{\text{rec}}), \tag{5.22}$$

$$\frac{\partial Q}{\partial \log\beta} = \frac{\beta}{N}(\mathcal{S}^+_{\text{inf}} \otimes \mathcal{I}^+_{\text{inf}}) - \frac{\beta}{N}(\mathcal{S}^-_{\text{inf}} \otimes \mathcal{I}^-_{\text{inf}}). \tag{5.23}$$

Finally, differentiated uniformization requires a differentiable upper bound $\gamma$ on the absolute diagonal entries of $Q$. For the SIR model we choose the exact maximum

$$\begin{aligned}\gamma = \max_x |Q_{x,x}| &= \max\left\{|Q_{(N-1,N-1),(N-1,N-1)}|, |Q_{(N,N),(N,N)}|\right\} \\ &= \max\left\{N(N-1)\frac{\beta}{N} + (N-1)\alpha, N\alpha\right\} \\ &= \max\{(N-1)\beta + (N-1)\alpha, \alpha + (N-1)\alpha\} \\ &= (N-1)\alpha + \max\{(N-1)\beta, \alpha\}. \end{aligned} \tag{5.24}$$

It is differentiable[5] for $\alpha \neq (N-1)\beta$ with

$$\frac{\partial\gamma}{\partial\log\alpha} = \begin{cases} N\alpha & \text{if } \alpha > (N-1)\beta, \\ (N-1)\alpha & \text{if } \alpha < (N-1)\beta, \end{cases} \tag{5.25}$$

$$\frac{\partial\gamma}{\partial\log\beta} = \begin{cases} 0 & \text{if } \alpha > (N-1)\beta, \\ (N-1)\beta & \text{if } \alpha < (N-1)\beta. \end{cases} \tag{5.26}$$

Overall, differentiated uniformization performs $\mathcal{O}(\gamma)$ matrix-vector products and thus has a total runtime complexity in $\mathcal{O}(\gamma N^2) = \mathcal{O}(N^3)$ for the SIR model. It requires storage of the result $\mathbf{p}(t)$, which has complexity $\mathcal{O}(N^2)$.

For parameter inference we are typically only interested in the likelihood that an earlier data point $(S, I)$ is followed by a later data point $(S + \Delta S, I + \Delta I)$ after time $t$. Since the number of susceptibles cannot increase ($\Delta S \leq 0$) and the number of recovered cannot decrease ($\Delta R = -\Delta S - \Delta I \geq 0$) along a trajectory, it is sufficient to compute $\mathbf{p}(t)$ and $\mathbf{p}(t)'$ on the restricted state space

$$\{S + \Delta S, \ldots, S\} \times \{I - \Delta R, \ldots, I - \Delta S\},$$

as explained in appendix B. Following Ho et al. [45] we use this state-space restriction to reduce the time complexity of our algorithm to $\mathcal{O}\big((I + |\Delta S|)(\Delta S^2 + |\Delta S|\Delta R)\big)$ and its storage complexity to $\mathcal{O}(\Delta S^2 + |\Delta S|\Delta R)$.

[5] For $\alpha = (N-1)\beta$ a differentiable upper bound for $\max\{(N-1)\beta, \alpha\}$ is $\log(e^{(N-1)\beta} + e^{\alpha})$.

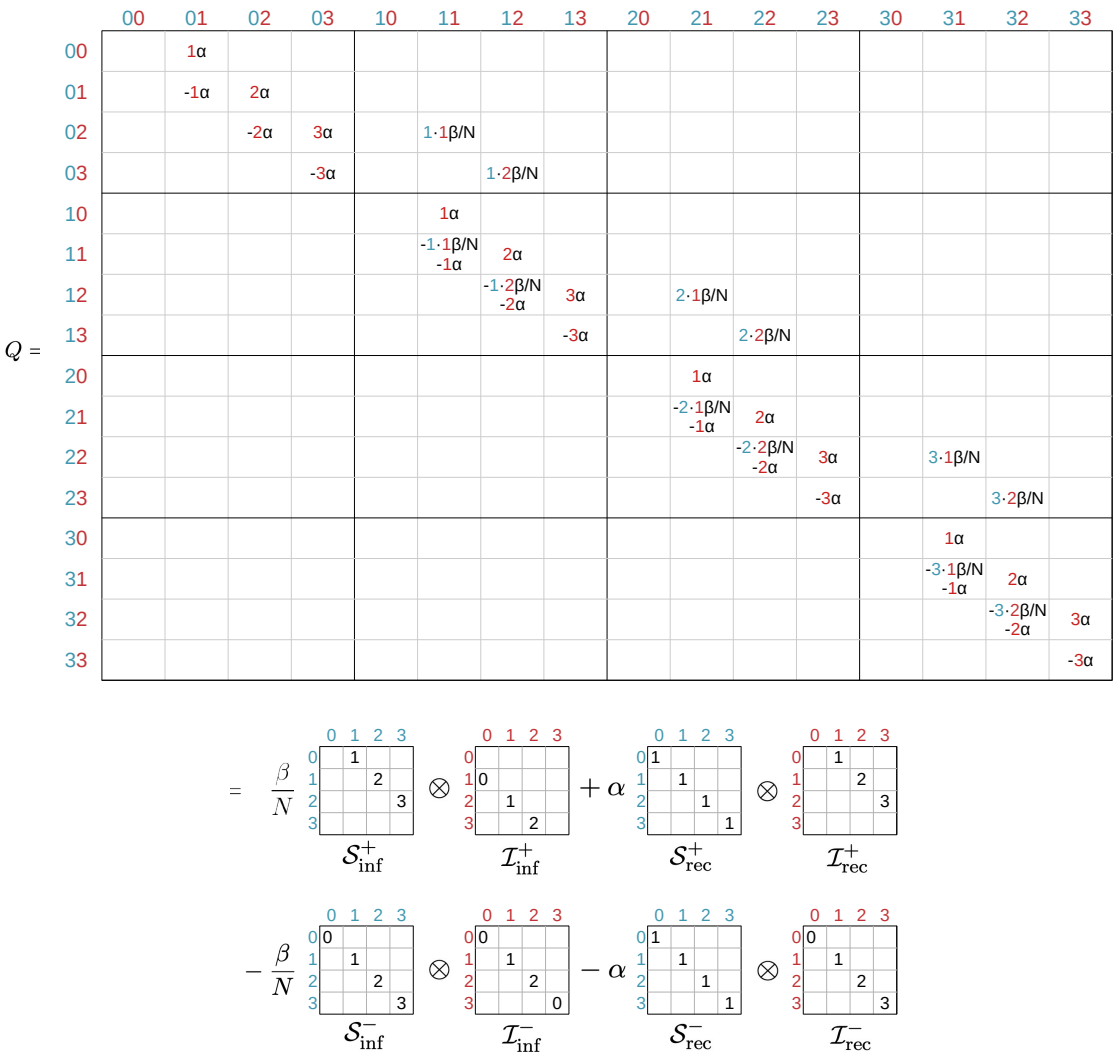



**Figure 5.2.:** Illustration of $Q$ for a population of size $N = 3$ given by its entry-wise representation in eq. (5.19) (top) and its tensor representation in eq. (5.21) (bottom). Blue numbers indicate susceptibles, red numbers indicate infected and blank entries in the matrices are zero. Transitions should be read from columns to rows.
$\mathcal{S}^+_{\text{inf}}$: An infection decreases the number of susceptibles by one and happens proportionally to the current number of susceptibles.
$\mathcal{I}^+_{\text{inf}}$: At the same time, an infection increases the number of infected by one and happens proportionally to the current number of infected. The tensor product $\otimes$ combines both these transitions for a single infection. Moreover, an infection happens inversely proportional to the total population size $N$ and proportionally to the parameter $\beta$.
$\mathcal{S}^+_{\text{rec}}$: A recovery does not change the number of susceptibles.
$\mathcal{I}^+_{\text{rec}}$: At the same time, a recovery decreases the number of infected by one and happens proportionally to the current number of infected. The tensor product $\otimes$ combines both these transitions for a single recovery. Moreover, a recovery happens proportionally to the parameter $\alpha$.
The matrices $\mathcal{S}^-_{\text{inf}}$, $\mathcal{I}^-_{\text{inf}}$, $\mathcal{S}^-_{\text{rec}}$, $\mathcal{I}^-_{\text{rec}}$ generate corresponding negative entries for the diagonal of $Q$.

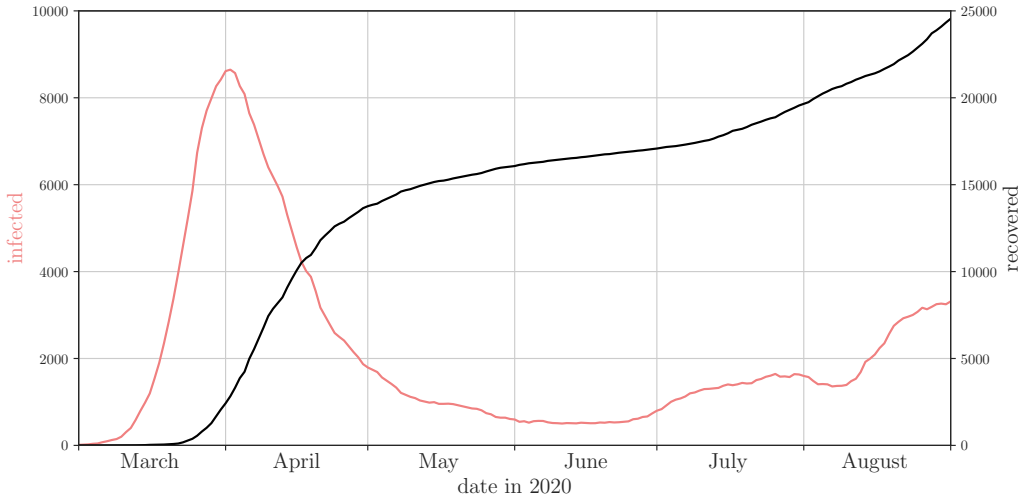



**Figure 5.3.:** Daily reported numbers of people infected by and recovered from SARS-CoV-2 in Austria.

## 5.4. COVID-19 pandemic

Here we model the first wave of the COVID-19 pandemic in Austria as a stochastic SIR model. We employ differentiated uniformization to estimate the parameters $\alpha$ and $\beta$ and quantify their uncertainty. We use daily numbers on $S$, $I$ and $R$ between 2020-03-01 and 2020-09-01 from public health data provided by the Austrian Bundesministerium für Soziales [17] (Figure 5.3). $I$ and $R$ are given directly, and we set $S = N - I - R$ assuming that the initial population size $N = 8{,}932{,}664$ stays constant. People who have died from COVID-19 are counted under "recovered" in a technical sense as they are no longer infectious. We do not correct for undiscovered cases and biases in testing and reporting. We also assume that parameters are piecewise constant for each month.

We do a full Bayesian analysis for parameter pairs $(\log\alpha, \log\beta)$ with a uniform prior. Following Ho et al. [45] we sample from the joint posterior using a Hamiltonian Monte Carlo (HMC) scheme [26, 56]. Unlike a standard Metropolis-Hastings scheme, HMC makes use of the gradient of the likelihood, which we compute using differentiated uniformization. This makes sampling more efficient with less samples needed to cover the posterior distribution [30]. We estimated the joint posterior of $(\log\alpha, \log\beta)$ for every month between March 2020 and August 2020 separately. For each month we performed 10 parallel Monte Carlo chains with length 100, where we discarded the first 10 points each, resulting in 900 points per month. These calculations were done on the QPACE 4 cluster [32] and took about 30 minutes for each of May and June, 2 hours for July, 16 hours for each of April and August and 4 days for March.

