Causal Models over Infinite Graphs and their Application to the Sensorimotor Loop - General Stochastic Aspects and Gradient Methods for Optimal Control

Der Fakultät für Mathematik und Informatik
der Universität Leipzig
eingereichte

DISSERTATION

zur Erlangung des akademischen Grades

DOCTOR RERUM NATURALIUM
(Dr.rer.nat.)

im Fachgebiet

Mathematik

vorgelegt

von Diplomphysiker Holger Bernigau
geboren am 12.08.1983 in Potsdam (Deutschland)

Die Annahme der Dissertation wurde empfohlen von:

1. Professor Dr. Shun-ichi Amari (Saitama, Japan)
2. Professor Dr. Nihat Ay (Leipzig)

Die Verleihung des akademischen Grades erfolgt mit Bestehen
der Verteidigung am 07.04.2015 mit dem Gesamtprädikat magna cum laude.

To my great love, Ying

Bibliographische Daten

Causal Models over Infinite Graphs and their Application to the Sensorimotor Loop - General Stochastic Aspects and Gradient Methods for Optimal Control
(Kausale Modelle über unendlichen Grafen und deren Anwendung auf die sensomotorische Schleife - stochastische Aspekte und gradientenbasierte optimale Steuerung)
Bernigau, Holger
Universität Leipzig, Dissertation, 25.07.2014
175 Seiten, 54 Abbildungen, 197 Referenzen

Introduction and summary of results

The enormous amount of capabilities that every human learns throughout his life, is probably among the most remarkable and fascinating aspects of life. Learning has therefore drawn lots of interest from scientists working in very different fields like philosophy, biology, sociology, educational sciences, computer sciences and mathematics. We will focus on the mathematical and information theoretical aspects of learning within this thesis.

We are interested in the learning process of an agent (which can be for example a human, an animal, a robot, an economical institution or a state) that interacts with its environment. The formulation of a learning problem in the sensorimotor loop is far from being obvious, since the dynamic of the process (i.e. the distribution of the sensor values, memory values, action values etc.) depends crucially on the learning algorithm used by the agent. The definition of the problem must therefore include the complex dependencies within this learning process:

- The agent's actions depend on the sensor input, since this is the only source of information about the environment that is accessible to the agent;

- future sensor input depends on current actions, if for example a young dog nibbles his owner's favorite shoes he is likely to receive a high volume signal with his ears in near future;

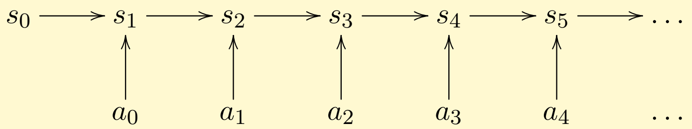
- adjustable learning parameters and memory values might complicate this causal structure further.

This thesis is organized in two parts. In Part I we revisit and develop the mathematical framework and in Part II we use this framework to formulate and solve a general class of learning problems in the sensorimotor loop.

Motivation and state of the art for Part I

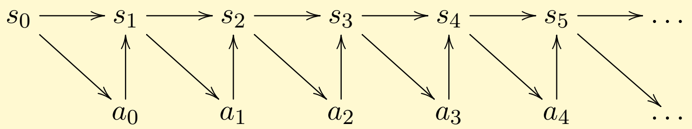
Part I is organized in two chapters that together provide the theoretical foundation for Part II. Chapter 1 is dedicated to a proper stochastic description of an agent interacting with its environment. The description must finally contain both, causal dependencies within the agent-environment system that restrict the dynamic and degrees of freedom in the system that can be adjusted in a learning process. An appropriate mathematical tool to describe causal dependencies and learnable degrees of freedom is the well-known theory of causal networks (also known as Bayesian networks or graphical models, see for example Lauritzen [114], Murphy [133] or Rückert et al. [158]). Unfortunately the well-known theory is not immediately applicable in our setup, since we are interested in the process over an infinite time horizon. This is why we provide a generalization of the known results from the theory of graphical models in Chapter 1. The theory developed in Chapter I can also be seen as a generalization of discrete-time Markov-processes to processes with more involved causal dependencies.

In order to illustrate the idea of graphical models and its relation to the sensorimotor loop we will discuss the sensor-action process in a reinforcement learning setup (compare for example Sutton and Barto [179]). Assume an agent to receive a sensor value s_i at time i and to perform an action a_i (that we assume to be independent of the former sensor values and actions at this point). Both the sensor value, s_i, and the action, a_i causally influence the next sensor value s_{i+1}. These dependencies can be illustrated by the following causal graph:



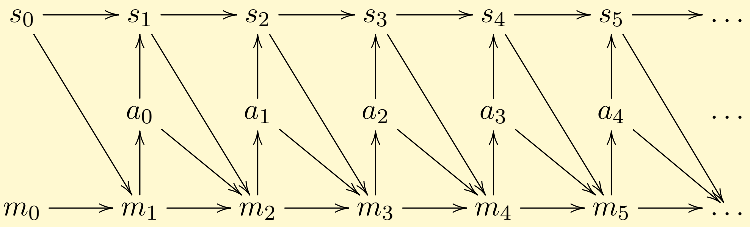
Caus. mod. 1 - *Causal structure of a simple open-loop controlled MDP*

A graphical model roughly speaking consists of a graph indicating the causal dependencies and a probability distribution that is compatible with this graph. The learnable degrees of freedom mentioned above are given by the free actions, a_i. The causal dependencies (and the resulting dynamic) of the state-action process change significantly when the process is altered into a closed-loop dynamic. Formally a feedback can be included by changing the causal graph, Caus. mod. 1, into the following one:



Caus. mod. 2 - *Causal structure of a simple closed-loop controlled MDP*

The resulting process is a controlled version of the former one. By this we mean that the probabilistic transition rules from the pair (s_i, a_i) to s_{i+1} remains exactly the same as before. On the level of graphs this requires, that the open-loop controlled MDP is a subgraph of the closed-loop MDP with the further property that new arrows either point to newly introduced vertices or to input vertices of the former graph (by input vertices we mean vertices without parents). There exist other controlled dynamics over the graph Caus. mod. 1 that are of great relevance for practical applications. One extension is what we will refer to as (simple) sensorimotor loop. It describes an agent, that receives a sensor value, updates its memory and reacts to the state of its memory. Therefore we introduce some memory values, m_i, and assume the following causal relationship:



Caus. mod. 3 - *Causal structure of a simple sensorimotor loop*

Once a mathematical model for the dynamic of the agent-world system is settled, the learning objective has to be formulated and appropriate solution algorithms have to be introduced. Mathematically this issue is commonly known as optimization theory (or optimal control, if the optimization variables are paths subject to dynamical constraints). Therefore Chapter 2 is devoted to optimization theory and stochastic gradient algorithms (for good references see for example Clarke [50], Jahn [93], Aubin [7], Aubin and Frankowska [8], Anger, Aubin, and Cellina [6], Troutman [185], Troutman [185], Borkar [38], Bharath and Borkar [25], Kushner and Clark [109] and Kushner and Yin [110]). Since we will finally optimize over compact constraint sets, we focus on projected stochastic gradient algorithms as a tool to find stationary points of an optimization problem in a stochastic setup.

Summary and results from Chapter 1

In order to specify a model of an agent interacting with the environment, the following questions have to be answered:

- Which observables exist and which of them are adjustable?
- How are these observables causally related over an infinite time-horizon?
- What is the stochastic dynamic of these variables, if the adjustable input variables are fixed before the process starts (by stochastic dynamic we mean a probability distribution on the collection of all variables)?
- What is the stochastic dynamic , if the adjustable input variables depend causally on previous observables and maybe on some new memory variables?

These questions can be modelled using the theory of causal models as known from the machine learning literature (see for example Lauritzen [114], Murphy [133] or Rückert et al. [158]). In order to describe an agent interacting with the environment over an infinite time horizon there remain two challenges:

- The theory is needed for infinite graphs.
- Continuous state spaces and deterministic transition laws should be permitted, such that the standard construction of the theory of directed models as a special case of the theory for undirected models does not work.

First of all an appropriate class of graphs needs to be defined that is on the one hand powerful enough to contain at least the sensorimotor loop and on the other hand should be restrictive enough to yield a convenient mathematical theory. We introduce an appropriate class of graphs that satisfies these requirements. We will refer to this class of graphs as recursively constructible graph (see Definition 1.1.1, especially Definition 1.1.1). We illustrate this class of graphs by appropriate examples and counterexamples and proceed with a proof of existence of a unique process law (i.e. a probability distribution on the space of possible configurations over this graph) that is compatible with a given configuration of initial vertices and the causal transition rules in Theorem 1.2.1. The main results from this chapter are two conditional independence results for directed graphical models (Theorem 1.3.1 and Theorem 1.4.1). Stochastic independence in graphical models over finite graphs has been investigated intensively. A fundamental result has been proven by Hammersley and Clifford in 1971 (see Hammersley and Clifford [80]), showing that a probability distribution with positive density satisfies some Markov property with respect to a given finite graph if and only if it factorizes over the cliques of the graph. There are at least three different reasonable versions of Markov properties in undirected graphs, namely the pairwise Markov property, the local Markov property and the global Markov property - see for example Lauritzen [114]. The Hammersley-Clifford theorem immediately gives a similar statement for finite directed graphs with positive density. We go beyond the known results in three main aspects:

- We consider causal models over infinite graphs (this is unavoidable to describe a Markov decision process or the sensorimotor loop over infinitely many time steps, which is again necessary to investigate stochastic convergence in these models).
- For Theorem 1.3.1 we provide an alternative proof for the relation between graph separation and conditional independence. Our proof needs less technical prerequisites and is therefore applicable in more general cases (unlike the classical result, as provided in Lauritzen [114] for example, we do not need the existence of a positive density. Our result includes continuous state spaces and deterministic transition rules for example - hence all statements remain true for (controlled) deterministic dynamical systems with random initial values).
- In Theorem 1.4.1 we formulate and prove strong conditional independence properties, i.e. conditional independence results for randomly chosen vertex sets. This is essential to understand certain conditional independence statements that we need in Chapter 4. To our knowledge strong conditional independence properties have not

been considered systematically in a causal model setup so far. Even for the theory of ordinary Markov processes, where we borrowed the terminology from, our result extends the common statements (see Kallenberg [99], Rogers and Williams [154] or Bauer [22]) in a non-trivial way: Our theorem allows conditional independence statements given certain pairs of stopping times for example (compare Example 1.4.2).

The second point is worth a further comment. It is a well-known fact that the existence of a positive density (or another slightly weaker constraint) is crucial for the Hammersley-Clifford theorem to hold true. There exist examples of probability distributions showing that the three Markov properties mentioned above are actually different in general (see Lauritzen [114] for a good summary). Directed models on the other hand, under very mild technical restrictions, always satisfy the strongest Markov property with respect to the moral graph (for a clarification of these concepts see Section 1.3 in Chapter 1). The existence of a positive density is usually not satisfied in these models and henceforth nothing can be factorized over the cliques. The recursive construction of the probability law from appropriate transition kernels (similar to the standard construction in the theory of finite time Markov processes) assures the desired conditional independence results.

We think that causal models on recursively constructible graph, and especially the (strong) independence results offer a very convenient technical tool to investigate the stochastic properties of systems with Markovian dependencies with and without controls.

From a practical prospective, the first chapter offers a clear language to define what is meant by convergence of a learning algorithm in a non-IID setup, where it is well-possible that no state-action pair is observed more than once, that no pair of sensor values (or actions) have identical distributions and that no pair of variables are stochastically independent. The definitions and the conditional independence results also simplify many arguments in consecutive chapters of this thesis. Structural robustness is most naturally described in the language of statistics. A statistical model is commonly known as a parameterized (or non-parameterized) family of probability measures on a given probability space. In the case of graphical models the probability space consists of all configurations on the graph equipped with an appropriate σ-algebra . Any collection of free initial values defines a unique process law and a collection of different initial configurations therefore naturally defines a statistical model over the causal model. Structural robustness of some property then means that this property holds for any probability measure in the given family. In Definition 1.2.2 we therefore also define the concept of a statistical model over a causal model. The statistical language is also very useful for the formulation of learning problems - it is usually desirable that a learning algorithm converges for all possible laws from a given family.

Summary and results from Chapter 2

Chapter 2 introduces some mostly well-known concepts and terminology from theoretical optimization theory and set-valued analysis. The main ideas and results from this chapter are:

- We suggest to back-project onto the constraint set using quasi-projectors different from the Euclidean best-approximation (for a definition see for example: Aubin [7]). The problem with general quasi-projectors is that they might introduce spurious stationary points at the boundary. We will show that this can be compensated by using an appropriate metric for the gradient ascent. As a possible application we will describe a suitable quasi-projector and a compatible metric for an optimization problem over the unit ball in the set of matrices equipped with operator norm (see: example Example 2.3.2 - an example that we will built upon in Chapter 4).

- We provide a detailed proof for the asymptotic behavior of an iterative stochastic approximation sequence that is back-projected onto the constraint set by a general quasi-projector. (see: Theorem 2.4.1)

- We give a proof of convergence for a stochastic gradient algorithm with back-projection by a general quasi-projector where the gradient ascent is performed with respect to some possibly discontinuous metric (or several metrics where the actual choice at each step depends on the history, see Theorem 2.4.3).

Motivation and state of the art for Part II

The second part of this thesis deals with application of the theory of causal models on infinite graphs to learning problems in the sensorimotor loop. This topic is strongly inspired by current developments in robotics and artificial intelligence. Therefore we will provide an overview about recent developments and targets in this field before we continue with a short summary of the last two chapters of this thesis.