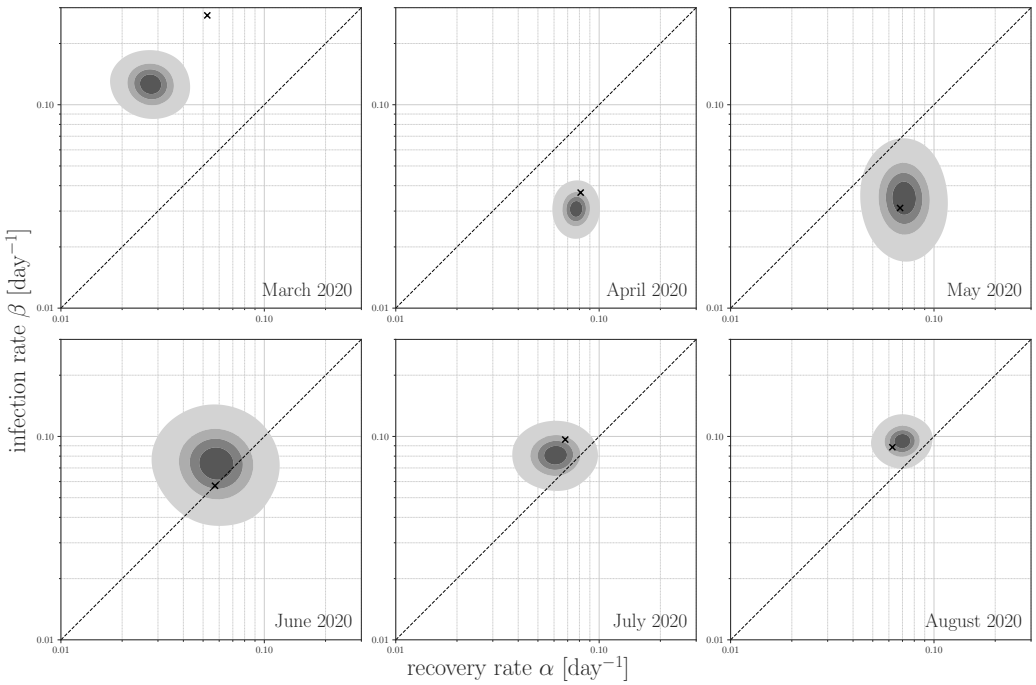



**Figure 5.4.:** Posterior probability densities over parameter pairs $(\alpha, \beta)$ for separate stochastic SIR models of the first six months of the COVID-19 pandemic in Austria. The dashed lines indicate parameters where the basic reproduction number $\mathcal{R}_0 = \beta/\alpha = 1$. The crosses mark the least-squares estimators of the corresponding deterministic SIR models.

Figure 5.4 shows the results of this analysis. The estimated posterior is plotted for $(\alpha, \beta)$ on logarithmic scales. The gray shaded areas were generated using Gaussian-kernel density estimation applied to the posterior samples. The crosses mark the least-squares estimators of the corresponding deterministic SIR models. The dashed lines represent parameter constellations where $\alpha = \beta$ and thus $\mathcal{R}_0 = 1$. Here the epidemic switches between growing and decreasing numbers of infected. From April-August 2020 the posterior of the recovery rate $\alpha$ varies around a value of 0.07 per day, corresponding to the realistic mean time to recovery of about 2 weeks [27]. In contrast, the posterior of $\alpha$ in March 2020 appears to be off, with a mean of about 0.03 per day corresponding to a mean time to recovery of one month. Inspecting the original numbers, we observed that the numbers of recovered are unexpectedly low (less than 100 people until 2020-03-23) possibly due to lagging declaration of recoveries because of cautious hospital policies in the beginning of the pandemic.

Overall, we observe large uncertainties associated with the parameters in several but not all months. These might hint at epidemic courses that are not perfectly in line with an SIR model or with the assumption of piecewise constant parameters. Such deviations from the model assumptions are much less readily apparent in deterministic approaches.

## 5.5. Stochastic predator-prey models

We consider the following deterministic predator-prey equations, based on [50, 75],

$$\frac{\mathrm{d}X(t)}{\mathrm{d}t} = -\beta X(t)Y(t) \overbrace{+ \alpha X(t) - \alpha \frac{X(t)^2}{X_{\max}}}^{\text{prey birth}} \tag{5.27}$$

$$\frac{\mathrm{d}Y(t)}{\mathrm{d}t} = \underbrace{+ \beta X(t)Y(t)}_{\substack{\text{prey consumption} \\ \& \text{ predator birth}}} \underbrace{- \delta Y(t)}_{\text{predator death}} \tag{5.28}$$

which describe how the population size $X(t) \in \mathbb{R}^+$ of a prey species and the population size $Y(t) \in \mathbb{R}^+$ of a predator species change continuously over time as the species interact (black curve in Figure 5.5a). The parameter $\alpha$ is the birth rate of prey, $\delta$ is the death rate of predators and $\beta$ is the contact rate between predators and prey. Upon contact a prey is consumed and, we assume for simplicity, exactly one predator is born. $X_{\max}$ is a finite carry capacity representing the available plant resources for the prey species, which would result in logistic growth in the absence of predators. This is neither necessary nor commonly assumed in the literature on the deterministic model, since the prey population is always limited by a nonzero number of predators in $\mathbb{R}^+$.

Here we are interested in a corresponding stochastic model [58] (blue curves in Figure 5.5a) in which the number of predators is an integer and may drop to zero, which would lead to exponential growth of the prey population without finite carry capacity. We define a stochastic predator-prey model as a CTMC over the state space

$$\{0, \dots, X_{\max}\} \times \{0, \dots, Y_{\max}\}$$

with transition-rate matrix

$$Q_{(X+\Delta X, Y+\Delta Y),(X,Y)} = \begin{cases} \beta XY & \text{if } \Delta X = -1, \Delta Y = +1, \\ \delta Y & \text{if } \Delta X = 0, \Delta Y = -1, \\ \alpha X - \alpha X^2 / X_{\max} & \text{if } \Delta X = +1, \Delta Y = 0, \\ -\beta XY - \delta Y - \alpha X + \alpha X^2 / X_{\max} & \text{if } \Delta X = 0, \Delta Y = 0, \\ 0 & \text{otherwise} \end{cases} \tag{5.29}$$

whose columns sum to less than zero. This is because the upper limit $Y_{\max}$ of the predator population is a computational cutoff and not enforced by the model. Transitions that leave the state space result in missing probability mass, which must be mitigated by choosing a sufficiently high $Y_{\max}$.

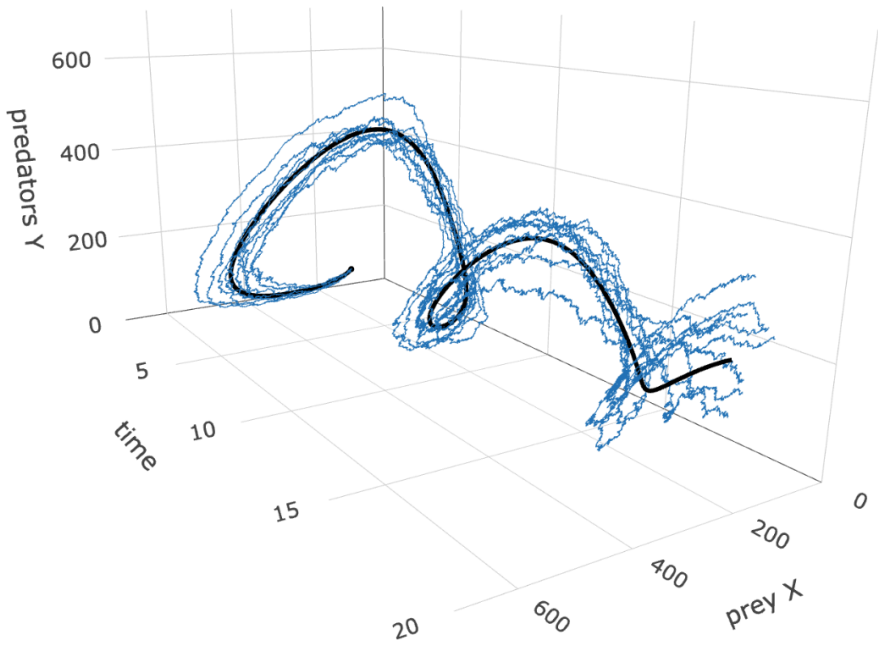
We introduce the band matrices

$$\underbrace{\begin{aligned} \mathcal{X}^+_{\text{cons}} &= \text{superdiag}(1, \dots, X_{\max}), \\ \mathcal{X}^-_{\text{cons}} &= \text{diag}(0, \dots, X_{\max}), \\ \mathcal{X}^+_{\text{birth}} &= \text{subdiag}(0, \dots, X_{\max} - 1), \\ \mathcal{X}^-_{\text{birth}} &= \text{diag}(0, \dots, X_{\max} - 1, 0), \\ \mathcal{X}^+_{\text{cap}} &= \text{subdiag}(0^2, 1^2, 2^2, \dots, (X_{\max} - 1)^2), \\ \mathcal{X}^-_{\text{cap}} &= \text{diag}(0^2, 1^2, 2^2, \dots, (X_{\max} - 1)^2, 0), \\ \mathcal{X}^+_{\text{death}} &= \text{diag}(1, 1, \dots, 1) = \text{I}, \\ \mathcal{X}^-_{\text{death}} &= \text{diag}(1, 1, \dots, 1) = \text{I}, \end{aligned}}_{(X_{\max}+1) \times (X_{\max}+1)} \qquad \underbrace{\begin{aligned} \mathcal{Y}^+_{\text{cons}} &= \text{subdiag}(0, \dots, Y_{\max} - 1), \\ \mathcal{Y}^-_{\text{cons}} &= \text{diag}(0, \dots, Y_{\max}), \\ \mathcal{Y}^+_{\text{birth}} &= \text{diag}(1, 1, \dots, 1) = \text{I}, \\ \mathcal{Y}^-_{\text{birth}} &= \text{diag}(1, 1, \dots, 1) = \text{I}, \\ \mathcal{Y}^+_{\text{cap}} &= \text{diag}(1, 1, \dots, 1) = \text{I}, \\ \mathcal{Y}^-_{\text{cap}} &= \text{diag}(1, 1, \dots, 1) = \text{I}, \\ \mathcal{Y}^+_{\text{death}} &= \text{superdiag}(1, \dots, Y_{\max}), \\ \mathcal{Y}^-_{\text{death}} &= \text{diag}(1, \dots, Y_{\max}), \end{aligned}}_{(Y_{\max}+1) \times (Y_{\max}+1)} \tag{5.30}$$