The construction of robots that are able to perform more than a little number of highly-specialized tasks remains a very challenging problem. It is practically impossible to specify all relevant situations that the robot might ever encounter during interaction with its environment and to implement all possible reactions right from the beginning into the software of the robot. Current research in robotics usually tackles this problem by leaving non-fixed parameters in the software architecture of the robot and by applying sophisticated methods from machine learning and/or (stochastic) optimal control theory to learn these parameters. Examples for robotic tasks that have attracted lots of interest include table tennis (see for example Muelling, Kober, and Peters [131] or Muelling et al. [132] for two recent papers containing a good overview over the subject), baseball (see Peters and Schaal [142] or Senoo et al. [169]). Another important application is grasping (see Peters and Schaal [152] for application of Gaussian process implicit shape potential to robot grasping or see the recent paper Dragiev, Toussaint, and Gienger [66] that applies the approximate inference control framework to a novel coordinate representation of physically stable grasps). Typically the degree of freedom of the robot exceeds the dimension of the physical space by far. Therefore target positions can usually be reached in many different ways. Coping with this redundancy is known as operational state control , which is another important research field in robotics (see for example Nakanishi et al. [182] or Zarubin et al. [197] for an approach to operational space control using reward weighted regression). The paradigm shift away from robots with a totally pre-implemented software towards partially self-learning robots came along with another important insight: It is well possible that the robot can operate successfully without having an explicit model of the environment and its own interaction with the environment. For many tasks this explicit model representation within the robot is superfluous, since the relevant physical constraints are already encoded in the physical laws that the robot is automatically subjected to. This phenomena and related ideas are known as embodiment. A very famous example for embodiment is the passive walker (see also example Collins, Wisse, and Ruina [53], and Hoffmann and Pfeifer [87] for a recent case study of embodiment in robotics and biology).

A significant portion of the literature in robotics deals with an appropriate representation of the state spaces, the action spaces and appropriate representations of the policies. Common representation include spline-based approaches (see Miyamoto et al. [129]), an encoding by appropriate dynamical systems (see Ijspeert, Nakanishi, and Schaal [116] and Paraschos et al. [139] for a probabilistic version) and many other task-dependent representations (see Zarubin et al. [197] for different representation with a special focus on grasping for example). The construction of good policy search algorithms is very actively investigated in robotics (see Deisenroth, Neumann, and Peters [60] for a recent survey on this topic).

There has also been lots of successful attempts in applying methods from reinforcement learning (see for example Kaelbling, Littman, and Moore [98] or Sutton and Barto [179]) to problems in robotics. Our work focuses on the agent-environment interaction, and is therefore strongly connected to the theory of reinforcement learning. A key concept in this field is the Markov decision problem. A Markov decision problem usually consists of two ingredients, a stochastic model describing how the upcoming sensor values are influenced by current actions (see Caus. mod. 2 and Caus. mod. 1) and an optimization problem. The optimization problem is the maximization of the expectation of a reward function that maps the state-action trajectories to a real number (see Eugene and Feinberg [71] for a good reference). The value of the expected reward can be modified via the policy, i.e. the way of choosing new actions. Important instances of expected reward problems that have been discussed extensively in the literature include the immediate rewards, discounted rewards and average rewards (see Eugene and Feinberg [71] for a good overview). In reinforcement learning some parts of the input data for the Markov decision problem are unknown (this might include the reward function and/or the transition probabilities). Most of the policy optimization algorithms use gradient ascent methods (see for example Peters and Schaal [143], Sutton et al. [180] or Peters, Mülling, and Altün [145]) or some EM-inspired algorithms (see Murphy [133]) applied to an equivalent statistical inference problem (see for example Kober and Peters [103], Rawlik, Toussaint, and Vijayakumar [150], Vlassis et al. [189], Toussaint and Goerick [183]).

A very important recent target in robotics is to increase the level of autonomy. Present robots are still far from interacting with the world, gathering knowledge with relevance for different tasks and successfully extracting appropriate behavorial routines for the current situation automatically. One problem lies in a special feature of many algorithms from machine learning and statistics: they strongly rely on the observation of IID samples. In many tasks in robotics this is reached by repeating the following steps:

- put the robot into a well-defined initial state with some well-defined set of parameters;
- perform a roll-out;
- evaluate the roll-out, return to the initial position and restart the procedure (maybe with changed parameters).

For a step towards more autonomy it would be desirable to let the robot follow a certain trajectory without the reset step. Any learning algorithm then necessarily requires a clear model about the influence of the robot on the environment and vice versa. This is very different from the situation in standard computing. In standard computing the machine transfers input into output and does usually not have to care about the influence of its output on future input. In robotics, however, this influence is crucial (as for example an instable position might cause the robot to fall down and break etc.).

The search for a way to make robots more autonomous also led to reinforcement learning algorithms with non-reward like objective functions, like the predictive information (see Ay et al. [17], Zahedi, Ay, and Der [196] and Ay et al. [16]). The underlying idea is that learning of a robot (or another agent) might be driven by a comparatively simple information theoretical principle. A maximization of the predictive information is one such principle. The predictive information has the appealing property that it is a compromise between a high entropy on the one hand and a very coordinated transition on the other hand. Applications to physically realistically simulated robots indeed show a very interesting coordinated behavior of the robot (see Ay et al. [17], Zahedi, Ay, and Der [196]) and might therefore be an appropriate ingredient to guide the exploration of a self-learning robot.

The idea of maximizing information measures is tightly related to another important cornerstone of modern research in artificial intelligence, theoretical biology, theoretical neurosciences, coding theory etc.: information theory (for good monographs on information theory and information geometry see Amari [2], Amari, Nagaoka, and Harada [5], Cover and Thomas [54], Csiszár and Korner [58], Shannon [170] and Liese and Vajda [118]). Beside their theoretical importance, information theoretical insights can often improve known algorithms. One example for this is the natural gradient. To our best knowledge, the idea of using the so-called Fisher metric in a gradient ascent algorithms in the parameter space of a statistic models dates back to Amari (see for example Amari [3] and Amari and Douglas [4]). The natural gradient outperforms the plain-vanilla gradient (which is the gradient with the underlying metric being the standard Euclidean metric on the chosen parameter set) in many applications. An intuitive explanation is that the Fisher-metric is associated to statistical properties of the model (it lower bounds the variance of an unbiased estimator in the Cramer-Rao inequality for example, see Amari, Nagaoka, and Harada [5]), whereas the Euclidean metric in parameter space is rather arbitrary and highly depends on a good choice of parameters. A problem with plain-vanilla gradients is often that they do not relate to the probabilistic properties of the statistical model. In Peters and Schaal [143] the authors give an example where the plain-vanilla gradient suppresses exploration too quickly, whereas the Fisher gradient performs reasonably well. The issue has also been addressed in the recent paper Peters, Mülling, and Altün [145] where the authors suggest to update the parameters with bounded information loss (measured by the Kullback-Leibler divergence between the observed empirical distribution of state-action pairs and the distribution resulting from the new policy).

The causal models over infinite graphs and the language developed in the first chapter of this thesis provide an appropriate tool for describing an agent interacting with the environment and rigorously defining reinforcement learning problems with non-reward-like objective functions as statistical problems. This is what we do in the second part of this thesis.

Summary and results from Chapter 3

In Chapter 3 we describe our model of an agent interacting with the environment with and without learning. The concepts originate from the theory of Markov decision problems and will appear familiar to everyone working in related fields. What differs from most definitions is that we split the dynamical part of a Markov decision problem from the optimization part. The reason is twofold. First of all we want to define our model rigorously, meaning that we want to have a clearly defined probabilistic model of the process for different exterior parameters and different learning algorithms. For the convergence proofs to follow in Chapter 4 we need a clear language for what is actually mean by learning - it is the convergence of the learning algorithm to a specified context-dependent optimizer for every measure in the given statistical model. There exists a collection of world parameters, each of which determines a well-defined process law for a given learning algorithm. We want that the agent reaches to find the right, possibly parameter-dependent learning objective with probability one. A rigorous definition of the stochastic model of an agent interacting with the environment needs a certain amount of care and further explanation. The second reason for a separation of dynamical model and the optimization problem is that we will optimize more general process functionals than the expected reward.

The main results from Chapter 3 are a clarification of the models to be considered in Chapter 4 (see section 3.1), the introduction of a reinforcement learning problem that goes beyond expected reward maximization (see: Problem 3.2.1 and the special instances Problem 3.2.2 and Problem 3.2.3), a presentation and motivation of numerous interesting instances of this problem with a collection of relevant policy gradient formulas (see section 3.3). We prove the relation between discounted functionals of the process law (a generalization of the discounted expected reward) and their ergodic counterparts (a generalization of the long-time-average reward) for finite state space models (see Theorem 3.2.1). The gradient formulas use some results from perturbation theory for finite state Markov chains. The idea to apply this theory to gradient ascent algorithms in machine learning is also new to our knowledge.

Summary and results of Chapter 4

In Chapter 4 we finally prove convergence of a collection of learning algorithms that are supposed to find local optima for the functionals discussed in Chapter 3. Applying this to the predictive information we are able to improve an algorithm given in Zahedi, Ay, and Der [196] and to show its convergence for the first time. We start with a convergence result for finite state spaces and proceed with a general convergence result. The general result is based on a (in an appropriate sense) consistent estimator of the world parameters and the existence of a suitable back-projection onto the constraint set. The proofs and the algorithms strongly depend on the definitions and results in Chapter 2. For finite state space MDPs and for linear Gaussian MDPs we explicitly provide these tools, such that the algorithm can be implemented with a computer without further theoretical work. The main outcomes of this section are Theorem 4.1.1, Theorem 4.2.1 and the application to linear Gaussian dynamics in section4.3

Concluding remarks

We have collected several mathematical preliminaries from probability theory, differential geometry of submanifolds of $\mathbb{R}^n$ and information theory in the Appendix.

To improve readability we have collected a list of own definitions or definitions that differ slightly from their common usage in a glossary. We have also included a list of abbreviations and a symbol index.

Acknowledgments

I would like to thank Prof. Dr. Nihat Ay for supervising my thesis. Beside the stimulating scientific discussions he was always very supportive and I strongly benefited from his very helpful hints to structure my thoughts and ideas in a clear and transparent way.

I would also like to thank Dr. Johannes Rauh and Dr. Arleta Szkoła for the fruitful discussions and the scientific feedback. I would also like to thank Prof. Dr. Ralf Der, who had many helpful hints for me, and Dr. Keyan Ghazi-Zahedi for their great collaborations.

In general I would like to thank all members of the research groups of Prof. Dr. Jürgen Jost and of the research group of Prof. Dr. Nihat Ay. They all made my stay at the Max Planck Institute for Mathematics in the Sciences a wonderful experience. I enjoyed the many discussions and activities with Dr. Eckehard Olbrich, Dr. Nils Bertschinger, Martin Kell, Wiktor Młynarski, Dr. Georg Martius, Özge Erdem, Dr. Pierre Yves Bourguignon, Stephan Poppe, Dr. Stephan Weis, Dr. Felix Effenberger, Dr. Oliver Pfante, Philipp Benner, Philipp-Jens Ostermeier, Yangjing Long, Dr. Timo Ehrig, Enno Keßler, Dr. Guido Montúfar and many more.

I would also like to thank the International Max Planck Research School (IMPRS) for providing the funding and creating a very good academic environment for my Ph.D. project, organizing work shops and interesting lecture series. Moreover I would like to thank the DFG for funding the Priority Program Autonomous Learning, that I was a member of. I gained many interesting stimulating insights from the symposiums provided by this program.

Furthermore I would like to thank the administration and the staff of the Max Planck Institute for Mathematics in the Sciences for creating a perfect background for the scientific work. I especially would like to thank Antje Vandenberg for her great support in all administrative issues.

Last but not least I would like to take the opportunity to thank some people for their private support. I would like to thank my parents, my grandparents and my sister for their great support during my entire life. Finally I would like to thank my beloved wife, Ying, for being a great enrichment for my life, for her great support and understanding especially in the final period of my Ph.D. project. Besides she also provided many useful work-related comments and had very good suggestions for improvements in style and presentation.

Contents

Part I

Mathematical fundamentals

Chapter 1

Causal models on recursively constructible graphs

In the first section we introduce a class of stochastic models, that we will refer to as causal models in this thesis. On the one hand this class of models is broad enough to capture a huge variety of different models, on the other hand this class is narrow enough to posses some interesting non-trivial properties. For a mathematical treatment of causal models we combine some concepts from probability theory, namely from the theory of Markov chains (with general state space) with the theory of causal models as it is known in the machine learning community.

In machine learning causal models are a frequently used tool (compare Murphy [133] or Bishop [31] for example). They consist of a graph that illustrates the causal relationship between variables. The model is then constructed by assigning a state space to each vertex and constructing a probability measure that respects the causal structure of that graph in a certain sense. Even though there exists a huge amount of literature on causal models (compare for example Lauritzen [114]) we need to extend the standard results in order to make statements about the sensorimotor loop (a definition of our model of the sensorimotor loop follows in Chapter 3).

Our definition generalizes the usual setup in two aspects. First of all we do not assume that the overall probability measure has a density with respect to some product measure (compare Lauritzen [114]) and secondly we allow certain infinite graphs. The latter typically arise whenever ergodic or asymptotic properties of processes are of interest. We are interested in asymptotic probabilistic statements about an agent (usually a robot) interacting with the environment by reacting to sensor inputs via outputs to a motor controller. The class of models that we describe is in our opinion also appropriate for a stochastic model of multiple agents interacting via a specified interface for example in biology or economy. We include non-discrete state spaces and transition kernels whose transition probabilities are not dominated by a single measure. As a simple example for a kernel with the latter property, consider a deterministic transition, i.e. a kernel of the form $K(x, A) := \delta_{f(x)}(A)$ where $x \in \mathbb{R}$, f is a measurable real-valued function and δ_y is the Dirac measure at point y, i.e.

$$\delta_y(A) = \begin{cases} 1 \text{ if } y \in A \\ 0 \text{ else.} \end{cases}$$

The measures $\delta_{f(x)}$ are dominated by a single σ-finite measure μ if and only if the image of f is countable. Let X be a random variable with uniform distribution on $[0, 1]$ and define $Y := f(X)$ and $Z := g(Y)$ for some continuous non-single valued functions $f, g : [0, 1] \to [0, 1]$. The standard theorems on conditional stochastic independence (compare Lauritzen [114], Bishop [31] and Murphy [133]) do not guarantee that Z is independent from X given Y. The reason is that it is impossible to find a product measure on $[0, 1] \times [0, 1] \times [0, 1]$ that dominates the common distribution of (X, Y, Z).

For the purpose of this thesis it is conceptually advantageous to distinguish between the causal model (by this we mean the state space on the graph together with the transition rules) and a specific stochastic dynamic on the graph (or more generally a statistical model of several possible dynamics). This is very common in the theory of Markov chains, where the law of the stochastic process contains two ingredients - the transition kernels (this is what we called transition rules) and the initial distribution. Most questions in the theory of Markov chains are concerned with the dependence of certain stochastic quantities on the starting point, or more generally on the initial measure (like mean passage times of certain points or sets, absorption probabilities for traps, expectation values of first hitting times of certain sets, return probabilities and many more). These statements are statistical statements, since one is interested in the behavior of the model for different measures from a certain class. We will prove the existence of a unique probability law on the entire state space, that is compatible with the initial measure and the transition kernel structure - a result very similar to the theory of Markov chains again.

Afterwards we will focus on the conditional independence properties of causal models. The main result of this chapter is a generalization of the usual conditional independence results known from the standard theory of causal models in machine learning. Our result generalizes the known theorems in two aspects:

- The requirement of the existence of a product measure that dominates the probability law (as assumed in all proofs of the conditional independence results known to the authors) is dispensable.
- The underlying graph is not required to be finite but can rather be chosen from a certain class of infinite graphs (which we will call recursively constructible). This extension is essential for our analysis of the sensorimotor loop later on. Generally speaking this extension is crucial whenever the conditional independence properties are used to describe an infinite process of several variables that influence each other.

Moreover we will prove the strong Markov property of causal models and give some examples for possible applications. To our knowledge this problem has not been addressed so far. The formulation of strong conditional independence statements for general causal models is a little bit technical but we think that once established it provides a very clear view onto the subject. Some new results for Markov chains follow immediately (as strong independence statements for certain pairs of stopping times for example). We will also motivate the definitions and theorems of this section by numerous examples.

In our opinion the theory of causal models on recursively constructible graphs as developed in this chapter is a versatile tool to describe conditional stochastic dependencies between several state variables of a stochastic process that mutually influence each other. In probability theory these dependencies are often described by the underlying process being adapted to an appropriate filtration. This assumption can easily be translated into the language of causal models on recursively constructible graphs. However the full strength of the machinery pops up whenever certain process values do not depend on all the past values but only on a certain collection. In this case a careful graphical representation reveals many more non-trivial (strong) conditional independence results between the state variables.

1.1 Recursively constructible graphs

For this chapter we need some concepts from probability theory, most of which we listed in the Appendix, A.1. First we will summarize some graph-theoretical concepts and define what we will refer to as recursively constructible graph. Afterwards we will justify the definitions and give some examples to illustrate the concept (for a reference on graph theory see for example the first chapter of Lauritzen [114], Diestel [65], Bang-Jensen and Gutin [20] or Bondy and Murty [35]).

Definition 1.1.1 - (**Concepts from graph theory and definition of recursively constructible graphs**)

► **Definition 1.1.1.1:** *A directed graph is a pair (V, E) where V denotes the set of vertices and $E \subseteq V \times V$ denotes the set of edges. This relation can be illustrated graphically by plotting the vertices and linking v and w by an arrow if $(v, w) \in E$.*

► **Definition 1.1.1.2:** *Let (V, E) be a graph. For a given vertex $v \in V$ we define the set of its parents*

$$\mathrm{Par}\,(v) := \{w \in V \,|(w, v) \in E\}$$

and the set of its children

$$\mathrm{Child}\,(v) := \{w \in V \,|(v, w) \in E\}$$

► **Definition 1.1.1.3:** *A subset $A \subseteq V$ is called ancestrally closed, if $v \in A$ implies $\mathrm{Par}(v) \subseteq A$. The intersection of an arbitrary collection of ancestrally closed sets is ancestrally closed again, such that there exists a smallest ancestrally closed set containing a given set $A \subseteq V$. This set is called ancestral closure of A:*

$$\mathrm{An}(A) := \cap_{B \in \{X \in 2^V | X \supseteq A \text{ ; } X \text{ ancestrally closed}\}} B \tag{1.1}$$

► **Definition 1.1.1.4:** *We will write $u \rightsquigarrow v$ if there exists a path from u to v, i.e. there exists a finite sequence $(v_1, v_2, \ldots, v_n) \in V^n$ with $v_1 = u, v_n = v$ and $(v_i, v_{i+1}) \in E$ for $1 \leq i < n$.*

► **Definition 1.1.1.5:** *A directed graph (V, E) is called acyclic if $vu \rightsquigarrow vv$ does not hold true for every $v \in V$.*

► **Definition 1.1.1.6:** *Let (V, E) be a directed acyclic graph. The set of vertices without parents will be called input vertices, i.e.:*

$$V_0 := \{v \in V \,|\mathrm{Par}\, v = \emptyset\}$$

inductively define the set of "vertices with information available from vertices of degree i with $i \leq n$" only, i.e. if $V_0, V_1, \ldots V_n$ have already been defined then:

$$V_{n+1} := \{v \in V \setminus (\cup_{0 \leq i \leq n} V_i) \,|\mathrm{An}\,(\{v\}) \setminus \{v\} \subseteq \cup_{0 \leq i \leq n} V_i\}$$

► **Definition 1.1.1.7:** *A directed acyclic graph (V, E) with vertex set V will be called recursively constructible if $V = \cup_{i \in \mathbb{N}_0} V_i$.*

It is clear that a graph describing a causal structure should be acyclic, since an effect cannot be its own cause. The last condition (the recursive constructability assumption) in Definition 1.1.1 is worth a closer look. The usefulness of this statement will become more obvious later on, when we investigate the probabilistic properties of causal models. Intuitively imagine the graph to illustrate some evaluation scheme and the kernels to express

calculation rules. In a zeroth step assign some value to each input vertex $v \in V_0$. In a first evaluation step calculate the value of each $v \in V_1$ from the values of the vertices V_0 (which is possible since all parents of vertices in V_1 are elements of V_0). All other vertices have non-evaluated parents and therefore cannot be evaluated yet. In a second step the values of vertices in V_2 can be calculated since they have parents in V_1 and V_0 only. Proceeding this way one can evaluate vertices in V_i in the i-Th step. In this picture the recursive constructability condition implies that any given vertex is finally evaluated.
Here is an examples of an infinite graph that is directed but is not recursively constructible: $V = \mathbb{Z}$ and $E := \big\{(i, i+1) \in V^2 \,|\, i \in \mathbb{Z}\big\}$ or graphically:

$$\begin{tikzcd} \cdots \arrow[r] & (i) \arrow[r] & (i+1) \arrow[r] & (i+2) \arrow[r] & (i+3) \arrow[r] & \cdots \end{tikzcd}$$

Caus. mod. 4 - *First example of an acyclic, directed graph that is not recursively constructible*

An example of a directed, non-recursively constructible graph with non-empty set of input vertices is the following one:
$V = \mathbb{Z} \times \{0, 1\}$, $E = \big\{((i, 0), (i+1, 0)) \in V^2 \,|\, i \in \mathbb{Z}\big\} \cup \big\{((i, 1), (i, 0)) \in V_2^2 \,|\, i \in \mathbb{Z}\big\}$ or graphically:

$$\begin{tikzcd} \cdots \arrow[r] & (i,0) \arrow[r] & (i+1,0) \arrow[r] & (i+2,0) \arrow[r] & (i+3,0) \arrow[r] & \cdots \\ \cdots & (i,1) \arrow[u] & (i+1,1) \arrow[u] & (i+2,1) \arrow[u] & (i+3,1) \arrow[u] & \cdots \end{tikzcd}$$

Caus. mod. 5 - *Second example of an acyclic, directed graph that is not recursively constructible*

Recursively constructible graphs can also be characterized by their communication structure. This requires the following concepts about ordering relations (compare Kemeny and Snell [101] for example):

Definition 1.1.2 - (**Ordering relations**)

Let V be some set and let $\leq$ be a binary relation on V.
▶ **Definition 1.1.2.1:** *The binary relation $\leq$ is called* ***weak ordering*** *relation (often also denoted as preorder or quasi order) if*

- *it is transitive, i.e. $x \leq y$ and $y \leq z$ implies $x \leq z$*
- *and reflexive, i.e. $x \leq x$ for all $x \in V$*

▶ **Definition 1.1.2.2:** *Let V be a set, let $a \in V$ and assume $\leq$ to be a weak ordering relation. Then a is a* ***maximal element*** *of A if for all $x \in V$: $a \leq x$ implies $x \leq a$. Moreover a is called minimal element if for all $x \in V$ the identity $x \leq a$ implies $a \leq x$.*
▶ **Definition 1.1.2.3:** *The binary relation $\leq$ is called partial order relation if it is a weak ordering relation and it is antisymmetric, i.e. whenever $x \leq y$ and $y \leq x$ then $x = y$.*
▶ **Definition 1.1.2.4:** *Let $A \subseteq V$ and assume $\leq$ to be a partial ordering relation. Then A is called totally ordered with respect to $\leq$ if any two elements $x, y \in A$ are comparable, i.e. either $x \leq y$ or $y \leq x$.*

For any directed graph (V, E) the set of edges induce a canonical weak ordering relation on the set of vertices indicating whether it is possible to go from one vertex to another, namely

$$a \leq b \text{ if } a = b \text{ or } a \rightsquigarrow b$$

These concepts allow an alternative definition of recursively constructible graphs:

Theorem 1.1.1 - (**Alternative definition of recursively constructible graph**)

Let $G = (V, E)$ be a directed graph and let $\leq$ denote the canonical weak ordering relation induced by E. Then G is recursively constructible if and only if

- *$\leq$ is a partial ordering (alternative statement of G being acyclic) and*
- *For every $a \in V$ there exists some $N \in \mathbb{N}$ such that every totally ordered subset $A \subseteq V$ with maximal element a has cardinality $|A| \leq N$ (equivalent formulation of recursively constructability).*

Proof of Theorem 1.1.1. As usual we write $a < b$ for $a \leq b$ and $a \neq b$.
Let (V, E) be a directed graph. If for $x, y \in V$ both $x \leq y$ and $y \leq x$ are satisfied then either $x = y$ or there exists a path from x to y and from y to x. Therefore $\leq$ is a partial ordering if and only if the graph is acyclic.
Assume (V, E) to be recursively constructible and let $v \in V$. By definition there exists some $n \in \mathbb{N}_0$ such that $v \in V_n$ (compare Definition 1.1.1). Then necessarily $|A| \leq n + 1$ for every totally ordered subset $A \subset V$ with maximal element v, for otherwise there exist elements $w_i \in V$ such that $w_1 < w_2 <, \ldots, < w_k < v$ for some $k > n$ which contradicts $v \in V_n$.
Now assume that for every $v \in V$ there exists some $N \in \mathbb{N}$ such that every totally ordered subset, $A \subseteq V$, with maximal element v has cardinality at most N. Then $v \in \cup_{0 \leq n < N} V_n$. To see this assume that the contrary holds true. Then there exists some element $w_1 \in V$ with $w_1 \in \operatorname{Par}(v)$ and $w_1 \notin \cup_{0 \leq n < N-1} V_n$. Recursively one can construct a sequence $(w_1, w_2, \ldots, w_{N-1})$ such that $w_k \notin \cup_{0 \leq n < N-k} V_n$ and $w_{i+1} \in \operatorname{Par}(w_i)$. Since $w_{N-1} \notin V_0$ there exists $w_N < w_{N-1}$ and

$$\{w_1, w_2, \ldots, w_N\} \cup \{v\}$$

is a totally ordered subset of V with maximal element v and cardinality $N + 1$ contradicting the assumption. ■

This characterization of Definition 1.1.1 can be formulated in a more colloquial language as the statement, that the "ancestral tree" of every vertex v has finite depth. Note that every finite acyclic graph is automatically recursively constructible.

1.2 Causal models and laws on causal models

Beside the graph describing the causal structure, the state spaces and the transition rules have to be specified. The algebraic product (or Cartesian product) of sets $\mathbf{B}_i$ where $i \in I$, is the set of choice functions:

$$\prod_{i \in I} \mathbf{B}_i := \{f : I \to \cup_{i \in I} \mathbf{B}_i \,|\, f(i) \in \mathbf{B}_i\} \,. \tag{1.2}$$

We will frequently need the projections onto the individual factors:

$$\pi_i : \prod_{i \in I} \mathbf{B}_i \to \mathbf{B}_i \; ; f \mapsto f(i) \tag{1.3}$$

For each $i \in I$ let $(\mathbf{B}_i, \mathcal{F}_i)$ be a measurable space. Then the product $\prod_{i \in I} \mathbf{B}_i$ equipped with the product σ-algebra,

$$\otimes_{i \in I} \mathcal{F}_i := \sigma(\{\pi_i\}_{i \in I}), \tag{1.4}$$

is a measurable space. The product σ-algebra is obviously the coarsest σ-algebra that renders all projections π_i measurable. In this thesis we will always equip algebraic products with the product σ-algebra if not stated elsewise.