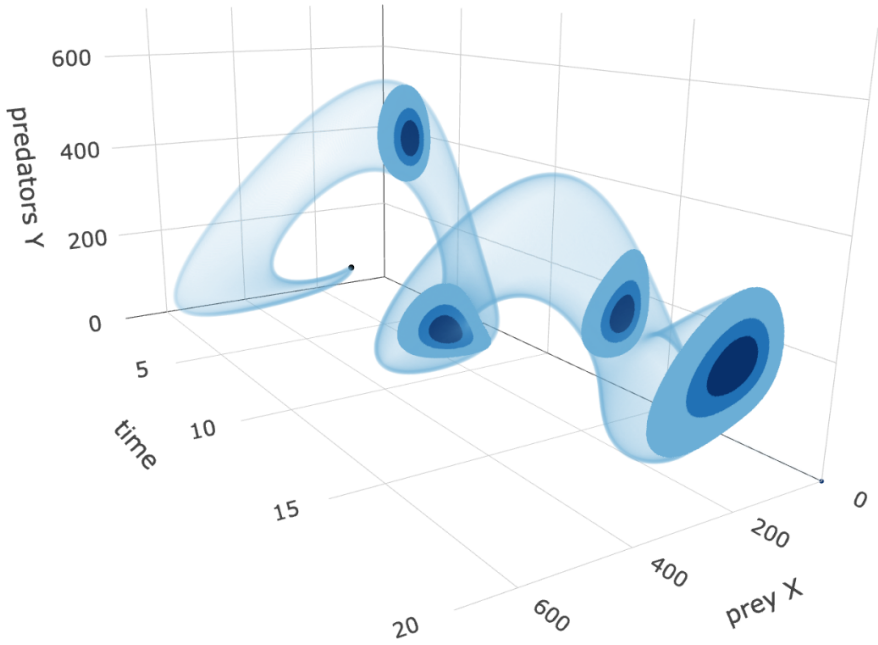
in order to represent the transition-rate matrix

$$\begin{aligned} Q = &+ \beta(\mathcal{X}^+_{\text{cons}} \otimes \mathcal{Y}^+_{\text{cons}}) + \alpha(\mathcal{X}^+_{\text{birth}} \otimes \mathcal{Y}^+_{\text{birth}}) - \frac{\alpha}{X_{\max}}(\mathcal{X}^+_{\text{cap}} \otimes \mathcal{Y}^+_{\text{cap}}) + \delta(\mathcal{X}^+_{\text{death}} \otimes \mathcal{Y}^+_{\text{death}}) \\ &- \beta(\mathcal{X}^-_{\text{cons}} \otimes \mathcal{Y}^-_{\text{cons}}) - \alpha(\mathcal{X}^-_{\text{birth}} \otimes \mathcal{Y}^-_{\text{birth}}) + \frac{\alpha}{X_{\max}}(\mathcal{X}^-_{\text{cap}} \otimes \mathcal{Y}^-_{\text{cap}}) - \delta(\mathcal{X}^-_{\text{death}} \otimes \mathcal{Y}^-_{\text{death}}) \end{aligned} \tag{5.31}$$

as a sum of tensor products. This allows us to efficiently compute solutions $\mathbf{p}(t)$ of the Kolmogorov equation for the stochastic predator-prey model (see Figure 5.5b).




**(a)** Solution of a deterministic predator-prey model (black curve) and 10 randomly sampled trajectories (blue) of the corresponding stochastic model.




**(b)** Analytic solution of the Kolmogorov equation for a stochastic predator-prey model. The time slices show distributions $\mathbf{p}(t)$ where the shades of blue show the smallest areas that contain 30%, 60% and 90% of the probability mass at time $t$.

**Figure 5.5.:** Illustration of predator-prey models with $\alpha = 1$, $\beta = 0.004$, $\delta = 0.8$, $X(0) = 100$, $Y(0) = 40$, $X_{\max} = Y_{\max} = 1200$.

# 6. Extending MHNs by pivotal events

*In this chapter we extend Mutual Hazard Networks by certain events that play a pivotal role in tumor progression, namely the diagnosis of the tumor itself, the death of the patient, seeding of a metastasis and temporary events such as inflammation. For these extended MHNs we provide tensor representations of $Q$ and algorithms for computing the likelihood of corresponding observations and their derivative, largely based on differentiated uniformization.*

## 6.1. Diagnosis event

One major limitation that MHNs share with CBNs and other tumor progression models is the assumption that the time of diagnosis is a random variable that is independent of the tumor state. However, in reality the diagnosis does occur precisely because and when the tumor becomes noticable, for example by growing over a certain size. Hence we extend the MHN framework by *diagnosis of the tumor* as its own event which may be influenced by any other progression event.

We believe this will not only enhance the explanatory power of our models, but also improve their causal interpretability by explaining away spurious anti-correlations: any event that causes the diagnosis of the tumor will look to a standard MHN mistakenly as if it were inhibiting all other events (because it leaves no more time for their occurance).

Formally, we introduce and treat diagnosis (say, event $n$) nearly the same as any other progression event. It has free parameters $\Theta_{nj}$ which can be learned from data and represent the multiplicative effect of any other event $j$ on the rate of the diagnosis. However we fix the outgoing effects of the diagnosis on all other events $i$ to $\Theta_{in} = 0$ and do not change these during the learning. This means that as soon as the diagnosis occurs, it halts all other events from ever occuring by multiplying their rates with zero, thereby "freezing" the tumor progression. A tumor will have the same state in the infinite future which it had at the moment of diagnosis (see Figure 6.1).

We can therefore read off the likelihood $\mathbf{p}(\infty)_{\mathbf{x}}$ of a tumor state $\mathbf{x}$ at the time of its

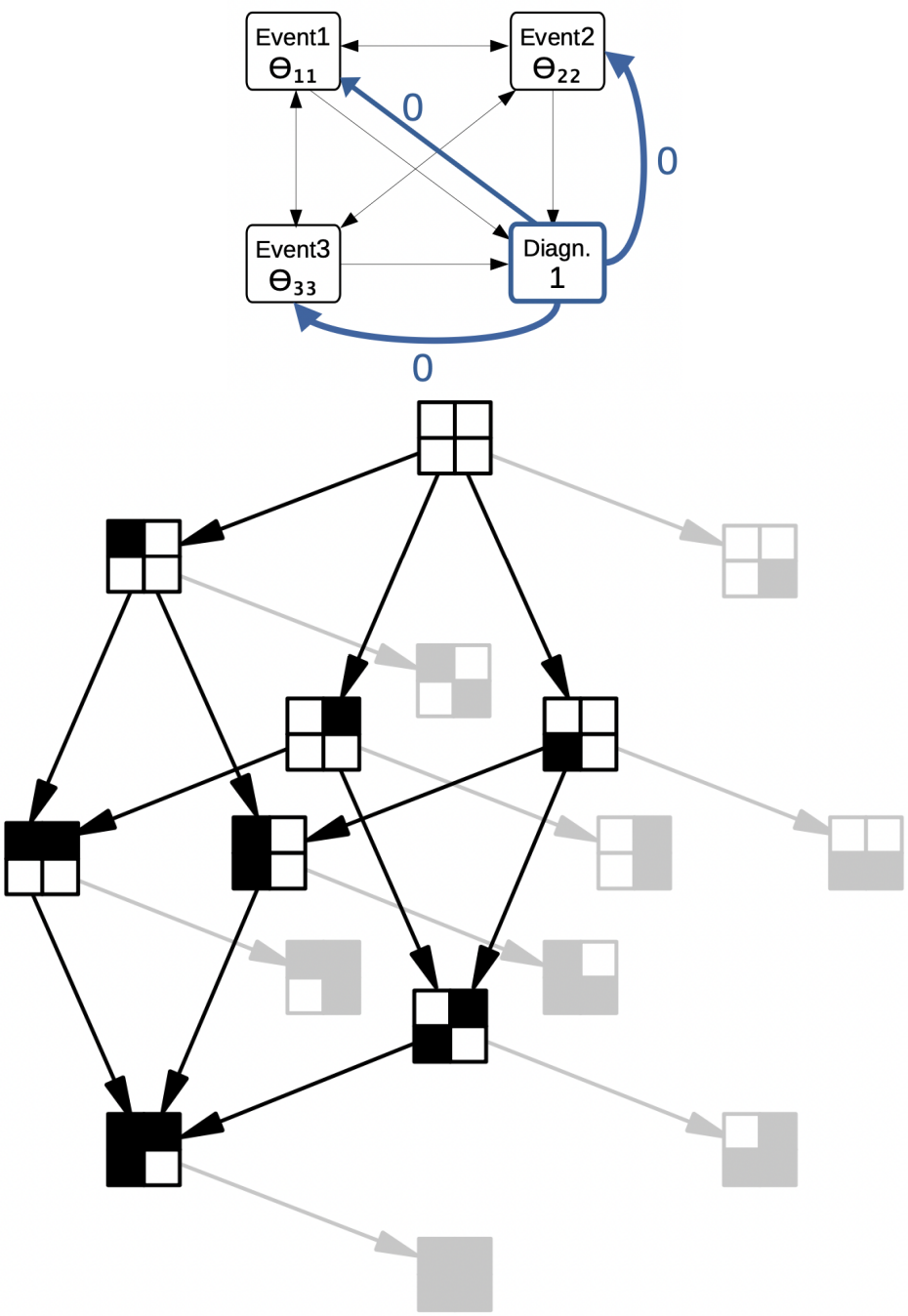



**Figure 6.1.:** Illustration how learning a diagnosis event can be implemented. Top: The occurance of a diagnosis multiplies the rates of all other events by zero (blue). Bottom: This freezes tumor progression such that the tumor will remain forever in the same state (greyed out) it had at the moment of diagnosis. Its probability can then be read off from the stationary distribution of the Markov chain.

diagnosis from the stationary distribution

$$\mathbf{p}(\infty) := \underbrace{\left(\frac{1}{\gamma}Q + I\right)}_{=:P}^{m} \mathbf{p}(0) \qquad \text{for } m \to \infty, \gamma \geq \max_x |Q_{x,x}| \tag{6.1}$$

This distribution and its derivative

$$\frac{\partial \mathbf{p}(\infty)}{\partial \theta} = \left[\frac{\partial P}{\partial \theta} P^{m-1} + P \frac{\partial P^{m-1}}{\partial \theta}\right] \mathbf{p}(0) \qquad \text{for } m \to \infty \tag{6.2}$$

can be computed by algorithm 3, which a slight adaptation of algorithm 2, see [63].

**Algorithm 3:** Computing the stationary distribution and its derivative

**input :** $\mathbf{p}(0), P, P', \varepsilon$
**output:** $\mathbf{p}(\infty)$, $\mathbf{p}(\infty)'$

$\mathbf{p}(\infty) \leftarrow \mathbf{0}$
$\mathbf{p}(\infty)' \leftarrow \mathbf{0}$
**repeat**
  $\mathbf{p}(\infty)' \leftarrow P'\mathbf{p}(\infty) + P\mathbf{p}(\infty)'$
  $\mathbf{p}(\infty) \leftarrow P\mathbf{p}(\infty)$
**until** $1 - \mathbf{d}^T\mathbf{p}(\infty) < \varepsilon$;
**return** $\mathbf{p}(\infty), \mathbf{p}(\infty)'$

The required number $m$ of iterations can be chosen such that the total probability $\mathbf{d}^T\mathbf{p}(\infty)$ of diagnosed states is close to 1, where

$$\mathbf{d} := \bigotimes_{i=1}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{6.3}$$

is the vector whose entries are 1 for all diagnosed states and 0 elsewhere.

## 6.2. Death event and survival analysis

Another desired goal of tumor progression models is the prediction of survival or death of a patient, particularly in order to guide therapeutic interventions. To this end we extend the MHN framework by *death of the patient* as another special event (say, again event $n$) which may be influenced by any other progression event.