Now we will define what we mean by a causal model over a given recursively constructible graph. Again we will write down a definition first and then we will give an example to illustrate the concepts.

Definition 1.2.1 - (**Causal model**)

A causal model is a triple $C := ((V, E), \mathfrak{S}, \mathfrak{T})$ *where*

- (V, E) *is a recursively constructible graph encoding the causal structure of the model*
- $\mathfrak{S}$ *is a collection of measurable state spaces indexed by the vertex set of* V*:*

  $$\mathfrak{S} = \prod_{v \in V} \mathbf{S}_v,$$

  where $(\mathbf{S}_v, \mathcal{F}_v)$ *are measurable spaces*
- $\mathfrak{T}$ *is a collection of transition rules, i.e. probability kernels from the parent vertices to their children:*

  $$\mathfrak{T} \in \prod_{v \in V \setminus V_0} \Lambda^{(\mathbf{S}_v, \mathcal{F}_v)}_{(\prod_{w \in \mathrm{Par}(v)} \mathbf{S}_w, \otimes_{w \in \mathrm{Par}(v)} \mathcal{F}_w)}$$

Remark 1.2.1 - (**Remark on Definition 1.2.1**)

Even though the choice of the σ*-algebras* $\mathcal{F}_v$ *is part of the specification of the model we do not explicitly mention this in the definition of* C *to keep notation simple. Very frequently* $\mathbf{S}_v$ *is a finite set (and* $\mathcal{F}_v$ *is the entire power set,* $2^{\mathbf{S}_v}$*) or* $\mathbf{S}_v = \mathbb{R}^n$ *(more generally a topological space) such that* $\mathcal{F}_v$ *is canonically the corresponding Borel* σ*-algebra.*

To keep notation short we will use the following abbreviations:

$$\mathfrak{S}_A := \prod_{v \in A} \mathbf{S}_v \text{ and } \mathcal{F}_A := \sigma(\{\pi_v\}_{v \in A}) \subseteq \otimes_{v \in V} \mathcal{F}_v, \tag{1.5}$$

where $A \subseteq V$. Note that in the definition we assumed $\pi_v : \mathfrak{S} \to \mathbf{S}_v$ such that strictly speaking $\mathcal{F}_A$ is different from $\otimes_{v \in A} \mathcal{F}_v$. The former consists of subsets of $\prod_{v \in V} \mathbf{S}_v$ whereas the latter consists of subsets of $\prod_{v \in A} \mathbf{S}_v$. Of course $\mathcal{F}_A$ and $\otimes_{v \in A} \mathcal{F}_v$ are naturally isomorphic to each other but we prefer $\mathcal{F}_A$ to be a sub σ-algebra of $\mathcal{F}_V$.

We will provide an example to illustrate the concepts now: Imagine a robot that receives input values from it's sensors and can output values to a motor controller (in order to move his arm for example). The causal structure is described by the following recursively constructible graph, denoted by (V, E) in the sequel:

$$\begin{tikzcd}
 & v_{m,1} \arrow[d] & v_{m,2} \arrow[d] & v_{m,3} \arrow[d] & v_{m,4} \arrow[d] & v_{m,5} \arrow[d] & \dots \\
v_{e,0} \arrow[r] & v_{e,1} \arrow[r] \arrow[d] & v_{e,2} \arrow[r] \arrow[d] & v_{e,3} \arrow[r] \arrow[d] & v_{e,4} \arrow[r] \arrow[d] & v_{e,5} \arrow[r] \arrow[d] & \dots \\
 & v_{s,1} & v_{s,2} & v_{s,3} & v_{s,4} & v_{s,5} & \dots
\end{tikzcd}$$

Caus. mod. 6 - *Example: Non-observing robot*

Here vertex $v_{e,i}$ denotes the state of the environment at time $i \in \mathbb{N}_0$, vertex $v_{m,i}$ denotes the state of the motor controller at time $i \in \mathbb{N}$ and vertex $v_{s,i}$ denotes the sensor value at time $i \in \mathbb{N}$. The set of input vertices is

$$V_0 = \{v_{e,0}\} \cup \left(\cup_{i \in \mathbb{N}} \{v_{m,i}\}\right).$$

Furthermore

$$V_1 = \{v_{e,1}\} \ ; V_i = \{v_{e,i}, v_{s,i-1}\} \text{ for } i \geq 2.$$

Let $(\mathbf{S}, \mathcal{F}_S)$ denote the state space for the sensor values and let $(\mathbf{M}, \mathcal{F}_M)$ be the state space for the motor controller. Let the environment contain both the body of the agent (described mathematically by coordinates for all relevant degrees of freedom - this includes coordinates for the center of mass, angles of arm joints etc.) and all degrees of freedom of the environment that are necessary to describe the interaction with the agent's body. Denote the entire space of the environment by $(\mathbf{E}, \mathcal{F}_E)$. Note that we do not impose any strong conditions on the state space of the environment. We only assume it to be a measurable space. If the environment has a finite amount of relevant states only (a situation that naturally arises from approximating the dynamic by coarse-graining), the state space can be chosen to be a finite set. If the system's interaction with the environment can be described by Newtonian mechanics, the state space can be chosen to be a subset of $\mathbb{R}^n$ or more generally it can be chosen to be a differentiable manifold equipped with its Borel σ-algebra. If the robot can be described as a test-particle interacting with a fluid (if the robot steers an aircraft for example) the environment can also be chosen to be an appropriate infinite dimensional topological vector space equipped with it's Borel σ-algebra.
To sum up:

$$\mathfrak{S}_v = \begin{cases} \mathbf{E} & \text{if } v = v_{e,i} \text{ for some } i \in \mathbb{N}_0 \\ \mathbf{S} & \text{if } v = v_{s,i} \text{ for some } i \in \mathbb{N} \\ \mathbf{M} & \text{if } v = v_{m,i} \text{ for some } i \in \mathbb{N} \end{cases}$$

Assuming a time homogeneous, deterministic, Markovian dynamic, the new environment state depends on the current motor value and the previous state of the environment only. Let $T : \mathbf{E} \times \mathbf{M} \to \mathbf{E}$ be a measurable map encoding the transition from the old environment state and a given state of the motor controller to a new state of the environment. Then

$$\mathfrak{T}_{v_{e,i}}((e,m), A) = \delta_{T(e,m)}(A) \text{ where } i \in \mathbb{N}; e \in \mathbf{E}; m \in \mathbf{M} \text{ and } A \in \mathcal{F}_{\mathbf{E}}.$$

The robot can get information about the environment via its sensor values only. For simplicity we assume again that the sensor value depends deterministically on the current environmental state. If we denote this (measurable) map by $T' : \mathbf{E} \to \mathbf{S}$ then

$$\mathfrak{T}_{v_{s,i}}(e, A) = \delta_{T'(e)}(A) \text{ for } i \in \mathbb{N}; e \in \mathbf{E} \text{ and } A \in \mathcal{F}_{\mathbf{S}}.$$

Then the triple $((V, E), \mathfrak{S}, \mathfrak{T})$ is a causal model according to Definition 1.2.1.

The notion of causal model captures all dynamical restrictions on the system that hold for all possible dynamics. For the specification of a specific dynamic, the values of all initial vertices must be specified (more generally we allow initial probability measures

$p \in M_1\left(\otimes_{v\in V_0}\mathcal{F}_v\right)$. To ensure compatibility with the Markovian structure of the graph, we require these probability measures to be product measures, i.e. we require the family $\left(\pi_v : \mathfrak{S}_{V_0} \to \mathfrak{S}_v\right)_{v\in V_0}$ to be independent under p. A product measure $p \in M_1\left(\otimes_{v\in V_0}\mathcal{F}_v\right)$ is completely determined by the distribution of the projections π_v, denoted by

$$\pi_{v*}p \tag{1.6}$$

(also referred to as marginal distributions of p). On the other hand every collection of measures $p_v \in M_1\left(\mathcal{F}_v\right)$ where $v \in V_0$ determines a unique product measure $p \in M_1\left(\otimes_{v\in V_0}\mathcal{F}_v\right)$ with marginals p_v. This existence theorem does not require any further regularity assumptions on the state spaces. For countable index sets this follows directly from the Ionescu-Tulcea extension theorem (compare Lemma 1.2.1) and the extension to uncountable index sets is straightforward (compare Kallenberg [99], Corollary 6.18 on p. 117 for example).

Definition 1.2.2 - (**Admissible initial measures and causal statistical models over a causal model**)

Let $C := ((V,E),\mathfrak{S},\mathfrak{T})$ *be a causal model.*

▶ **Definition 1.2.2.1:** *An admissible initial measure on* C *is a probability measure,* $p \in M_1\left(\otimes_{v\in V_0}\mathcal{F}_v\right)$*, such that the family* $\left(\pi_v : \mathfrak{S}_{V_0} \to \mathfrak{S}_v\right)_{v\in V_0}$ *is independent under the measure* p*.*
▶ **Definition 1.2.2.2:** *A causal statistical model over* C *is a set of admissible initial measures on* C*.*
▶ **Definition 1.2.2.3:** *A parametric causal statistical model over* C *is a pair* $(\mathfrak{Z},\hat{p})$ *where*

- $\mathfrak{Z}$ *is the parameter set.*
- $\hat{p}$ *is an injective map from* $\mathfrak{Z}$ *to the admissible initial measures on* C*.*

In the robot example above there are several canonical ways to define a statistical model over the causal model describing the environment-agent system. One canonical choice is

$$Q := \left\{p \in M_1\left(\otimes_{v\in V_0}\mathcal{F}_v\right) \middle| p \text{ admissible initial measure }; \pi_{v*}p = \delta_{s_v}; v \in V_0; s \in \mathfrak{S}_{V_0}\right\},$$

where δ_x denotes the Dirac measure centered at x. This choice corresponds to arbitrary initial states and a fixed, deterministic sequence of motor controller values. A second canonical choice corresponds to the time-homogeneous situation. By this we mean that the value of the motor controller is the same for all instances of time, i.e.

$$\begin{aligned} Q' \;:=\; & \left\{p \in M_1\left(\otimes_{v\in V_0}\mathcal{F}_v\right) \middle| p \text{ admissible initial measure };\right. \\ & \left.\pi_{v_{e,0}*}p = \delta_z; \pi_{v_{m,i}*}p = \delta_q; z \in \mathbf{E}; q \in \mathbf{M}\right\} \end{aligned}$$

Our restrictions on the graph imply that any initial measure on $\otimes_{v\in V_0}\mathcal{F}_v$ extends to a unique probability law on $\mathcal{F} := \mathcal{F}_V$ that is compatible with the initial measure, the transition kernels and satisfies a further independence assumption to be stated later on. The proof is based on the following lemma, a proof of which can be found in Kallenberg [99], p.116:

Lemma 1.2.1 - (**Extension theorem by Ionescu-Tulcea**)

Let $(\mathbf{S}_n,\mathcal{F}_n)$ *be measurable spaces, let* $\mu_1 \in M_1\left(\mathcal{F}_1\right)$ *and let*

$$\mu_n \in \Lambda^{(\mathbf{S}_n,\mathcal{F}_n)}_{\left(\prod_{1\le i<n}\mathbf{S}_i,\otimes_{1\le i<n}\mathcal{F}_i\right)} \text{ for every } n > 1.$$

For every cylinder set $A \in \otimes_{n\in\mathbb{N}}\mathcal{F}_n$ *(by (finite) cylinder set we mean that* A *can be written as* $A = A' \times \prod_{k>n} \mathbf{S}_k$ *where* $A' \in \otimes_{1\leq k\leq n}\mathcal{F}_k$*) define*

$$P(A) = \int_{A'} \mu_1(d\omega_1)\mu_2(\omega_1, d\omega_2)\ldots\mu_n((\omega_1,\ldots,\omega_{n-1}), d\omega_n) \tag{1.7}$$

Then P *is a well defined map from the (finite) cylinder sets to the positive real numbers and possesses a unique extension to a probability measure on* $\otimes_{n\in\mathbb{N}}\mathcal{F}_n$*.*

Definition 1.2.3 - (**Some notation**)

Let $G := (V, E)$ *be a recursively constructible graph. Let* V_n *be the sets from Definition 1.1.1. Set:*

$$V_{<n} := \cup_{0\leq k<n} V_k \text{ and } V_{\leq n} := \cup_{0\leq k\leq n} V_k \tag{1.8}$$

We need the following lemma, that follows immediately from existence and uniqueness of the product measure:

Lemma 1.2.2 - (**Independent combination of kernels**)

Let $C := ((V, E), \mathfrak{S}, \mathfrak{T})$ *be a causal model and let* $A \subseteq V_n$ *for some* $n \geq 1$ *(see Definition 1.2.3). Then there exists a unique kernel*

$$\mathfrak{T}_A \in \Lambda^{(\mathfrak{S}_A, \otimes_{v\in A}\mathcal{F}_v)}_{(\mathfrak{S}_{V_{<n}}, \otimes_{v\in V_{<n}}\mathcal{F}_v)}$$

such that for every set of the form

$$B := \left(\prod_{v\in J} B_v\right) \times \mathfrak{S}_{A\setminus J} \text{ where } J \subseteq A; |J| < \infty; B_v \in \mathcal{F}_v \tag{1.9}$$

we have

$$\mathfrak{T}_A(s, B) := \prod_{v\in V} \mathfrak{T}_v\left(\left(s_w\right)_{w\in \mathrm{Par}(v)}, B_v\right) \tag{1.10}$$

For the proof we need the monotone class argument, based on the following definition:

Definition 1.2.4 - (**π-systems, λ-systems**)

▶ **Definition 1.2.4.1:** *Let* Ω *be a set and let* $\mathcal{C} \subseteq 2^{\Omega}$*. Then* $\mathcal{C}$ *is called* π*-system if* $A, B \in \mathcal{C}$ *implies* $A \cap B \in \mathcal{C}$*.*
▶ **Definition 1.2.4.2:** *Let* Ω *be a set and let* $\mathcal{C} \subseteq 2^{\Omega}$*. Then* $\mathcal{C}$ *is called* λ*-system if*

- $\Omega \in \mathcal{C}$
- $\mathcal{C}$ *is closed under proper differences, i.e. whenever* $A, B \in \mathcal{C}$ *and* $B \subseteq A$ *then* $A \setminus B \in \mathcal{C}$
- $\mathcal{C}$ *is closed under increasing limits, i.e. whenever* $A_n \in \mathcal{C}$ *and* $A_n \subseteq A_{n+1}$ *then* $\cup_{n\in\mathbb{N}} A_n \in \mathcal{C}$

The monotone class theorem is a useful tool to extend certain statements from a π-system to the σ-algebra generated by this π-system. It is used in the standard proof of Fubini's theorem for example. The following version originates from Kallenberg [99], p.2:

Lemma 1.2.3 - (**Monotone class argument**)

Let $\mathcal{C}$ be a π-system, let $\mathcal{D}$ be a λ-system such that $\mathcal{C} \subseteq \mathcal{D}$. Then $\sigma(\mathcal{C}) \subseteq \mathcal{D}$.

Proof of Lemma 1.2.2. Fix $s \in \mathfrak{S}_{V_{<n}}$. By the existence and uniqueness of product measures over arbitrary index sets (compare the comment prior to Definition 1.2.2) there exists a unique measure $P_s \in M_1(\otimes_{v\in A}\mathcal{F}_v)$ such that the family $\pi_v : \mathfrak{S}_A \to \mathfrak{S}_v$ is independent under P_s and

$$\pi_{v*}P_s := \mathfrak{T}_v\left((s_j)_{j\in\mathrm{Par}(v)}, \cdot\right)$$

Let $\mathcal{C}$ denote the class of subsets of the form Eq. 1.9. Then $\mathcal{C}$ is a π-system generating the σ-algebra $\otimes_{v\in A}\mathcal{F}_v$.

It remains to show that $P_s(B)$ is a measurable function of s for every $B \in \otimes_{v\in A}\mathcal{F}_v$. This is clear for every $B \in \mathcal{C}$. Moreover the collection of all sets $B \in \otimes_{v\in A}\mathcal{F}_v$ for which $s \mapsto P_s(B)$ is measurable is a λ-system (it is closed with respect to proper differences by measurability of the addition in $\mathbb{R}$ and is closed with respect to increasing limits by σ-additivity of the measure P_s and the measurability of the pointwise limit of a sequence of real-valued, measurable functions). By the monotone class argument $s \mapsto P_s(B)$ is measurable for all $B \in \sigma(\mathcal{C})$. This is the statement. ■

From here on we will always write $\mathfrak{T}_A$ (where $A \subseteq V_n$) for the kernel from Lemma 1.2.2.

Theorem 1.2.1 - (**Law on a causal model**)

Let $C := ((V,E), \mathfrak{S}, \mathfrak{T})$ be a causal model. Then:

▶ **Theorem 1.2.1.1:** *For every measure $p \in M_1(\otimes_{v\in V_0}\mathcal{F}_v)$ there exists exactly one probability law $\hat{P}$ on $\mathcal{F}_V$ that satisfies the following two conditions:*

- *It is compatible with the initial measure, p, i.e.*

  $$\left((\pi_v)_{v\in V_0}\right)_* \hat{P} = p \tag{1.11}$$

  where

  $$\pi_v : \mathfrak{S} \to \mathfrak{S}_v; f \to f(v)$$

  denotes the canonical projection from $\mathfrak{S}$ to the individual factors.

- *The kernels, $\mathfrak{T}_v$, describe the transition probabilities and the process is memoryless in the sense that every sensor value, is independent from its ancestors given all its parent's values. Moreover we require that for every $n \in \mathbb{N}$ the family $(\pi_v)_{v\in V_n}$ is "maximally uncorrelated", in the sense that this family is independent given all the "past data", $\mathcal{F}_{V_{<n}}$. To sum up - using the notation from Lemma 1.2.2 - we assume that for every $A \subseteq V_n$ with $n > 0$ there exists a null set $N \in \mathcal{F}_V$ such that*

  $$\hat{P}\left[(\pi_v)_{v\in A} \in B \,|\mathcal{F}_{V_{<n}}\right](\omega) = \mathfrak{T}_A\left((\pi_w(\omega))_{w\in V_{<n}}, B\right) \tag{1.12}$$

  For every $B \in \otimes_{v\in A}\mathcal{F}_v$ whenever $\omega \in \mathcal{F}_V \setminus N$.

▶ **Theorem 1.2.1.2:** *Let $\mathbf{s} \in \mathfrak{S}_{V_0}$ and let $\hat{P}_{\mathbf{s}}$ be the unique law from Theorem 1.2.1 associated to the initial measure $\delta_{\mathbf{s}}$. Then the map*

$$\hat{K} : \mathfrak{S}_{V_0} \times \otimes_{v\in V}\mathcal{F}_v \to [0,1] \ ; (\mathbf{s}, B) \mapsto \hat{P}_{\mathbf{s}}[B]$$

is a probability kernel from $(\mathfrak{S}_{V_0}, \otimes_{v\in V_0}\mathcal{F}_v)$ to $(\mathfrak{S}, \mathcal{F}_V)$ and for an arbitrary initial measure $p \in M_1(\otimes_{v\in V_0}\mathcal{F}_v)$ the associated law, $\hat{P}$, satisfies:

$$\hat{P}[B] = \int K(s,B)p(ds) \text{ and } \hat{P}[B\,|\mathcal{F}_{V_0}] = \hat{K}\left[(\pi_v)_{v\in V_0}, B\right] \text{ a.s.} \tag{1.13}$$

for every $B \in \mathcal{F}_V$.

Proof of Theorem 1.2.1. Set $\tilde{S}_n := (\pi_v)_{v\in V_n}$. The two conditions of the first part of the theorem are satisfied if and only if

$$\hat{P}\left[\left\{\tilde{S}_0 \in A\right\}\right] = p\left[A\right] \tag{1.14}$$

for every $A \in \otimes_{v\in V_0}\mathcal{F}_v$ and

$$\hat{P}\left[\left\{\tilde{S}_n \in A\right\}\Big|\tilde{S}_0, \tilde{S}_1, \dots, \tilde{S}_{n-1}\right] = \mathfrak{T}_{V_n}\left(\left(\tilde{S}_{k,v}\right)_{k<n;v\in V_k}, A\right) \text{ a.s.} \tag{1.15}$$

for every $A \in \otimes_{v\in V_n}\mathcal{F}_v$. Thus existence and uniqueness in Theorem 1.2.1 follows immediately from the Ionescu-Tulcea extension theorem and Lemma 1.2.2.

For Theorem 1.2.1 let $s \in \mathfrak{S}_{V_0}$ and let $\hat{K}(s, B)$ be the unique law induced by the admissible initial measure δ_s. Let $C \in \mathcal{F}_V$ be a set of the form

$$C = \left(\prod_{0\le k\le n} B_k\right) \times \mathfrak{S}_{V\setminus V_{\le n}}; B_k \in \otimes_{v\in V_k}\mathcal{F}_k. \tag{1.16}$$

Then $\hat{K}(s,\cdot)$ satisfies

$$\begin{aligned} &\hat{K}\left(s, C\right) \qquad (1.17)\\ = \;&\delta_{(s_v)_{v\in V_0}}\left[B_0\right]\int_{\prod_{i=1}^n B_i} \mathfrak{T}_{V_1}\left[s_0, ds_1\right]\mathfrak{T}_{V_2}\left[\left(s_0, s_1\right), ds_2\right]\dots\mathfrak{T}_{V_n}\left[\left(s_0,\dots,s_{n-1}\right), ds_n\right]. \end{aligned}$$

It remains to show that $s \mapsto \hat{K}\left(s, B\right)$ is measurable for every $B \in \mathcal{F}_V$. Note that for any $K \in \Sigma^{(\mathbf{Y},\mathcal{F}_Y)}_{(\mathbf{X},\mathcal{F}_X)}$ and any bounded $\mathcal{F}_Y/\mathcal{B}_{\mathbb{R}}$-measurable function $f : \mathbf{Y} \to \mathbb{R}$, the function:

$$s \mapsto \int f(s')K(s, ds') \tag{1.18}$$

is measurable. This is a standard result and follows from the following argument: Whenever $f = \mathbb{1}_A$ with $A \in \mathcal{F}_Y$ then the result is true since $K(s, A)$ is measurable in s. For general bounded, measurable f the result follows from approximating f by indicator functions and applying Lebesgue's dominated convergence theorem (or alternatively the monotone convergence theorem and linearity).
We will show that $s \mapsto \hat{K}\left(s, C\right)$ is measurable for every $C \in \mathcal{F}_V$ of the form Eq. 1.16. The proof uses induction over $n \in \mathbb{N}_0$. For $n = 0$ let $J \subseteq V_0$ such that $|J| < \infty$. Then for every C' of the form

$$C' = \left(\prod_{v\in J} B_v\right) \times \mathfrak{S}_{V\setminus J}; B_v \in \mathcal{F}_v \tag{1.19}$$

the map

$$s \mapsto K\left(s, C'\right) = \prod_{v\in J} \mathbb{1}_{B_v}\left(s_v\right)$$

is clearly measurable. By a monotone class argument this extends to all $C' \in \mathcal{F}_{V_0}$, such that the claim is true for $n = 0$.
Now assume that the claim is true for some index $n \in \mathbb{N}_0$. Then for every C' of the form

$$C' = B_n \times B_{n+1} \times \mathfrak{S}_{V\setminus\left(V_{\le n+1}\right)}; B_n \in \otimes_{v\in V_{\le n}}\mathcal{F}_v, B_{n+1} \in \otimes_{v\in V_{n+1}}\mathcal{F}_v \tag{1.20}$$

we have:

$$\begin{aligned} \hat{P}_s\left[C'\right] &= E_s\left[\mathbb{1}_{\left\{(\pi_v)_{v\in V_{\le n}}\in B_n\right\}}\hat{P}_s\left[\left\{(\pi_v)_{v\in V_n} \in B_{n+1}\right\}\middle|\mathcal{F}_{V_{\le n}}\right]\right] \\ &= \int_{B_n} \mathfrak{T}_{V_{n+1}}\left[s', B_{n+1}\right] d\left(\pi_v\right)_{v\in V_{\le n}*}\hat{P}_s\left(ds'\right) \qquad (1.21) \end{aligned}$$

By the inductive assumption $K(s,\cdot) := \left((\pi_v)_{v\in V_{\leq m}}\right)_* \hat{P}_s$ is a probability kernel, such that, by the general remark, Eq.1.18, the inductive assumption implies the measurability of $s \mapsto \hat{K}\left[s, C'\right]$ for every $C' \in \mathcal{F}_{\leq n+1}$. Hence the statement is true for all $n \in \mathbb{N}_0$.
The measurability of $\hat{K}(\cdot, B)$ for general $B \in \mathcal{F}_V$ follows by a monotone class argument again. This proves that $\hat{K}$ is indeed a probability kernel.

To show the last statement (Eq. 1.13) consider some event $C \in \mathcal{F}_V$ of the form Eq. 1.16 and some $p \in M_1\left(\otimes_{v\in V_0}\mathcal{F}_v\right)$ inducing the law $\hat{P}$. Iterating the conditioning in Eq. 1.21, using the identities Eq. 1.17 and Eq. 1.15 gives

$$\hat{P}\left[C\right] \;=\; \int_C p(ds_0)\hat{K}\left[s, ds'\right]$$

By a monotone class argument this relation extends to all $C \in \mathcal{F}_V$. Eq. 1.22 clearly implies:

$$\hat{P}\left[C\left|\mathcal{F}_{V_0}\right.\right] = \hat{K}\left((\pi_v)_{v\in V_0}, C\right) \tag{1.22}$$

■

Sometimes it is useful to write the transition kernels as a deterministic transition of former state variables and a randomization variable. This will be advantageous for the analysis of stochastic gradient algorithms for example. The relation between the kernel approach and a description by a randomized transition is given by the following theorem a proof of which can be found in Kallenberg [99], p.112. We also use the definition of Borel spaces used in this book. They are defined to be measure spaces that are measure isomorphic (i.e. there exists a measurable bijection between them with measurable inverse) to a measurable subset of the unit interval. As a classical result in advanced measure theory every Borel subset of a Polish space is a Borel space (compare Breiman [39], Kechris [100] and Rogers and Williams [154] for example).