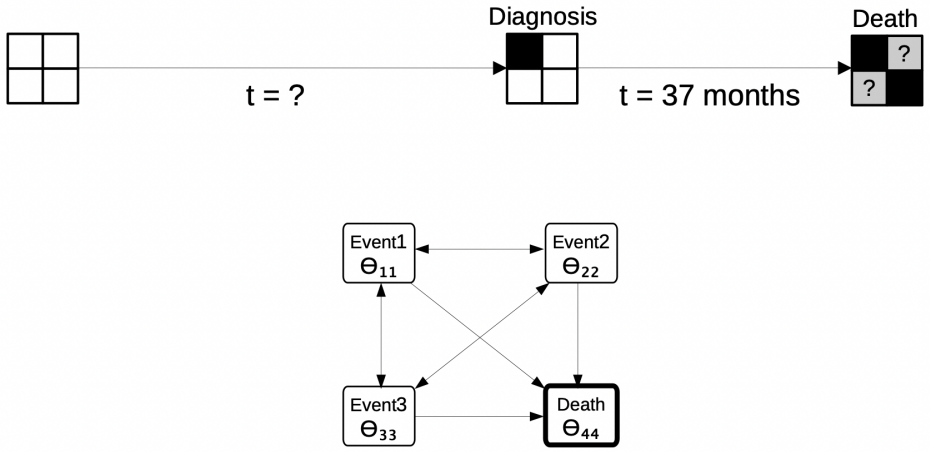



**Figure 6.2.:** Illustration of data that can be used to learn an MHN with a death event. While primary tumor data do not tell the elapsed time from the onset of tumor progression, they show the full state of the tumor. Survival data on the other hand tell the time elapsed until death, but do not show any other events that might have accumulated after diagnosis of the primary tumor.

The multiplicative effect $\Theta_{nj}$ of any other event $j$ on the rate of death can be learned from data. This requires a patient's survival status after a given follow up time from the diagnosis of the primary tumor, which is available for example in some datasets from TCGA.

Let $\mathbf{p}(0)$ be a vector with $\mathbf{p}(0)_{\mathbf{x}} = 1$ and all other entries zero. Let

$$\mathbf{d} := \bigotimes_{i=1}^{n-1} \begin{pmatrix} 1 \\ 1 \end{pmatrix} \otimes \begin{pmatrix} 0 \\ 1 \end{pmatrix} \tag{6.4}$$

be the vector whose entries are 1 for all dead states and 0 elsewhere. The likelihood that a patient who had a tumor with state $\mathbf{x}$ at time 0 is still alive by time $t$ is

$$1 - \mathbf{d}^T \exp(tQ)\mathbf{p}(0) \tag{6.5}$$

and the likelihood that he is dead **by** time $t$ is

$$F(t) = \mathbf{d}^T \exp(tQ)\mathbf{p}(0). \tag{6.6}$$

Survival datasets typically state that a patient has died **at** time $t$, for which the likelihood is

$$f(t) = \frac{d}{dt} F(t) \tag{6.7}$$

$$= \mathbf{d}^T \frac{d}{dt} \exp(tQ)\mathbf{p}(0) \tag{6.8}$$

$$= \mathbf{d}^T Q \exp(tQ)\mathbf{p}(0). \tag{6.9}$$

Its derivative is

$$\frac{\partial f(t)}{\partial \theta} = \mathbf{d}^T \left[ \frac{\partial Q}{\partial \theta} \exp(tQ) + Q \frac{\partial \exp(tQ)}{\partial \theta} \right] \mathbf{p}(0). \tag{6.10}$$

The expressions involving the matrix exponential and its derivative can be computed using algorithm 2. Note, however, that primary tumor models cannot be directly combined with survival data, since the former treat time in units relative to the mean tumor age, while the latter treat time in absolute units of days. Hence, when using survival data one should also include a diagnosis event and learn its base rate in terms of absolute time units as well.

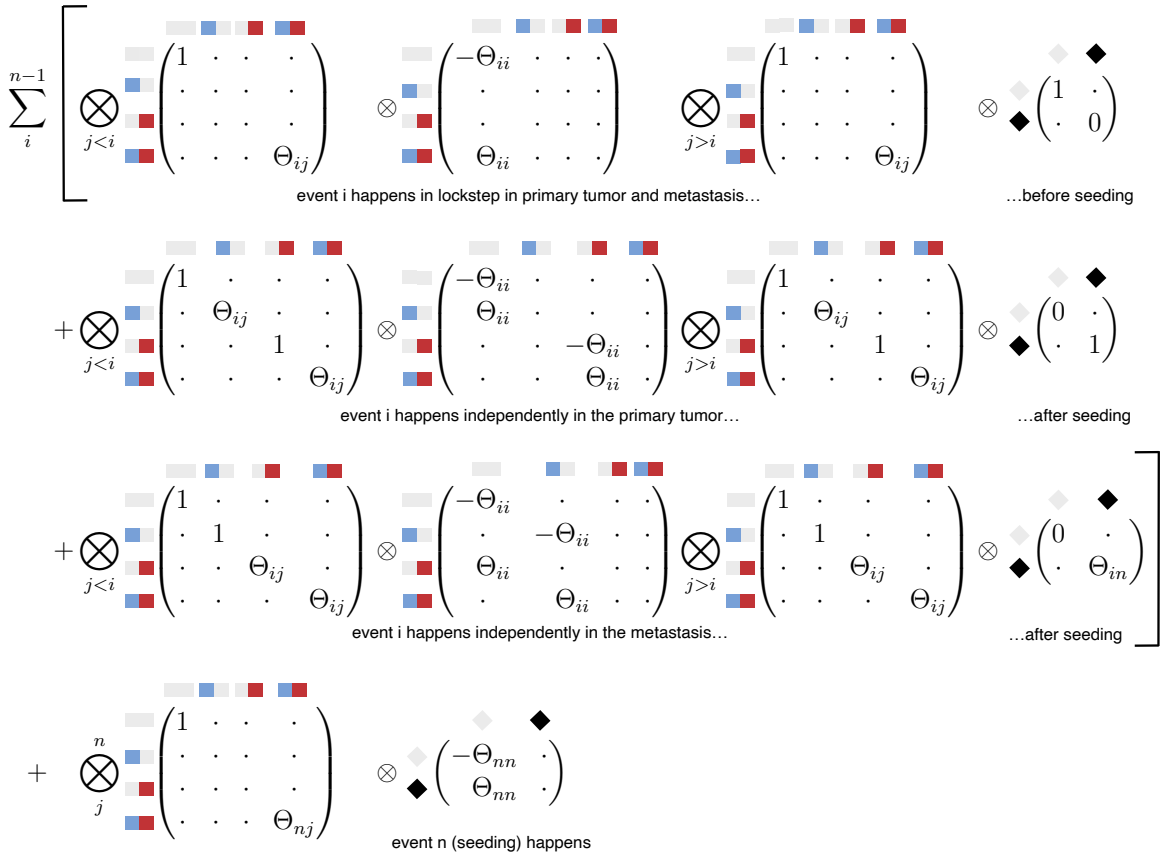
## 6.3. Seeding of metastasis

Since most human cancer deaths are due to a metastasis, we would like to model its genome in addition to the genome of the primary tumor as well as the seeding event that spawns the metastasis.

To this end we introduce for each event that may be present or not present in the primary tumor a copy of this event that may be independently present or not present in the metastasis. At first we treat the genome of the (soon-to-be) metastasis however as part of the homogenous primary tumor. That is, it has exactly the same events as the primary tumor and also acquires new events in lockstep with the primary tumor.

In addition, we introduce a seeding event which represents the metastasis breaking away from the primary tumor. After the seeding event has occured, the primary tumor and metastasis no longer acquire events in lockstep but independently, and may from then on differ in their genomes fig. 6.3.

Following [48] we can derive the following tensor structure for Q [64]:




This model can be learned by computing the time-marginal distribution for the primary tumor and then propagating it forward using the matrix exponential for the metastasis.

## 6.4. Inflammation and other temporary events

Finally, we would like to model immunological events such as an inflammation of the tumor which can arise and then abate after some time.

We account for temporary events by introducing additional parameters $(\Theta^-_{ij}) \in \mathbb{R}^{n\times n}$, where $\Theta^-_{ii}$ is the baseline removal rate of event $i$ and $\Theta^-_{ij}$ is the multiplicative effect of (the presence of) event $j$ on the removal rate of event $i$. In order to avoid overfitting one could learn only the baseline removal rates $\Theta^-_{ii}$ as free parameters and allow no dependencies by setting all other $\Theta^-_{ij} = 1$. Another natural choice would be to set $\Theta^-_{ij} = \Theta^{-1}_{ij}$, i.e., if event $j$ doubles the occurance rate of event, then event $j$ necessarily halves the removal rate of event $i$.

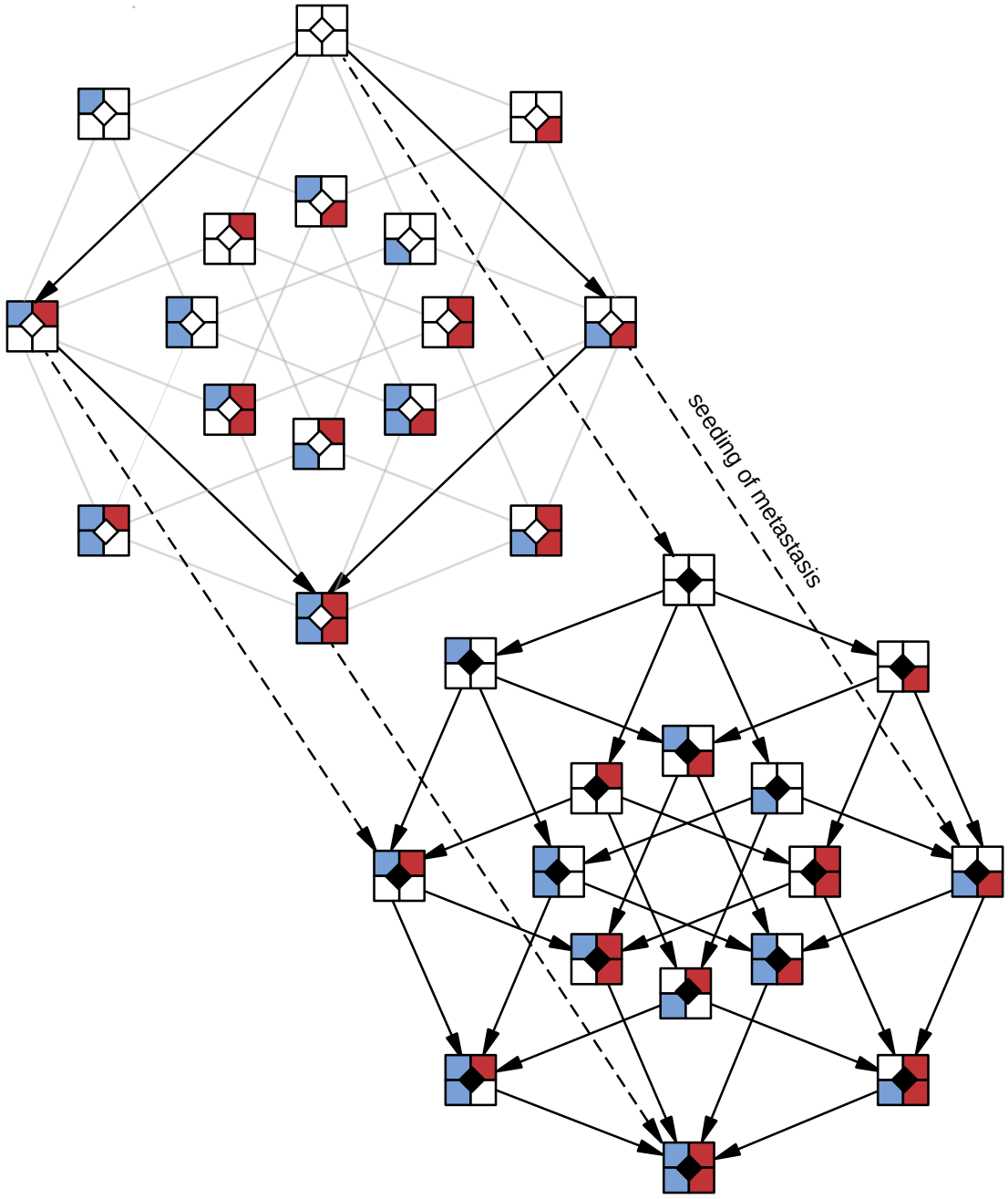



**Figure 6.3.:** Illustration of an MHN with Metastasis for 2 normal events + 1 seeding event. The tiled squares show possible states of a paired primary tumor and (potential) metastatis: Each event comes in two copies, one in the primary tumor (left, blue), one in the metastasis (right, red). The black diamond in the center indicates the seeding event and the corresponding transitions are the dashed arrows. Before the seeding (upper 4-cube) the potential metastasis accumulates events in lockstep with the primary tumor. After the seeding (lower 4-cube) all events happen independently.

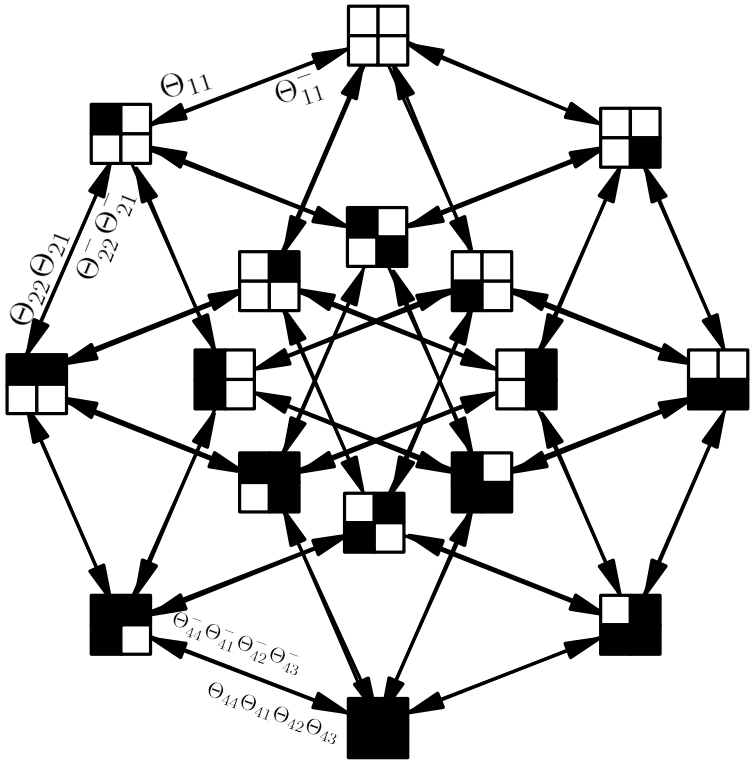



**Figure 6.4.:** Illustration of eq. (6.11) for $n = 4$. Labels correspond to arrowheads.

Any of these choices admit a tensor representation of

$$
\begin{aligned}
Q = &\sum_{i=1}^{n} \bigotimes_{j<i} \begin{pmatrix} 1 & 0 \\ 0 & \Theta_{ij} \end{pmatrix} \otimes \begin{pmatrix} -\Theta_{ii} & 0 \\ \Theta_{ii} & 0 \end{pmatrix} \otimes \bigotimes_{j>i} \begin{pmatrix} 1 & 0 \\ 0 & \Theta_{ij} \end{pmatrix} \\
&+ \sum_{i=1}^{n} \bigotimes_{j<i} \begin{pmatrix} 1 & 0 \\ 0 & \Theta^-_{ij} \end{pmatrix} \otimes \begin{pmatrix} 0 & \Theta^-_{ii} \\ 0 & -\Theta^-_{ii} \end{pmatrix} \otimes \bigotimes_{j>i} \begin{pmatrix} 1 & 0 \\ 0 & \Theta^-_{ij} \end{pmatrix}. \qquad (6.11)
\end{aligned}
$$

Note however that this $Q$ is no longer lower-triangular unless $\Theta^-_{ii} = 0$ for all $i$, which would be equivalent to a standard MHN. This means that we can no longer compute

$\mathbf{p}_\Theta$ according to eq. (2.11). Instead we use the series expansion of

$$\mathbf{p}_\Theta = \int_0^\infty \exp(-t)\mathbf{p}(t)dt \tag{6.12}$$

$$= \int_0^\infty \sum_{m=0}^{\infty} \frac{(\gamma t)^m}{m!} e^{-\gamma t} P^m \mathbf{p}(0) dt \tag{6.13}$$

$$= \sum_{m=0}^{\infty} \frac{\gamma^m}{m!} \int_0^\infty t^m e^{-(\gamma+1)t} dt P^m \mathbf{p}(0) \tag{6.14}$$

$$= \sum_{m=0}^{\infty} \frac{\gamma^m \Gamma(m+1)}{m!(1+\gamma)^{m+1}} P^m \mathbf{p}(0) \tag{6.15}$$

$$= \sum_{m=0}^{\infty} \frac{\gamma^m}{(1+\gamma)^{m+1}} P^m \mathbf{p}(0) \tag{6.16}$$

[48] and compute it together with its derivative

$$\frac{\partial \mathbf{p}_\Theta}{\partial \theta} = \sum_{m=0}^{\infty} \frac{\gamma^{m-1}}{(1+\gamma)^{m+2}} \left( \frac{\partial \gamma}{\partial \theta}(m-\gamma)P^m \mathbf{p}(0) + \gamma(1+\gamma)\frac{\partial P^m}{\partial \theta}\mathbf{p}(0) \right) \tag{6.17}$$

using algorithm 4, an adaptation of algorithm 2.

**Algorithm 4:** Marginalized Differentiated Uniformization

**input :** $\mathbf{p}(0), P, P', \gamma, \gamma', \varepsilon$
**output:** $\mathbf{p}_\Theta, \mathbf{p}'_\Theta$

$m \leftarrow 0$
$w \leftarrow \frac{\gamma}{1+\gamma}$
$\mathbf{p}_\Theta \leftarrow \mathbf{0}$
$\mathbf{p}'_\Theta \leftarrow \mathbf{0}$
$\mathbf{p} \leftarrow \mathbf{p}(0)$
$\mathbf{p}' \leftarrow \mathbf{0}$
**repeat**
  $\mathbf{p}_\Theta \leftarrow \mathbf{p}_\Theta + w\mathbf{p}$
  $\mathbf{p}'_\Theta \leftarrow \mathbf{p}'_\Theta + \frac{w}{\gamma(1+\gamma)}(\gamma'(m-\gamma)\mathbf{p} + \gamma(1+\gamma)\mathbf{p}')$
  $m \leftarrow m + 1$
  $\mathbf{p}' \leftarrow P'\mathbf{p} + P\mathbf{p}'$
  $\mathbf{p} \leftarrow P\mathbf{p}$
  $w \leftarrow \frac{\gamma}{1+\gamma}w$
**until** $|1 - \mathbf{p}_\Theta|_1 < \varepsilon$;
**return** $\mathbf{p}_\Theta, \mathbf{p}'_\Theta$

# 7. Discussion and outlook

We have presented Mutual Hazard Networks, a new framework for modeling tumor progression from cross-sectional bulk observations. MHNs build and improve upon Conjunctive Bayesian Networks [8] and Network Aberration Models [44] by simultaneously allowing for stochastic dependencies between progression events, as well as inhibiting dependencies and cyclic dependencies.

We have shown that our models consistently outperform the previous state of the art in terms of cross-validated model fit. In particular, MHN was the only model that inferred from a glioblastoma dataset that despite their rare occurance, IDH1 mutations precede the much more frequent TP53 mutations, which is endorsed by independent data from consecutive biopsies [76]. Where in the training data was the evidence for this interpretation? We found it in the four *IDH1(M)* positive / *TP53(M)* negative cases (Fig. 4.4A, purple). All of them had at most one mutation in addition to *IDH1(M)*.

This greater flexibility compared to earlier models comes however at the cost of a greater computational complexity. We formulate and learn MHNs as general Markov Chain models on a huge combinatorial state space, which seems at first glance intractable. Our key observation was that the transition-rate matrix of such a model can be written as a sum of tensor products, which allows us to cheaply compute matrix-vector products without storing the matrix itself. This operation alone is sufficient to compute the time-marginal distribution and its derivative with a computational complexity that is "only" exponential in the number modeled events. By applying a state space restriction we were able to further reduce the complexity to exponential in the number of events that have occured per patient.

Moreover, we adapted the same tensor approach to computing the transient distribution of a large Markov chain and its derivative. We demonstrated its general usefulness by performing a full Bayesian analysis of the stochastic SIR model of epidemic spread, which was commonly seen as intractable. We then used this operation as a building block for extending Mutual Hazard Networks with special events that play an important role in tumor progression.

These include the diagnosis of the tumor, the death of the patient, and seeding of a metastasis. While we have laid the groundwork for such advanced models by finding a tensor representation for their transition-rate matrices, they still remain out of reach

in terms of computational cost. This is because the state space restriction no longer for survival analysis (since one is looking into the future and all states may be reachable) and modeling of metastasis (since a metastasis is often not observed, yet may be still present and all possible states must be accounted for).

Therefore we will have to break to curse of dimensionality and reduce the complexity from exponential to polynomial in the number of events. This is possible by treating not only the transition-rate matrix of a Markov chain as a tensor, but also its probability distribution. Modern low-rank tensor formats then allow for an efficient arithmetic, which has already been explored for MHN by [31, 48] and may enable more advanced models in the near future.

# Bibliography


[1] H. Akaike. "A new look at the statistical model identification". In: *IEEE Transactions on Automatic Control* 19.6 (Dec. 1974), pp. 716–723. DOI: 10.1109/tac.1974.1100705.

[2] L. J. S. Allen. "A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis". In: *Infectious Disease Modelling* 2.2 (2017), pp. 128–142. DOI: https://doi.org/10.1016/j.idm.2017.03.001.

[3] V. Amoia, G. De Micheli, and M. Santomauro. "Computer-Oriented Formulation of Transition-Rate Matrices via Kronecker Algebra". In: *IEEE Transactions on Reliability* R-30.2 (June 1981), pp. 123–132. ISSN: 0018-9529. DOI: 10.1109/TR.1981.5221004.

[4] Galen Andrew and Jianfeng Gao. "Scalable Training of L1-regularized Log-linear Models". In: *Proceedings of the 24th International Conference on Machine Learning*. ICML '07. Corvalis, Oregon, USA: ACM, 2007, pp. 33–40. ISBN: 978-1-59593-793-3. DOI: 10.1145/1273496.1273501.

[5] "Assessing the significance of chromosomal aberrations in cancer: methodology and application to glioma". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 104.50 (Dec. 2007), pp. 20007–20012.

[6] Michael Baudis and Michael L. Cleary. "Progenetix.net: an online repository for molecular cytogenetic aberration data". In: *Bioinformatics* 17 12 (2001), pp. 1228–9.

[7] N. Beerenwinkel and S. Sullivant. "Markov models for accumulating mutations". In: *Biometrika* 96.3 (2009), pp. 645–661. DOI: https://doi.org/10.1093/biomet/asp023.

[8] Niko Beerenwinkel, Nicholas Eriksson, and Bernd Sturmfels. "Conjunctive Bayesian networks". In: *Bernoulli* 13.4 (Nov. 2007), pp. 893–909. DOI: 10.3150/07-BEJ6133.

[9] Niko Beerenwinkel, Jörg Rahnenführer, Martin Däumer, Daniel Hoffmann, Rolf Kaiser, Joachim Selbig, and Thomas Lengauer. "Learning Multiple Evolutionary Pathways from Cross-Sectional Data". In: *Journal of Computational Biology* 12.6 (July 2005), pp. 584–598. DOI: 10.1089/cmb.2005.12.584.