Theorem 1.2.2 - (**Randomization of transition kernels**)

Let $(\mathbf{S}_1, \mathcal{B}_1)$ and $(\mathbf{S}_2, \mathcal{B}_2)$ be Borel spaces and let $K \in \Lambda^{(\mathbf{S}_2,\mathcal{B}_2)}_{(\mathbf{S}_1,\mathcal{B}_1)}$. Let X be a random variable on some probability space $(\Omega, \mathcal{F}, P)$ with values in $\mathbf{S}_1$ Then there exists a measurable transition function

$$T : \mathbf{S}_1 \times [0,1] \to \mathbf{S}_2$$

such that

$$\hat{P}\left[\left\{T(X(\pi_1), \pi_2) \in B\right\} \left| X(\pi_1)\right.\right] = K(X(\pi_1), B) \text{ a.s.}$$

where

- $\hat{P} \in M_1\left(\mathcal{F} \otimes \mathcal{B}_{[0,1]}\right)$ *is the product measure of P and the Lebesgue measure,* $\nu_{\text{Leb.}}$.
-
  $$\pi_1 : \Omega \times [0,1] \to \Omega \,;\, \pi_1 : \Omega \times [0,1] \to [0,1]$$

  denote the projections onto the first and second factor of the Cartesian product.

One of our main interest in causal models over recursively constructible graphs lies in a mathematical rigorous definition of a model of learning algorithms over a Markov decision processes. This type of question is very general: Given a causal model with a certain inbuilt dynamic (specified by the transition kernels) - what is a good model for a controlled dynamic over this causal model? Abstractly a controlled dynamic can be seen as an extension of the causal model that preserves the old transition rules. For reasons that will be explained later on we require the new kernels of the extended model to be deterministic (compare Lemma 1.2.4 and Remark 1.2.3).

Definition 1.2.5 - (**Controller graphs and controlled model over a given causal model**)

▶ **Definition 1.2.5.1:** *Let* (V, E) *be a recursively constructible graph. A controller graph over* (V, E) *is another recursively constructible graph ,* (V', E')*, with the following properties:*

- (V, E) *is a subgraph of* (V', E')
- *For every* $v \in V \setminus V_0$ *the parents of* v *in* V *are equal to the parents of* v *in* V'*:*

$$\mathrm{Par}'(v) = \mathrm{Par}(v) \tag{1.23}$$

▶ **Definition 1.2.5.2:** *Let* $C := ((V, E), \mathfrak{S}, \mathfrak{T})$ *be a causal model. A controller extension over* C *is a causal model* $C' := ((V', E'), \mathfrak{S}', \mathfrak{T}')$ *satisfying*

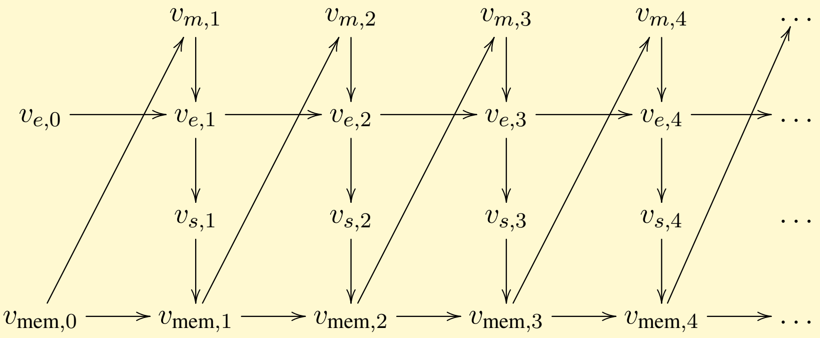
- $\mathfrak{S}'_v = \mathfrak{S}_v$ *whenever* $v \in V$.
- $\mathfrak{T}'_v = \mathfrak{T}_v$ *for every* $v \in V \setminus V_0$
- *For* $v \in V' \setminus (V'_0 \cup (V \setminus V_0))$ *the transitions are deterministic, i.e. there exist measurable maps* $T : \mathfrak{S}'_{\mathrm{Par}(v)} \to \mathfrak{S}'_v$ *such that*

$$\mathfrak{T}'[s, B] = \delta_{T(s)}[B] \text{ for every } s \in \mathfrak{S}'_{\mathrm{Par}(v)}; B \in \mathcal{F}_v$$

Before we comment on the definition, we given an illustration for the robot example described in Caus. mod. 6. We assume that the original dynamic is controlled in the following way:

- A sensor value is read and written to the memory (therefore we will add some vertices $v_{\mathrm{mem},i}$ where $i \in \mathbb{N}_0$ for the memory variables)
- From the current state of the memory an output value is calculated and sent to the motor controller.

This yields the following recursively constructible graph (which is indeed a controller graph over Caus. mod. 6):



Caus. mod. 7 - *Example: New causal structure of robot-world system*

The specification of a controller extension requires a specification of the memory state spaces, a specification of the transition functions describing the memory update and transition functions describing the motor update from the memory.

Remark 1.2.2 - (**Comment on Definition 1.2.5**)

In our definition we allow variables in $V' \setminus V$ to have children in V_0 such that feedback controls are included. Moreover causality is automatically satisfied (in the sense that data from future vertices cannot be used to manipulate a vertex in the past) through the requirement that (V', E') is a recursively constructible graph and therefore acyclic. A controller extension specifies the transition rules for the overall model. The restriction to deterministic transitions for the new vertices and V_0 will be justified in Lemma 1.2.4 and Remark 1.2.3.

We would like to model the controlled dynamics on the same probability space as the original dynamic, this is possible as the following lemma shows:

Lemma 1.2.4 - (**Controlled dynamic as a random variable over the original causal model**)

Let $C := ((V, E), \mathfrak{S}, \mathfrak{T})$ be a causal model and let $C' := ((V', E'), \mathfrak{S}', \mathfrak{T}')$ be a controller extension of C. Let $W_1 := V_0' \cap V$ denote the set of non-controlled input vertices of V and let $W_2 := V_0' \setminus V$ be the set of new input vertices. Let $K' \in \Lambda^{(\mathfrak{S}', \mathcal{F}_{V'})}_{\left(\mathfrak{S}'_{V_0'}, \otimes_{v \in V_0'} \mathcal{F}_v\right)}$ be the kernel from Theorem 1.2.1 for the model C'.

Then there exist a kernel $K'' \in \Lambda^{(\mathfrak{S}, \otimes_{v \in V} \mathcal{F}_v)}_{\left(\mathfrak{S}_{V_0'}, \otimes_{v \in V_0'} \mathcal{F}_v\right)}$ and a $((\otimes_{v \in W_2} \mathcal{F}_v) \otimes (\otimes_{v \in V} \mathcal{F}_v))/\mathcal{F}_{V'}$-measurable map

$$R : \mathfrak{S}'_{W_2} \times \mathfrak{S} \to \mathfrak{S}'$$

such that for every $s_1 \in \mathfrak{S}'_{W_1}$ and $s_2 \in \mathfrak{S}'_{W_2}$:

$$R(s_2, \cdot)_* K''((s_1, s_2), \cdot) = K'((s_1, s_2), \cdot) \tag{1.24}$$

In other words for every initial value $(s_1, s_2) \in \mathfrak{S}_{W_1} \times \mathfrak{S}_{W_2}$ the measure $K'[(s_1, s_2), \cdot]$ can be considered as the distribution of some $\mathfrak{S}'$-valued random variable $R(s_2, \cdot)$ on the probability space $(\mathfrak{S}, \otimes_{v \in V} \mathcal{F}_v, K''[(s_1, s_2), \cdot])$.

Proof of Lemma 1.2.4. Let $K' \in \Lambda^{(\mathfrak{S}', \mathcal{F}_{V'})}_{\left(\mathfrak{S}'_{V_0'}, \otimes_{v \in V_0'} \mathcal{F}_v\right)}$ be the kernel from Theorem 1.2.1 for the model C' and define for $B \in \otimes_{v \in V} \mathcal{F}_v$:

$$K''[(s_1, s_2), B] := K'\left[(s_1, s_2), \left\{(\pi_v)_{v \in V} \in B\right\}\right] \tag{1.25}$$

For $v \in W_2 \cup V$, $s_2 \in \mathfrak{S}'_{W_2}$ and $s \in \mathfrak{S}$ set

$$R_v(s_2, s) := \begin{cases} s_{2,v} & \text{for } v \in W_2 \\ s_v & \text{if } v \in V \end{cases}. \tag{1.26}$$

Then R_v is $(\otimes_{w \in W_2} \mathcal{F}_w) \otimes (\otimes_{w \in V} \mathcal{F}_w)/\mathcal{F}_v$-measurable.
We will construct R_v for $v \notin V \cup W_2$ recursively. Assume that, R_v, is a well-defined $(\otimes_{w \in W_2} \mathcal{F}_w) \otimes (\otimes_{w \in V} \mathcal{F}_w)/\mathcal{F}_v$ measurable map for every $v \in V \cup V'_{\leq n}$. Then define for $v \in V'_{n+1} \setminus V$:

$$R_v(s_2, s) := T_v\left[(R_w(s_2, s))_{w \in \mathrm{Par}(v)}\right] \tag{1.27}$$

where T_v is the transition function associated to vertex v.
By assumption R_w is $(\otimes_{u \in W_2} \mathcal{F}_u) \otimes (\otimes_{u \in V} \mathcal{F}_u)/\mathcal{F}_w$-measurable for every $w \in V'_{\leq n}$. This implies that the map $\tilde{R} : (s_2, s) \mapsto (R_w(s_2, s))_{w \in \mathrm{Par}'(v)}$ is $(\otimes_{w \in W_2} \mathcal{F}_w) \otimes (\otimes_{w \in V} \mathcal{F}_w)/\mathcal{F}_v$-measurable (compare for example Lemma 1.8 in Kallenberg [99]). Therefore R_v is measurable as a composition of measurable maps. Similarly the map $R : (s_2, s) \mapsto (R_v(s_2, s))_{v \in V'}$ is $(\otimes_{w \in W_2} \mathcal{F}_w) \otimes (\otimes_{w \in V} \mathcal{F}_w)/\otimes_{w \in V'} \mathcal{F}_w$-measurable.

Fix $s_1 \in \mathfrak{S}'_{W_1}$ and $s_2 \in \mathfrak{S}'_{W_2}$. We will show that

$$R(s_2,\cdot)_* K''((s_1,s_2),B) = K'\left((s_1,s_2),B\right) \tag{1.28}$$

for every $B \in \mathcal{F}_{V'}$. By definition of K'' and R this statement is true whenever $B \in \mathcal{F}_V$. We will show the validity 1.28 for all $B \in \mathcal{F}_{V \cup V'_{\leq n}}$ by induction over n. Then by a monotone class argument the validity of Eq. 1.28 extends to all $B \in \mathcal{F}_{V'}$.

For $n = 0$ consider a finite subset $J \subseteq V'_0 \setminus V$, subsets $B_v \in \mathcal{F}_v$ for every $v \in J$ and some $C_0 \in \otimes_{v \in V} \mathcal{F}_v$. Define $B' \in \mathcal{F}_{V'}$ via

$$B' := \left\{(\pi_v)_{v \in V} \in C_0\right\} \cap \cap_{v \in J} \left\{\pi_v \in B_v\right\} \tag{1.29}$$

Since for every $v \in J$ and $s \in \mathbf{S}$

$$R_v\left(s_2, s\right) \in B_v \text{ iff } s_{2,v} \in B_v$$

we have by definition of K'':

$$\begin{aligned} & R(s_2,\cdot)_* K''((s_1,s_2),B') \\ = & \prod_{v \in J} \delta_{B_v}\left(s_{2,v}\right) \cdot K'\left[\left(s_1,s_2\right), \left\{(\pi_w)_{w \in V} \in C_0\right\}\right] = K'\left[\left(s_1,s_2\right),B'\right] \end{aligned}$$

This result extends by a monotone class argument to all sets $B' \in \mathcal{F}_{v \cup V'_0}$. Now assume that Eq. 1.28 holds for all $B \in \mathcal{F}_{v \cup V'_{\leq n}}$ and let $J \subseteq V'_{n+1} \setminus V$ be a finite set, let $B_v \in \mathcal{F}_v$ for every $v \in J$, let $C_n \in \otimes_{v \in V'_{\leq n} \cup V} \mathcal{F}_v$ and consider the set

$$B' := \left\{(\pi_v)_{v \in V'_{\leq n} \cup V} \in C_n\right\} \cap \cap_{v \in J} \left\{\pi_v \in B_v\right\} \tag{1.30}$$

By definition of R for any $s \in \mathfrak{S}_V$

$$\begin{aligned} & R(s_2,s) \in B' \qquad (1.31) \\ \text{iff} \quad & \left(R_v(s_2,s)\right)_{v \in V'_{\leq n} \cup V} \in C_n \text{ and } T_v\left(\left(R_w\left(s_2,s\right)\right)_{w \in \mathrm{Par}(v)}\right) \in B_v \text{ for all } v \in J \end{aligned}$$

Thus by the inductive assumption:

$$\begin{aligned} & R\left(s_2,\cdot\right)_* K''\left[\left(s_1,s_2\right),B'\right] \\ = \; & K'\left[\left(s_1,s_2\right), \left\{(\pi_v)_{v \in V'_{\leq n} \cup V} \in C_n\right\} \cap \cap_{v \in J} \left\{(\pi_w)_{w \in \mathrm{Par}(v)} \in T_v{}^{-1}(B_v)\right\}\right] \\ = \; & K'\left[\left(s_1,s_2\right),B'\right] \qquad (1.32) \end{aligned}$$

where the last step followed from the definition of K. Thus by the monotone class argument the claim is also true for $n+1$ and Eq. 1.28 is indeed satisfied for any $B \in \mathcal{F}_{V'}$

■

Remark 1.2.3 - (Comment on the advantages of deterministic transition rules in Definition 1.2.5)

The proof of Lemma 1.2.4 works because the new transition rules are deterministic. Constructing the controlled dynamic over the same probability space as the original one has many technical advantages. For controlled dynamics it is usually assumed that any reasonable estimator depends on past observations only. This is typically incorporated by assuming that this estimator is adapted to an appropriate filtration. Formulations become significantly more complicated if this filtration lives on a probability space that is not known and has to be constructed together with the extension of the model.

If "new randomness" for the controlled dynamic is unavoidable (if the determination of the controls from former process values involves Monte-Carlo methods or noisy measurements for example) it is often possible to add some isolated vertices to the original causal model, C, and to use these isolated vertices as input for some vertices in $V' \setminus V$ (compare Theorem 1.2.2).

1.3 Conditional independence in causal models

Before we investigate the conditional independence in causal models we will revise some concepts from graph theory and probability theory. The first definition summarizes some necessary concept related to graph separation (see for example: Lauritzen [114]):

Definition 1.3.1 - (**Moral graph and d-separation**)

▶ **Definition 1.3.1.1:** *Let $G := (V, E)$ be a directed graph. Then the moral graph of G, denoted by G^m, is the undirected graph (V, E') that originates from deleting directions and marrying parents:*
$\{a, b\} \in E'$ if and only if one of the following statements holds true:

- $(a, b) \in E$
- $(b, a) \in E$
- *There exists $c \in V$ such that $(a, c) \in E$ and $(b, c) \in E$.*

▶ **Definition 1.3.1.2:** *Let $G := (V, E')$ be an undirected graph and let $A, B, S \subseteq G$ be disjoint. Then S separates A and B, if and only if every path from $v \in A$ to $w \in B$ contains an element $s \in S$.*
▶ **Definition 1.3.1.3:** *Let $G := (V, E)$ be a directed graph and let $A, B, S \subseteq G$ be disjoint. Then S d-separates A and B, if S separates A and B in $\left(G|_{\mathrm{An}(A \cup B \cup S)}\right)^m$. Here $G\,|_W$ denotes the restriction of the graph $G = (V, E)$ to the subset $W \subseteq V$, i.e. $G\,|_W = (W, E_W)$ with $(a, b) \in E_W$ if and only if $(a, b) \in E$.*

The next definition introduces the well-known concept of independence of σ-algebras.

Definition 1.3.2 - (**Conditional independence**)

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{G}_1, \mathcal{G}_2, \mathcal{G}_3 \subseteq \mathcal{F}$ be sub σ-algebras. Then $\mathcal{G}_1$ is independent of $\mathcal{G}_2$ given $\mathcal{G}_3$, written as

$$\mathcal{G}_1 \perp\!\!\!\perp \mathcal{G}_2\,|\mathcal{G}_3 \tag{1.33}$$

if for every $A \in \mathcal{G}_1$ and $B \in \mathcal{G}_2$:

$$P\left[A, B\,|\mathcal{G}_3\right] = P\left[A\,|\mathcal{G}_3\right] P\left[B\,|\mathcal{G}_3\right] \text{ a.s.} \tag{1.34}$$

this is equivalent to

$$E\left[XY\,|\mathcal{G}_3\right] = E\left[X\,|\mathcal{G}_3\right] E\left[Y\,|\mathcal{G}_3\right] \text{ a.s.} \tag{1.35}$$

for any pair of bounded random variables $X, Y : \Omega \to \mathbb{R}$ where X is $\mathcal{G}_1/\mathcal{B}_{\mathbb{R}}$ measurable and Y is $\mathcal{G}_2/\mathcal{B}_{\mathbb{R}}$ measurable.
The main theorem of this section is the following one:

Theorem 1.3.1 - (**Conditional stochastic independence in Causal models**)

Let $C := ((V, E), \mathfrak{S}, \mathfrak{T})$ be a causal model, let $p \in M_1(\otimes_{v \in V_0} \mathcal{F}_v)$ be an admissible initial measure and let $P \in M_1(\mathcal{F}_V)$ be the associated law (compare Theorem 1.2.1). Let $S, A, B \subseteq V$ be disjoint such that S d-separates A and B. Then

$$\mathcal{F}_A \perp\!\!\!\perp \mathcal{F}_B\,|\mathcal{F}_S \tag{1.36}$$

under P.

For the proof we will need more concepts and fundamental theorems from probability theory (as can be found in most textbooks on measure theoretic probability theory and measure theory like Kallenberg [99], Bauer [22], Jacod and Protter [92], Rudin [161], Breiman [39]). The following theorem can be found in Kallenberg [99], p. 111 for example:

Lemma 1.3.1 - (**Equivalent characterization of conditional independence**)

Let $\mathcal{F}$ be a σ-algebra and let $\mathcal{F}_n, \mathcal{H}, \mathcal{G} \subseteq \mathcal{F}$ be sub σ- algebras (where $n \in \mathbb{N}_0$). Then the following two conditions are equivalent:

1. $\mathcal{H} \perp\!\!\!\perp (\mathcal{F}_n)_{n\in\mathbb{N}} \,|\mathcal{G}$
2. $\mathcal{H} \perp\!\!\!\perp \mathcal{F}_{n+1} \,|\mathcal{G}, \mathcal{F}_1, \ldots, \mathcal{F}_n$ *for all* $n \geq 0$

We also need a simple version of the martingale convergence theorem:

Lemma 1.3.2 - (**Martingale convergence theorem**)

Let $(\Omega, \mathcal{F}, P, \mathbb{F})$ be a filtered probability space and let X be a bounded $\mathbb{F}$-martingale, then X_n converges almost surely (and in L_p for every $p \geq 1$) to some

$$\mathcal{F}_\infty := \cap_{n\in\mathbb{N}} \sigma\left(\mathbb{F}_k; k \geq n\right) \text{-measurable}$$

random variable X_∞. Moreover: $X_n = E\left[X_\infty \,|\mathbb{F}_n\right]$ almost surely.

Before proving the main theorem, we would like to transform the statement into an equivalent formulation that can be handled easier. We need the following sets:

Definition 1.3.3 - (**More concepts related to graph separation**)

Let $G := (V, E)$ be a directed graph and let $S, A, B \subseteq V$ be disjoint vertex sets.

► **Definition 1.3.3.1:** *Define the causal connected component of A by:*

$$\mathrm{CON}_{B\cup A\cup S,S}(A) \quad := \quad A \cup \{v \in \mathrm{An}(A \cup B \cup S)\,|v \leftrightsquigarrow_{m,S} a \text{ for some } a \in A\}$$

Here $v \leftrightsquigarrow_{m,S} a$ means that there exists a path between a and v in $\left(G|_{\mathrm{An}(A\cup B\cup S)}\right)^m$ that does not hit S.

► **Definition 1.3.3.2:** *The set of residual ancestors is:*

$$\mathrm{ResAnc}(A\,|S|\,B) \quad := \quad \mathrm{An}(A \cup B \cup S) \setminus (\mathrm{CON}_{B\cup A\cup S,S}(A) \cup \mathrm{CON}_{B\cup A\cup S,S}(B))$$

We will use the following equivalence:

Lemma 1.3.3 - (**Reformulation of causal separation**)

Let $G = (V, E)$ be a recursively constructible graph and let $S, A, B \subseteq V$ be disjoint. Then the following conditions are equivalent:

1. *S d-separates A and B*
2. $\mathrm{ResAnc}(A\,|S|\,B)$ *d-separates* $\mathrm{CON}_{B\cup A\cup S,S}(A)$ *and* $\mathrm{CON}_{B\cup A\cup S,S}(B)$

Proof of Lemma 1.3.3. All the following statements about the existence of certain paths are with respect to the graph $\left(G|_{\mathrm{An}(A\cup B\cup S)}\right)^m$.
Assume that S separates A and B and that there exists a path from some $a \in \mathrm{CON}_{B\cup A\cup S,S}(A)$ to some $b \in \mathrm{CON}_{B\cup A\cup S,S}(B)$ that does not pass through $\mathrm{ResAnc}(A\,|S|\,B)$. By definition of $\mathrm{CON}_{B\cup A\cup S,S}(A)$ there exists a path from a to some element $a' \in A$ that does not pass through S. Equivalently there exists a path from b to some element $b' \in B$ that does not pass through S. Hence there exists a path from a' to b' that does not pass through $S \subseteq \mathrm{ResAnc}(A\,|S|\,B)$. This is a contradiction to S d-separating A and B. Therefore the first statement implies the second one.
On the other hand assume that the second statement holds true and take an arbitrary path from $a \in A$ to $b \in B$. By validity of the second statement this path hat to pass some element $s \in \mathrm{ResAnc}(A\,|S|\,B)$. By definition of $\mathrm{ResAnc}(A\,|S|\,B)$ either the path from a to s has to pass S or the path from s to b has to pass S. Therefore the second statement implies the first one. ■

Therefore:

Lemma 1.3.4 - (**Towards an equivalent formulation of Theorem 1.3.1**)

Let $G = (V, E)$ be a directed graph, let $(\Omega, \mathcal{F}, P)$ be a probability space and let $(\mathcal{F}_A)_{A\subseteq V}$ be a monotonous family of sub σ-algebras of $\mathcal{F}$, i.e. $A \subseteq B$ implies that $\mathcal{F}_A$ is a sub-σ-algebra of $\mathcal{F}_B$. Then the following two statements are equivalent:

1. *For every sets $S, A, B \subseteq V$ such that S d-separates A and B we have*
$$\mathcal{F}_A \perp\!\!\!\perp \mathcal{F}_B\,|\mathcal{F}_S$$
2. *For every sets $S, A, B \subseteq V$ such that S d-separates A and B we have*
$$\mathcal{F}_{\mathrm{CON}_{B\cup A\cup S,S}(A)} \perp\!\!\!\perp \mathcal{F}_{\mathrm{CON}_{B\cup A\cup S,S}(B)}\,\big|\mathcal{F}_{\mathrm{ResAnc}(A|S|B)}$$

Proof. The first condition implies the second one by Lemma 1.3.3. Assume that the second condition holds and let $A, B, S \subseteq V$ be such that S d-separates A and B.
Define $B' := \mathrm{CON}_{B\cup A\cup S,S}(B) \cup \mathrm{ResAnc}(A\,|S|\,B) \setminus S$. Observe that $\mathrm{An}(A \cup B \cup S) = \mathrm{An}(A \cup B' \cup S)$. Moreover $\mathrm{ResAnc}(A\,|S|\,B') = S$. Therefore by assumption:

$$\mathcal{F}_{\mathrm{CON}_{B\cup A\cup S,S}(A)} \perp\!\!\!\perp \mathcal{F}_{\mathrm{CON}_{B\cup A\cup S,S}(B')}\,|\mathcal{F}_S \tag{1.37}$$

Since $A \subseteq \mathrm{CON}_{B\cup A\cup S,S}(A)$ and $B \subseteq \mathrm{CON}_{B\cup A\cup S,S}(B')$ the monotonicity of the family of σ-algebras gives the desired result. ■

After this preparation, we proceed with the proof of Theorem 1.3.1.

Proof of Theorem 1.3.1. By Lemma 1.3.4 and Lemma 1.3.3 we can assume that $A = \mathrm{CON}_{A\cup B\cup S,S}(A)$, $B = \mathrm{CON}_{A\cup B\cup S,S}(B)$ and $S = \mathrm{ResAnc}(A\,|S|\,B)$. In other words we can assume that any ancestor of some element $v \in A \cup B \cup S$ is contained in one of the three disjoint sets A, B and S.
Define the filtrations

$$\mathbb{F}_{A,n} := \mathcal{F}_{W_{A,n}}\,,\text{ where } W_{A,n} := A \cap \left(\cup_{0\leq k\leq n} V_k\right), \tag{1.38}$$

$$\mathbb{F}_{B,n} := \mathcal{F}_{W_{B,n}}\,,\text{ where } W_{B,n} := B \cap \left(\cup_{0\leq k\leq n} V_k\right), \tag{1.39}$$

and

$$\mathbb{F}_{S,n} := \mathcal{F}_{W_{S,n}}\,,\text{ where } W_{S,n} = S \cap \left(\cup_{0\leq k\leq n} V_k\right) \tag{1.40}$$

As a first step we will show inductively that for every $n \geq 0$:

$$\mathbb{F}_{A,n} \perp\!\!\!\perp \mathbb{F}_{B,n} \,|\mathbb{F}_{S,n} \tag{1.41}$$

If $n = 0$ this statement is trivially true, since we assumed p to be a product measure and therefore $\mathbb{F}_{A,0}$, $\mathbb{F}_{B,0}$ and $\mathbb{F}_{S,0}$ are independent (we set $\mathbb{F}_{S,n} = \{\emptyset, \Omega\}$ whenever $W_{S,n} = \emptyset$, such that this statement remains true in this case).
For the step from n to $n+1$ we define the sets

$$A_{n+1} := W_{A,n+1} \setminus W_{A,n} \text{ and } B_{n+1} := W_{B,n+1} \setminus W_{B,n}$$

Furthermore we divide the set $W_{S,n+1} \setminus W_{S,n}$ into the following two subsets:

$$S_{n+1,A} := \{v \in W_{S,n+1} \setminus W_{S,n} \,|\mathrm{Par}\,(v) \cap A \neq \emptyset\}$$

and

$$S_{n+1,B} := (W_{S,n+1} \setminus W_{S,n}) \setminus S_{n+1,A}$$

The fact that S d-separates A and B implies that

- elements $v \in A_{n+1} \cup S_{n+1,A}$ have parents in $W_{A,n} \cup W_{S,n}$ only
- elements $v \in S_{n+1,B} \cup B_{n+1}$ have parents in $W_{B,n} \cup W_{S,n}$ only