[10] Niko Beerenwinkel, Roland F Schwarz, Moritz Gerstung, and Florian Markowetz. "Cancer evolution: mathematical models and computational inference". en. In: *Syst. Biol.* 64.1 (Jan. 2015), e1–25.

[11] J. R. Bischoff. “A homologue of Drosophila aurora kinase is oncogenic and amplified in human colorectal cancers”. In: *The EMBO Journal* 17.11 (June 1998), pp. 3052–3065. DOI: 10.1093/emboj/17.11.3052.

[12] P. Buchholz. “Structured analysis approaches for large Markov chains”. In: *Applied Numerical Mathematics* 31.4 (1999), pp. 375–404. ISSN: 0168-9274. DOI: https://doi.org/10.1016/S0168-9274(99)00005-7.

[13] Peter Buchholz. “Structured analysis approaches for large Markov chains”. In: *Applied Numerical Mathematics* 31.4 (1999), pp. 375–404. ISSN: 0168-9274. DOI: 10.1016/S0168-9274(99)00005-7.

[14] Horst Buerger, Friedrich Otterbach, Ronald Simon, Karl-Ludwig Schï¿½fer, Christopher Poremba, Raihanatou Diallo, Christian Brinkschmidt, Barbara Dockhorn-Dworniczak, and Werner Boecker. “Different genetic pathways in the evolution of invasive breast cancer are associated with distinct morphological subtypes”. In: *The Journal of Pathology* 189.4 (Dec. 1999), pp. 521–526. DOI: 10.1002/(sici)1096-9896(199912)189:4<521::aid-path472>3.0.co;2-b.

[15] P E. Buis and Wayne R. Dyksen. “Efficient Vector and Parallel Manipulation of Tensor Products”. In: 22.1 (1996), 18â€“23. DOI: https://doi.org/10.1145/225545.225548.

[16] Paul E. Buis and Wayne R. Dyksen. “Efficient Vector and Parallel Manipulation of Tensor Products”. In: *ACM Trans. Math. Softw.* 22.1 (Mar. 1996), pp. 18–23. ISSN: 0098-3500. DOI: 10.1145/225545.225548.

[17] Pflege und Konsumentenschutz (BMSGPK) Bundesministerium fÃ¼r Soziales Gesundheit. *Open Data Ã–sterreich*. 2021. URL: https://www.data.gv.at/katalog/dataset/ef8e980b-9644-45d8-b0e9-c6aaf0eff0c0 (visited on 10/15/2021).

[18] Ethan Cerami, Jianjiong Gao, Ugur Dogrusoz, et al. “The cBio Cancer Genomics Portal: An Open Platform for Exploring Multidimensional Cancer Genomics Data.” In: *Cancer Discovery* 2.5 (May 2012), pp. 401–404. DOI: 10.1158/2159-8290.cd-12-0095.

[19] G. Ciriello, E. Cerami, C. Sander, et al. “Mutual exclusivity analysis identifies oncogenic network modules”. In: *Genome Research* 22.2 (Sept. 2011), pp. 398–406. DOI: 10.1101/gr.125567.111.

[20] Simona Constantinescu, Ewa Szczurek, Pejman Mohammadi, et al. “TiMEx: a waiting time model for mutually exclusive cancer alterations”. In: *Bioinformatics* 32.7 (July 2015), pp. 968–975. DOI: 10.1093/bioinformatics/btv400.

[21] C. L. Cowey and W. K. Rathmell. “VHL gene mutations in renal cell carcinoma: role as a biomarker of disease outcome and drug efficacy”. In: *Curr Oncol Rep* 11.2 (Mar. 2009), pp. 94–101.

[22] D. R. Cox. “Regression Models and Life-Tables”. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 34.2 (1972), pp. 187–220. ISSN: 00359246.

[23] Simona Cristea, Jack Kuipers, and Niko Beerenwinkel. “pathTiMEx: Joint Inference of Mutually Exclusive Cancer Pathways and Their Progression Dynam-

ics". In: *Journal of Computational Biology* 24.6 (June 2017), pp. 603–615. DOI: 10.1089/cmb.2016.0171.

[24] Richard Desper, Feng Jiang, Olli-P. Kallioniemi, et al. "Inferring Tree Models for Oncogenesis from Comparative Genome Hybridization Data". In: *Journal of Computational Biology* 6.1 (Jan. 1999), pp. 37–51. DOI: 10.1089/cmb.1999.6.37.

[25] Juan Diaz-Colunga and Ramon Diaz-Uriarte. "Conditional prediction of consecutive tumor evolution using cancer progression models: What genotype comes next?" In: *PLOS Computational Biology* 17.12 (Dec. 2021). Ed. by Rainer Spang, e1009055. DOI: 10.1371/journal.pcbi.1009055.

[26] S. Duane, A. D. Kennedy, B. P. J. Pendleton, and D. Roweth. "Hybrid Monte Carlo". In: *Physics Letters B* 195.2 (1987), pp. 216–222. DOI: https://doi.org/10.1016/0370-2693(87)91197-X.

[27] Christel Faes, Steven Abrams, Dominique Van Beckhoven, Geert Meyfroidt, Erika Vlieghe, and Niel Hens and. "Time between Symptom Onset, Hospitalisation and Recovery or Death: Statistical Analysis of Belgian COVID-19 Patients". In: *International Journal of Environmental Research and Public Health* 17.20 (Oct. 2020), p. 7560. DOI: 10.3390/ijerph17207560.

[28] Hossein Shahrabi Farahani and Jens Lagergren. "Learning Oncogenetic Networks by Reducing to Mixed Integer Linear Programming". In: *PLoS ONE* 8.6 (June 2013). Ed. by Jian-Xin Gao, e65773. DOI: 10.1371/journal.pone.0065773.

[29] Eric R. Fearon and Bert Vogelstein. "A genetic model for colorectal tumorigenesis". In: *Cell* 61.5 (June 1990), pp. 759–767. DOI: 10.1016/0092-8674(90)90186-i.

[30] A. Gelman, J. B. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. 3rd ed. Chapman and Hall/CRC, 2013. DOI: https://doi.org/10.1201/b16018.

[31] Peter Georg. *Tensor Train Decomposition for solving high-dimensional Mutual Hazard Networks*. 2022.

[32] Peter Georg, Nils Meyer, Stefan Solbrig, and Tilo Wettig. *Grid on QPACE 4*. 2021. arXiv: 2112.01852 [hep-lat].

[33] M. Gerstung, M. Baudis, H. Moch, et al. "Quantifying cancer progression with conjunctive Bayesian networks". In: *Bioinformatics* 25.21 (Aug. 2009), pp. 2809–2815. DOI: 10.1093/bioinformatics/btp505.

[34] Moritz Gerstung, Nicholas Eriksson, Jimmy Lin, et al. "The Temporal Order of Genetic and Pathway Alterations in Tumorigenesis". In: *PLoS ONE* 6.11 (Nov. 2011). Ed. by Amanda Ewart Toland, e27136. DOI: 10.1371/journal.pone.0027136.

[35] D. T. Gillespie. "A general method for numerically simulating the stochastic time evolution of coupled chemical reactions". In: *Journal of Computational*

*Physics* 22.4 (1976), pp. 403–434. DOI: https://doi.org/10.1016/0021-9991(76)90041-3.

[36] J.R. Gnarra, K. Tory, Y. Weng, L. Schmidt, M.H. Wei, H. Li, F. Latif, S. Liu, F. Chen, F.-M. Duh, I. Lubensky, D.R. Duan, C. Florence, R. Pozzatti, M. M. Walther, N.H. Bander, H.B. Grossman, H. Brauch, S. Pomer, J.D. Brooks, W.B. Isaacs, M.I. Lerman, B. Zbar, and W.M. Linehan. "Mutations of the VHL tumour suppressor gene in renal carcinoma". In: *Nature Genetics* 7.1 (May 1994), pp. 85–90. DOI: 10.1038/ng0594-85.

[37] D S Goodsell. "p53 Tumor Suppressor". In: *RCSB Protein Data Bank* (July 2002).

[38] D S Goodsell. "MDM2 and Cancer". In: *RCSB Protein Data Bank* (June 2019).

[39] David S Goodsell. "The molecular perspective: VEGF and angiogenesis". en. In: *Stem Cells* 21.1 (2003), pp. 118–119.

[40] David S Goodsell, Shuchismita Dutta, Maria Voigt, Christine Zardecki, and Stephen K Burley. "Molecular explorations of cancer biology and therapeutics at PDB-101". en. In: *Oncogene* (Aug. 2022).

[41] W.K. Grassmann. "Transient solutions in markovian queueing systems". In: *Computers & Operations Research* 4.1 (1977), pp. 47–53. DOI: https://doi.org/10.1016/0305-0548(77)90007-7.

[42] Katrin Hainke, Jörg Rahnenführer, and Roland Fried. "Cumulative disease progression models for cross-sectional data: A review and comparison". In: *Biometrical Journal* 54.5 (Aug. 2012), pp. 617–640. DOI: 10.1002/bimj.201100186.

[43] Douglas Hanahan and Robert A Weinberg. "Hallmarks of cancer: the next generation". en. In: *Cell* 144.5 (Mar. 2011), pp. 646–674.

[44] Marcus Hjelm, Mattias Höglund, and Jens Lagergren. "New Probabilistic Network Models and Algorithms for Oncogenesis". In: *Journal of Computational Biology* 13.4 (May 2006), pp. 853–865. DOI: 10.1089/cmb.2006.13.853.

[45] Lam Si Tung Ho, F.W. Crawford, and M. A. Suchard. "Direct likelihood-based inference for discretely observed stochastic compartmental models of infectious disease". In: *The Annals of Applied Statistics* 12.3 (2018), pp. 1993–2021. DOI: https://doi.org/10.1214/18-AOAS1141.

[46] W.O. Kermack and A.G. McKendrick. "A Contribution to the Mathematical Theory of Epidemics". In: *Proceedings of the Royal Society of London Series A* 115.772 (1927), pp. 700–721. DOI: https://doi.org/10.1098/rspa.1927.0118.

[47] Yoo-Ah Kim, Dong-Yeon Cho, Phuong Dao, et al. "MEMCover: integrated analysis of mutual exclusivity and functional network reveals dysregulated pathways across multiple cancer types". In: *Bioinformatics* 31.12 (June 2015), pp. i284–i292. DOI: 10.1093/bioinformatics/btv247.

[48] Maren Klever, Peter Georg, Lars Grasedyck, Rudolf Schill, Rainer Spang, and Tilo Wettig. "Low-rank tensor methods for Markov chains with applications to tumor progression models". In: (June 2020). arXiv: 2006.08135 [math.NA].

[49] Mark Leiserson, Dima Blokh, Roded Sharan, et al. "Simultaneous Identification of Multiple Driver Pathways in Cancer". In: *PLoS Computational Biology* 9.5 (May 2013). Ed. by Niko Beerenwinkel, e1003054. DOI: 10.1371/journal.pcbi.1003054.

[50] Alfred James Lotka. *Elements of physical biology*. Williams & Wilkins, 1925.

[51] A.G. McKendrick. "Applications of Mathematics to Medical Problems". In: *Proceedings of the Edinburgh Mathematical Society* 44 (1925), 98â€"130. DOI: https://doi.org/10.1017/S0013091500034428.

[52] Roger McLendon, Allan Friedman, Darrell Bigner, et al. "Comprehensive genomic characterization defines human glioblastoma genes and core pathways". In: *Nature* 455.7216 (Sept. 2008), pp. 1061–1068. DOI: 10.1038/nature07385.

[53] Navodit Misra, Ewa Szczurek, and Martin Vingron. "Inferring the paths of somatic evolution in cancer". In: *Bioinformatics* 30.17 (May 2014), pp. 2456–2463. DOI: 10.1093/bioinformatics/btu319.

[54] C. Moler and C. Van Loan. "Nineteen Dubious Ways to Compute the Exponential of a Matrix, Twenty-Five Years Later". In: *SIAM Review* 45.1 (2003), pp. 3–49. DOI: https://doi.org/10.1137/S00361445024180.

[55] Hesam Montazeri, Jack Kuipers, Roger Kouyos, Jürg Böni, Sabine Yerly, Thomas Klimkait, Vincent Aubert, Huldrych F. Günthard, and Niko Beerenwinkel and. "Large-scale inference of conjunctive Bayesian networks". In: *Bioinformatics* 32.17 (Sept. 2016), pp. i727–i735. DOI: 10.1093/bioinformatics/btw459.

[56] R. M. Neal. "MCMC using Hamiltonian dynamics". In: *Handbook of markov chain monte carlo* 2.11 (2011). DOI: https://doi.org/10.1201/b10905.

[57] P. C. Nowell. "The clonal evolution of tumor cell populations". In: *Science* 194.4260 (Oct. 1976), pp. 23–28.

[58] Jamie Owen, Darren J. Wilkinson, and Colin S. Gillespie. "Scalable inference for Markov processes with intractable likelihoods". In: *Statistics and Computing* 25.1 (Nov. 2014), pp. 145–156. DOI: 10.1007/s11222-014-9524-7.

[59] Daniele Ramazzotti, Giulio Caravagna, Loes Olde Loohuis, et al. "CAPRI: efficient inference of cancer progression models from cross-sectional data". In: *Bioinformatics* 31.18 (May 2015), pp. 3016–3026. DOI: 10.1093/bioinformatics/btv296.

[60] Benjamin J. Raphael and Fabio Vandin. "Simultaneous Inference of Cancer Pathways and Tumor Progression from Cross-Sectional Mutation Data". In: *Journal of Computational Biology* 22.6 (June 2015), pp. 510–527. DOI: 10.1089/cmb.2014.0161.

[61] A. Reibman and K. Trivedi. "Numerical transient analysis of markov models". In: *Computers & Operations Research* 15.1 (1988), pp. 19–36. DOI: https://doi.org/10.1016/0305-0548(88)90026-3.

[62] R. Roylance, P. Gorman, W. Harris, R. Liebmann, D. Barnes, A. Hanby, and D. Sheer. "Comparative genomic hybridization of breast tumors stratified by

histological grade reveals new insights into the biological progression of breast cancer". In: *Cancer Res.* 59.7 (Apr. 1999), pp. 1433–1436.

[63] Kevin Rupp. "Modelling Metastasis in an MHN". master's thesis. University of Regensburg, 2020.

[64] Kevin Rupp. *Accounting for metastasis in the MHN framework*. Workshop Mutual Hazard Networks and more. 2022.

[65] Kevin Rupp, Rudolf Schill, Jonas Süskind, Peter Georg, Maren Klever, Andreas Lösch, Lars Grasedyck, Tilo Wettig, and Rainer Spang. "Differentiated uniformization: A new method for inferring Markov chains on combinatorial state spaces including stochastic epidemic models". In: (Dec. 2021). arXiv: 2112.10971 [stat.ML].

[66] R. Schill, S. Solbrig, T. Wettig, and R. Spang. "Modelling cancer progression using Mutual Hazard Networks". In: *Bioinformatics* 36.1 (2019), pp. 241–249. DOI: https://doi.org/10.1093/bioinformatics/btz513.

[67] Rudolf Schill, Stefan Solbrig, Tilo Wettig, and Rainer Spang. "Modelling cancer progression using Mutual Hazard Networks". In: *Bioinformatics* 36.1 (June 2019). Ed. by Russell Schwartz, pp. 241–249. DOI: 10.1093/bioinformatics/btz513.

[68] Chris Sherlock. "Direct statistical inference for finite Markov jump processes via the matrix exponential". In: *Computational Statistics* 36.4 (Apr. 2021), pp. 2863–2887. DOI: 10.1007/s00180-021-01102-6.

[69] Ewa Szczurek and Niko Beerenwinkel. "Modeling Mutual Exclusivity of Cancer Mutations". In: *PLoS Computational Biology* 10.3 (Mar. 2014). Ed. by Amos Tanay, e1003503. DOI: 10.1371/journal.pcbi.1003503.

[70] Xiaoning Tang, Yongmei Huang, Jinli Lei, Hui Luo, and Xiao Zhu. "The single-cell sequencing: new developments and medical applications". en. In: *Cell Biosci.* 9.1 (June 2019), p. 53.

[71] Franck Toledo and Geoffrey M. Wahl. "MDM2 and MDM4: p53 regulators as targets in anticancer therapy". In: *The International Journal of Biochemistry & Cell Biology* 39.7-8 (July 2007), pp. 1476–1482. DOI: 10.1016/j.biocel.2007.03.022.

[72] Fabio Vandin. "Computational Methods for Characterizing Cancer Mutational Heterogeneity". In: *Frontiers in Genetics* 8 (June 2017). DOI: 10.3389/fgene.2017.00083.

[73] Roel G.W. Verhaak, Katherine A. Hoadley, Elizabeth Purdom, et al. "Integrated Genomic Analysis Identifies Clinically Relevant Subtypes of Glioblastoma Characterized by Abnormalities in PDGFRA, IDH1, EGFR, and NF1". In: *Cancer Cell* 17.1 (Jan. 2010), pp. 98–110. DOI: 10.1016/j.ccr.2009.12.020.

[74] Stefan Vocht. "Identifiability of Mutual Hazard Networks". bachelor's thesis. University of Regensburg, 2022.

[75] Vito Volterra. "Fluctuations in the Abundance of a Species considered Mathematically". In: *Nature* 118.2972 (Oct. 1926), pp. 558–560. DOI: 10.1038/118558a0.

[76] Takuya Watanabe, Sumihito Nobusawa, Paul Kleihues, and Hiroko Ohgaki. "IDH1 Mutations Are Early Events in the Development of Astrocytomas and Oligodendrogliomas". In: *The American Journal of Pathology* 174.4 (Apr. 2009), pp. 1149–1153. DOI: 10.2353/ajpath.2009.080958.

[77] Robert A Weinberg. *The biology of cancer*. 2nd ed. London, England: Garland Science, May 2013.

[78] V. Wolf. "Modelling of Biochemical Reactions by Stochastic Automata Networks". In: *Electronic Notes in Theoretical Computer Science* 171.2 (2007), pp. 197–208. DOI: https://doi.org/10.1016/j.entcs.2007.05.017.

[79] K. Yamanaka, M. Agu, and T. Miyajima. "A Continuous-Time Asynchronous Boltzmann Machine". In: *Neural Networks* 10.6 (1997), pp. 1103–1107. DOI: https://doi.org/10.1016/S0893-6080(97)00006-3.

[80] Chen-Hsiang Yeang, Frank McCormick, and Arnold Levine. "Combinatorial patterns of somatic gene mutations in cancer". In: *The FASEB Journal* 22.8 (Aug. 2008), pp. 2605–2622. DOI: 10.1096/fj.08-108985.

# A. State space restriction for an MHN

When computing the likelihood of observations from Markov Chains on large, combinatorial state spaces, it is often possible to substantially reduce the computational cost by focusing only on the reachable states

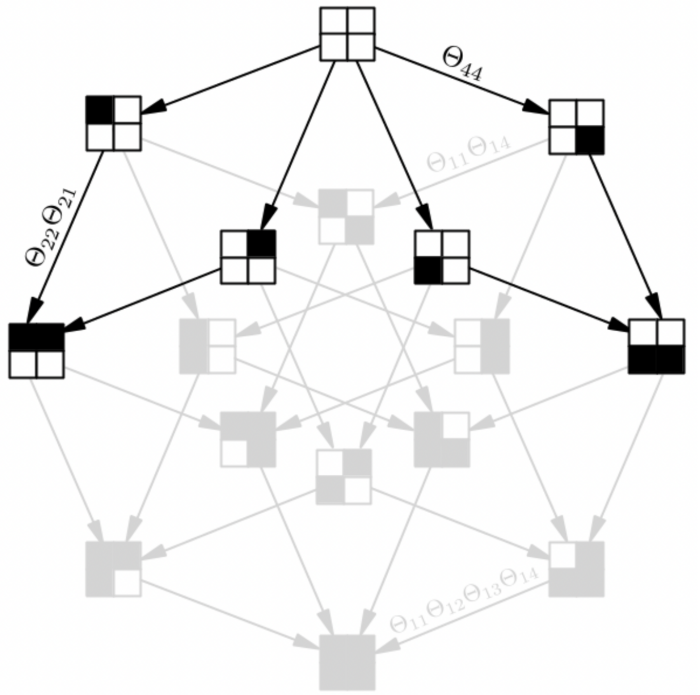
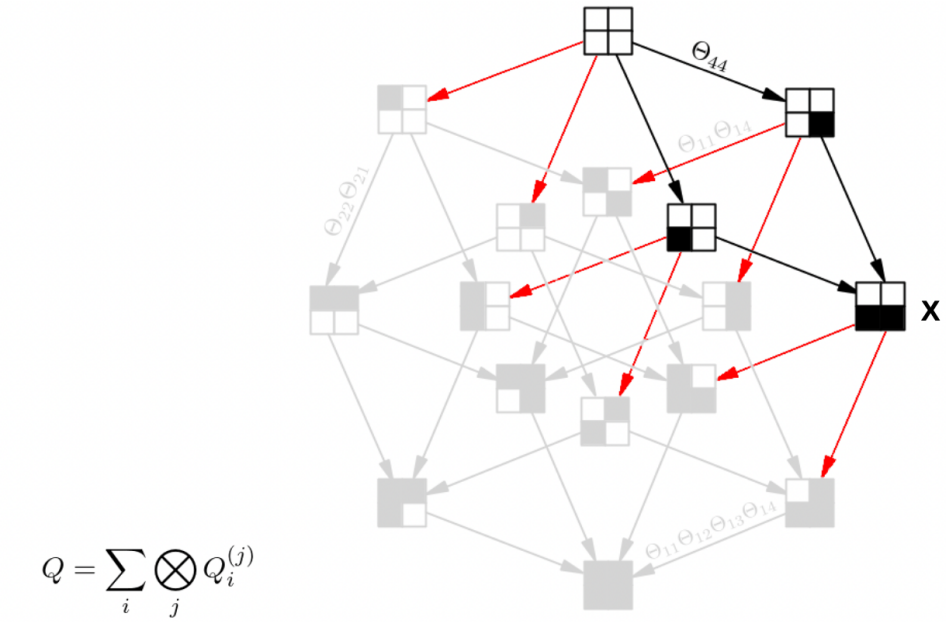



**Figure A.1.:** Motivating example for state space restriction in an MHN. Although the two observed tumors together have all 4 possible events, it is not necessary to compute the likelihood for all $2^4$ states. Since the transition rate matrix $Q$ is lower triangular, the likelihood of each state depends recursively only on those that are spanned from the starting state.




$$Q = \sum_i \bigotimes_j Q_i^{(j)}$$

$$Q_i^{(j)} = \begin{cases} \begin{pmatrix} 1 & 0 \\ 0 & \Theta_{ij} \end{pmatrix} & \text{if } i \neq j \text{ and } x_j = 1 \qquad \begin{pmatrix} 1 \end{pmatrix} \text{ if } i \neq j \text{ and } x_j = 0 \\ \begin{pmatrix} -\Theta_{ij} & 0 \\ +\Theta_{ij} & 0 \end{pmatrix} & \text{if } i = j \text{ and } x_j = 1 \qquad \begin{pmatrix} -\Theta_{ij} \end{pmatrix} \text{ if } i = j \text{ and } x_j = 0 \end{cases}$$

**Figure A.2.:** Formulation of Mutual Hazard Networks with state space restriction for a given observed tumor **x**. The crucial aspect is to include transitions that leave the restricted space (red) in order to "drain probability mass", so that the full model and the restricted model correspond to each other on the restricted space.

# B. State space restriction for SIR

In order to compute the likelihood that an earlier data point $(S, I)$ is followed by a later data point $(S + \Delta S, I + \Delta I)$ after time $t$, it is sufficient to compute $\mathbf{p}(t)$ on a small subset of the entire state space. Since the number of susceptibles cannot increase ($\Delta S \leq 0$) and the number of recovered cannot decrease ($\Delta R = -\Delta S - \Delta I \geq 0$), all possible trajectories from $(S, I)$ to $(S + \Delta S, I + \Delta I)$ must necessarily stay within the restricted state space

$$\{S_{\min}, \ldots, S_{\max}\} \times \{I_{\min}, \ldots, I_{\max}\},$$

where

$$\begin{aligned} S_{\min} &= S + \Delta S, & S_{\max} &= S, \\ I_{\min} &= I - \Delta R, & I_{\max} &= I - \Delta S. \end{aligned} \tag{B.1}$$

All probability mass that leaves this space must be accounted for, but we do not need to keep track of its destination. To this end, we introduce the modified band matrices

$$\begin{aligned} \tilde{\mathcal{S}}^{+}_{\text{inf}} &= \text{superdiag}(S_{\min} + 1, \ldots, S_{\max}) & \tilde{\mathcal{I}}^{+}_{\text{inf}} &= \text{subdiag}(I_{\min}, \ldots, I_{\max} - 1) \\ \tilde{\mathcal{S}}^{-}_{\text{inf}} &= \text{diag}(S_{\min}, \ldots, S_{\max}) & \tilde{\mathcal{I}}^{-}_{\text{inf}} &= \text{diag}(I_{\min}, \ldots, I_{\max}) \\ \tilde{\mathcal{S}}^{+}_{\text{rec}} &= \text{diag}(1, 1, \ldots, 1) = \mathrm{I} & \tilde{\mathcal{I}}^{+}_{\text{rec}} &= \text{superdiag}(I_{\min} + 1, \ldots, I_{\max}) \\ \underbrace{\tilde{\mathcal{S}}^{-}_{\text{rec}} = \text{diag}(1, 1, \ldots, 1) = \mathrm{I}}_{(|\Delta S| + 1) \times (|\Delta S| + 1)} & & \underbrace{\tilde{\mathcal{I}}^{-}_{\text{rec}} = \text{diag}(I_{\min}, \ldots, I_{\max}).}_{(|\Delta S| + \Delta R + 1) \times (|\Delta S| + \Delta R + 1)} & \end{aligned} \tag{B.2}$$

and define a smaller transition-rate matrix on the restricted state space as

$$\tilde{Q} = \frac{\beta}{N}(\tilde{\mathcal{S}}^{+}_{\text{inf}} \otimes \tilde{\mathcal{I}}^{+}_{\text{inf}}) + \alpha(\tilde{\mathcal{S}}^{+}_{\text{rec}} \otimes \tilde{\mathcal{I}}^{+}_{\text{rec}}) - \frac{\beta}{N}(\tilde{\mathcal{S}}^{-}_{\text{inf}} \otimes \tilde{\mathcal{I}}^{-}_{\text{inf}}) - \alpha(\tilde{\mathcal{S}}^{-}_{\text{rec}} \otimes \tilde{\mathcal{I}}^{-}_{\text{rec}}) \tag{B.3}$$

with derivatives

$$\frac{\partial \tilde{Q}}{\partial \log \alpha} = \alpha(\tilde{\mathcal{S}}^{+}_{\text{rec}} \otimes \tilde{\mathcal{I}}^{+}_{\text{rec}}) - \alpha(\tilde{\mathcal{S}}^{-}_{\text{rec}} \otimes \tilde{\mathcal{I}}^{-}_{\text{rec}}), \tag{B.4}$$

$$\frac{\partial \tilde{Q}}{\partial \log \beta} = \frac{\beta}{N}(\tilde{\mathcal{S}}^{+}_{\text{inf}} \otimes \tilde{\mathcal{I}}^{+}_{\text{inf}}) - \frac{\beta}{N}(\tilde{\mathcal{S}}^{-}_{\text{inf}} \otimes \tilde{\mathcal{I}}^{-}_{\text{inf}}). \tag{B.5}$$

Note that the columns of $\tilde{Q}$ sum to less than zero and that $\mathbf{p}(t)$ therefore sums to less than 1 on the restricted state space. Computing matrix-vector products using these operators has a time complexity in $\mathcal{O}(|\Delta S|^2 + |\Delta S|\Delta R)$.

The largest absolute diagonal entry of $\tilde{Q}$ is

$$\gamma = \max_x |\tilde{Q}_{x,x}| = \frac{\beta}{N} S_{\max} I_{\max} + \alpha I_{\max} \tag{B.6}$$

with derivatives

$$\frac{\partial \gamma}{\partial \log \alpha} = \alpha I_{\max}, \tag{B.7}$$

$$\frac{\partial \gamma}{\partial \log \beta} = \frac{\beta}{N} S_{\max} I_{\max}. \tag{B.8}$$

We perform $m$ iterations of algorithm 2 such that the entire probability mass (including that which left the restricted state space) according to eq. (5.8) reaches the required tolerance. Hence, the overall time complexity of the algorithm is

$$\mathcal{O}\big(\gamma(|\Delta S|^2 + |\Delta S|\Delta R)\big) = \mathcal{O}\big(I_{\max}(|\Delta S|^2 + |\Delta S|\Delta R)\big) = \tag{B.9}$$

$$\mathcal{O}\big((I + |\Delta S|)(|\Delta S|^2 + |\Delta S|\Delta R)\big). \tag{B.10}$$

Storing the result $\mathbf{p}(t)$ has complexity $\mathcal{O}(|\Delta S|^2 + |\Delta S|\Delta R)$.

# C. Computing the gradient of the score of an MHN

$$
\begin{aligned}
\frac{\partial \mathcal{S}_{\mathcal{D}}}{\partial \theta_{ij}} &= \frac{\partial \mathcal{S}_{\mathcal{D}}}{\partial \left[I-Q\right]^{-1}} \cdot \frac{\partial \left[I-Q\right]^{-1}}{\partial \theta_{ij}} \\
&= \frac{\mathbf{p}_{\mathcal{D}}}{\mathbf{p}_{\theta}} \mathbf{p}_0^T \cdot \left( - \left[I-Q\right]^{-1} \frac{\partial \left[I-Q\right]}{\partial \theta_{ij}} \left[I-Q\right]^{-1} \right) \\
&= \underbrace{- \left( \frac{\mathbf{p}_{\mathcal{D}}}{\mathbf{p}_{\theta}} \right)^T \left[I-Q\right]^{-1}}_{:=\mathbf{q}} \frac{\partial \left[I-Q\right]}{\partial \theta_{ij}} \underbrace{\left[I-Q\right]^{-1} \mathbf{p}_0}_{=\mathbf{p}_{\theta}} \\
&= -\mathbf{q} \frac{\partial \left[I-Q\right]}{\partial \theta_{ij}} \mathbf{p}_{\theta} \\
&= \mathbf{q} \frac{\partial Q}{\partial \theta_{ij}} \mathbf{p}_{\theta}
\end{aligned}
$$

$\mathbf{q}$ and $\mathbf{p}_\theta$ are computed once per gradient step and are constant for all $i, j \in \{1, ..n\}$. See Figure C.1 for an illustration of $\partial Q / \partial \theta_{ij}$, whose structure we can use to compute all $\partial \mathcal{S}_{\mathcal{D}} / \partial \theta_{ij}$ for each $i$ in one go:

**For each $i$ do:**

$$
\mathbf{r} \leftarrow \mathbf{q} \odot \left( \frac{\partial Q}{\partial \theta_{ii}} \mathbf{p}_{\theta} \right) \text{ (hadamard product)}
$$

$\frac{\partial \mathcal{S}_{\mathcal{D}}}{\partial \theta_{ii}}$ is then the sum over all entries in $\mathbf{r}$. $\frac{\partial \mathcal{S}_{\mathcal{D}}}{\partial \theta_{ij}}$ is then the sum over all entries in $\mathbf{r}$ where the corresponding state has event $x_j = 1$.

For example for $n = 3$, $i = 1$:

$$\frac{\partial \mathcal{S}_{\mathcal{D}}}{\partial \theta_{11}} = \sum \begin{bmatrix} 000 \\ 100 \\ 010 \\ 110 \\ 001 \\ 101 \\ 011 \\ 111 \end{bmatrix}, \frac{\partial \mathcal{S}_{\mathcal{D}}}{\partial \theta_{12}} = \sum \begin{bmatrix} 000 \\ 100 \\ 010 \\ 110 \\ 001 \\ 101 \\ 011 \\ 111 \end{bmatrix}, \frac{\partial \mathcal{S}_{\mathcal{D}}}{\partial \theta_{13}} = \sum \begin{bmatrix} 000 \\ 100 \\ 010 \\ 110 \\ 001 \\ 101 \\ 011 \\ 111 \end{bmatrix}$$

$$\frac{\partial Q}{\partial \theta} = \begin{bmatrix} \partial Q/\partial \theta_{11} & \partial Q/\partial \theta_{12} & \partial Q/\partial \theta_{13} \\ \partial Q/\partial \theta_{21} & \partial Q/\partial \theta_{22} & \partial Q/\partial \theta_{23} \\ \partial Q/\partial \theta_{31} & \partial Q/\partial \theta_{32} & \partial Q/\partial \theta_{33} \end{bmatrix} =$$

$$\begin{pmatrix} -e^{\theta_{11}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ e^{\theta_{11}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -e^{\theta_{11}+\theta_{12}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & e^{\theta_{11}+\theta_{12}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{11}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & e^{\theta_{11}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{11}+\theta_{12}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{11}+\theta_{12}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -e^{\theta_{11}+\theta_{12}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & e^{\theta_{11}+\theta_{12}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{11}+\theta_{12}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{11}+\theta_{12}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -e^{\theta_{11}+\theta_{12}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & e^{\theta_{11}+\theta_{12}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{11}+\theta_{12}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{11}+\theta_{12}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

$$\begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{11}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & e^{\theta_{11}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{11}+\theta_{12}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{11}+\theta_{12}+\theta_{13}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

$$\begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -e^{\theta_{22}+\theta_{21}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & e^{\theta_{22}+\theta_{21}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & e^{\theta_{22}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{22}+\theta_{21}+\theta_{23}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

$$\begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -e^{\theta_{33}+\theta_{31}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{33}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{33}+\theta_{31}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{33}+\theta_{31}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -e^{\theta_{33}+\theta_{31}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{33}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 0 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{33}+\theta_{31}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{33}+\theta_{31}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix} \begin{pmatrix} -e^{\theta_{33}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ e^{\theta_{33}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & -e^{\theta_{33}+\theta_{31}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & e^{\theta_{33}+\theta_{31}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & -e^{\theta_{33}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & e^{\theta_{33}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & -e^{\theta_{33}+\theta_{31}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & e^{\theta_{33}+\theta_{31}+\theta_{32}} & \cdot & \cdot & \cdot & \cdot & \cdot \end{pmatrix}$$

**Figure C.1.:** Illustration of the gradient of $Q$ for $n = 3$.