This implies that for every $C \in \otimes_{v \in S_{n+1,A} \cup A_{n+1}} \mathcal{F}_v$ the conditional expectation is

$$P\left[\left\{\pi_{S_{n+1,A} \cup A} \in C\right\} |\mathbb{F}_{A,n}, \mathbb{F}_{B,n}, \mathbb{F}_{S,n}\right] \;=\; \mathfrak{T}_{S_{n+1,A} \cup A_{n+1}}\left(\pi_{V_{\leq n}}, C\right) \text{ a.s.}$$

where we wrote π_W as a shorthand notation for $(\pi_v)_{v \in W}$ and used the notation from Lemma 1.2.2. This expression does not depend on vertex values of $v \in W_{B,n}$ and therefore possesses a $\sigma(\mathbb{F}_{A,n}, \mathbb{F}_{S,n})$-measurable version. Analogously for every $C \in \otimes_{v \in S_{n+1,B} \cup B_{n+1}} \mathcal{F}_v$ the conditional expectation

$$P\left[\left\{\pi_{S_{n+1,B} \cup B} \in C\right\} |\mathbb{F}_{A,n}, \mathbb{F}_{B,n}, \mathbb{F}_{S,n}\right] \;=\; \mathfrak{T}_{S_{n+1,B} \cup B_{n+1}}\left(\pi_{V_{\leq n}}, C\right) \text{ a.s.}$$

possesses a $\sigma(\mathbb{F}_{B,n}, \mathbb{F}_{S,n})$-measurable version. By the very construction of the process law (compare Theorem 1.2.1) we have for every $C_A \in \otimes_{v \in S_{n+1,A} \cup A_{n+1}} \mathcal{F}_v$ and $C_B \in \otimes_{v \in S_{n+1,B} \cup B_{n+1}} \mathcal{F}_v$:

$$\begin{aligned} & P\left[\left\{\pi_{S_{n+1,A} \cup A} \in C_A\right\} \cap \left\{\pi_{S_{n+1,B} \cup B} \in C_B\right\} |\mathbb{F}_{A,n}, \mathbb{F}_{B,n}, \mathbb{F}_{S,n}\right] \\ = \;& \mathfrak{T}_{S_{n+1,A} \cup A_{n+1}}\left(\pi_{V_{\leq n}}, C_A\right) \cdot \mathfrak{T}_{S_{n+1,B} \cup B_{n+1}}\left(\pi_{V_{\leq n}}, C_B\right) \end{aligned}$$

More generally fix any $C_{A,1} \in \otimes_{v \in W_{A,n}} \mathcal{F}_v$, $C_{A,2} \in \otimes_{v \in A_{n+1} \cup S_{n+1,A}} \mathcal{F}_v$, $C_{B,1} \in \otimes_{v \in W_{B,n}} \mathcal{F}_v$ and $C_{B,2} \in \otimes_{v \in B_{n+1} \cup S_{n+1,B}} \mathcal{F}_v$ and set

$$C_A := C_{A,1} \times C_{A,2} \text{ and } C_B := C_{B,1} \times C_{B,2}.$$

Then almost surely:

$$\begin{aligned} & P\left[\left\{\pi_{W_{A,n+1} \cup S_{n+1,A}} \in C_A\right\}, \left\{\pi_{W_{B,n+1} \cup S_{n+1,B}} \in C_B\right\} |\mathbb{F}_{S,n}, \mathbb{F}_{A,n}, \mathbb{F}_{B,n}\right] \\ = \;& P\left[\left\{\pi_{W_{A,n+1} \cup S_{n+1,A}} \in C_A\right\} |\mathbb{F}_{S,n}, \mathbb{F}_{A,n}\right] \cdot \\ & \cdot P\left[\left\{\pi_{W_{B,n+1} \cup S_{n+1,B}} \in C_B\right\} |\mathbb{F}_{S,n}, \mathbb{F}_{B,n}\right] \end{aligned} \tag{1.42}$$

where

$$\begin{aligned} & P\left[\left\{\pi_{W_{A,n+1} \cup S_{n+1,A}} \in C_A\right\} |\mathbb{F}_{S,n}, \mathbb{F}_{A,n}\right] \\ = \;& \mathbb{1}_{\{\pi_{A_n} \in C_{A,1}\}} \mathfrak{T}_{S_{n+1,A} \cup A_{n+1}}\left((\pi_v)_{v \in V_{\leq n}}, C_{A,2}\right) \end{aligned} \tag{1.43}$$

and

$$\begin{aligned} & P\left[\left\{\pi_{W_{B,n+1}\cup S_{n+1,B}} \in C_B\right\} \middle| \mathbb{F}_{S,n}, \mathbb{F}_{B,n}\right] \\ = \quad & \mathbb{1}_{\left\{\pi_{B_n} \in C_{B,1}\right\}} \mathfrak{T}_{S_{n+1,B}\cup B_{n+1}}\left(\left(\pi_v\right)_{v\in V_{\leq n}}, C_{B,2}\right) \end{aligned} \tag{1.44}$$

By the induction hypothesis

$$\left(\mathbb{F}_{A,n}, \mathbb{F}_{S,n}\right) \perp\!\!\!\perp \left(\mathbb{F}_{B,n}, \mathbb{F}_{S,n}\right) | \mathbb{F}_{S,n} \ . \tag{1.45}$$

This, Eq. 1.42 and the tower property of the conditional expectation imply:

$$\begin{aligned} & P\left[\left\{\pi_{W_{A,n+1}\cup S_{n+1,A}} \in C_A, \pi_{W_{B,n+1}\cup S_{n+1,B}} \in C_B\right\} \middle| \mathbb{F}_{S,n}\right] \\ = \quad & P\left[\left\{\pi_{W_{A,n+1}\cup S_{n+1,A}} \in C_A\right\} \middle| \mathbb{F}_{S,n}\right] \cdot \\ & \cdot P\left[\left\{\pi_{W_{B,n+1}\cup S_{n+1,B}} \in C_B\right\} \middle| \mathbb{F}_{S,n}\right] \end{aligned} \tag{1.46}$$

Sets of the form

$$C_A := \left\{\pi_{W_{A,n}} \in C_{A,1}, \pi_{A_{n+1}\cup S_{n+1,A}} \in C_{A,2}\right\}$$

form π-system that generates the σ-algebra $\mathcal{F}_{W_{A,n+1}\cup S_{n+1,A}}$. Moreover the collection of measurable sets C_A that satisfy Eq. 1.46 for fixed C_B forms a λ-system. So by the monotone class theorem Eq. 1.46 holds true for all $C_A \in \mathcal{F}_{W_{A,n+1}\cup S_{n+1,A}}$. Fixing an arbitrary C_A and using the same argument again, yields the validity of Eq. 1.46 for all $C_B \in \mathcal{F}_{W_{B,n+1}\cup S_{n+1,B}}$. This proves:

$$\left(\mathcal{F}_{W_{A,n+1}}, \mathcal{F}_{S_{n+1,A}}\right) \perp\!\!\!\perp \left(\mathcal{F}_{W_{B,n+1}}, \mathcal{F}_{S_{n+1,B}}\right) | \mathbb{F}_{S,n} \tag{1.47}$$

By Lemma 1.3.1 this implies

$$\mathbb{F}_{A,n+1} \perp\!\!\!\perp \mathbb{F}_{B,n+1} | \mathbb{F}_{S,n+1} \tag{1.48}$$

Hence Eq. 1.41 holds for every $n \in \mathbb{N}_0$.

Now fix some arbitrary large $n, m \in \mathbb{N}$, $C_A \in \mathbb{F}_{A,n}$ and $C_B \in \mathbb{F}_{B,m}$. Since indicator functions of sets are bounded, Lemma 1.3.2 implies that the martingales

$$X_k := E\left[\mathbb{1}_{C_A} | \mathbb{F}_{S,k}\right], \tag{1.49}$$

$$Y_k := E\left[\mathbb{1}_{C_B} | \mathbb{F}_{S,k}\right] \tag{1.50}$$

and

$$Z_k := E\left[\mathbb{1}_{C_B\cap C_A} | \mathbb{F}_{S,k}\right] \tag{1.51}$$

converge almost surely to some limit X (Y and Z respectively). Note that X is $\mathcal{F}_S$ measurable as a limit of $\mathcal{F}_S$ measurable functions. Moreover for any $C \in \mathbb{F}_{S,t}$ with $t > 0$ by conditional dominated convergence:

$$E\left[\mathbb{1}_C X\right] = \lim_{k\to\infty} E\left[\mathbb{1}_C E\left[X | \mathbb{F}_{S,k}\right]\right] = E\left[\mathbb{1}_C \mathbb{1}_{C_A}\right] \tag{1.52}$$

This is true for any t. Since $\left(\mathbb{F}_{S,t}\right)_{t\in\mathbb{N}}$ generates the σ-algebra $\mathcal{F}_S$ this implies:

$$X = P\left[C_A | \mathcal{F}_S\right]. \tag{1.53}$$

Analogously

$$Y = P\left[C_B | \mathcal{F}_S\right] \tag{1.54}$$

and

$$Z = P\left[C_B \cap C_A | \mathcal{F}_S\right] \tag{1.55}$$

By Eq. 1.41 and the product rule for limits of real-valued sequences:

$$P\left[C_A \cap C_B | \mathcal{F}_S\right] = P\left[C_A | \mathcal{F}_S\right] \cdot P\left[C_B | \mathcal{F}_S\right] \tag{1.56}$$

Finally by the monotone class argument Eq. 1.56 is true for any $C_A \in \mathcal{F}_A$ and $C_B \in \mathcal{F}_B$ showing that

$$\mathcal{F}_A \perp\!\!\!\perp \mathcal{F}_B \,|\mathcal{F}_S$$

■

Remark 1.3.1 - (**Comment on the proof of Theorem 1.3.1**)

The proof solely depends on the recursive construction of the model and therefore it strongly relies on the underlying graph being directed. If the underlying graph is directed but not recursively constructible, then the proof fails. However the proof technique above might still be useful to extend known conditional independence properties to larger sets. Imagine for example that we still have the causal ordering expressed by the sets V_n *but assume that* $n \in \mathbb{Z}$. *Assume that* S *d-separates* A *and* B. *As in the proof above one can construct the filtrations* $\mathbb{F}_{A,n}$, $\mathbb{F}_{B,n}$ *and* $\mathbb{F}_{S,n}$ *(this time with* $n \in \mathbb{Z}$*). Then the independence of* $\mathcal{F}_A$ *and* $\mathcal{F}_B$ *given* $\mathcal{F}_S$ *holds true if*

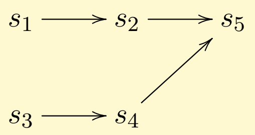
$$\mathbb{F}_{A,n} \perp\!\!\!\perp \mathbb{F}_{B,n} \,|\mathbb{F}_{S,n} \tag{1.57}$$

is satisfied for some $n \in \mathbb{Z}$.

Now we will give some examples of conditional independence for some specific causal models:

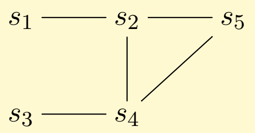
Example 1.3.1 - (**Graph separation and conditional independence in Causal models**)

Let $C = (G, \mathfrak{S}, \mathfrak{T})$ *be a causal model. We use the following convention for all upcoming examples: vertices of the graph will be expressed by lowercase letters and the corresponding vertex random variables,* π_v, *will be denoted by upper case letters. As an example consider the following graph:*



Caus. mod. 8 - *Example: example of a recursively constructible graph*

By our convention we will write S_1 *for the random variable* π_{s_1} *for example. The corresponding moral graph is*



Caus. mod. 9 - *Example: Moral graph of Caus. mod. 8*

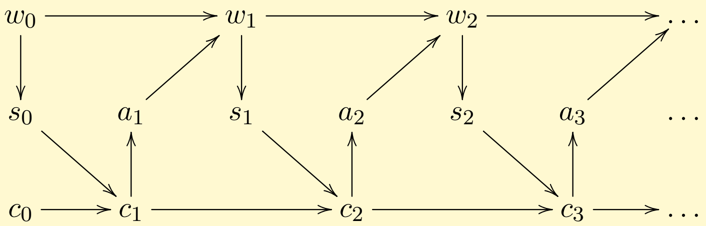
and the conditional independence theorem implies:

$$S_1 \perp\!\!\!\perp S_3 \,|S_2, S_4, S_5$$

and

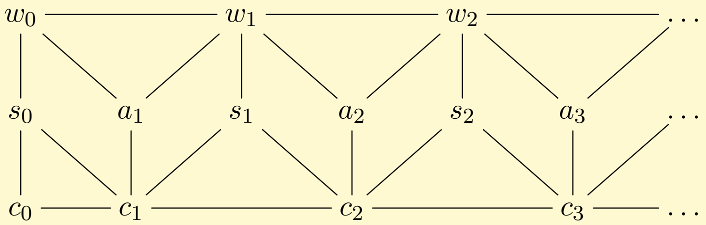
$$(S_1, S_2) \perp\!\!\!\perp S_3 \,|S_4, S_5$$

Consider an agent interacting with the environment by controlling some motor value. This can be modelled by the following causal model (where w_i *denotes the* i*-th state of the world,* s_i *denotes the* i*-th sensor value,* a_i *denotes the* i*-th action and* c_i *denotes the* i*-th state of the agent's memory):*



Caus. mod. 10 - *Example: Agent interacting with the world*

The corresponding moral graph is:



Caus. mod. 11 - *Example: Moral graph of Caus. mod. 10*

The independence theorem shows for example that the process of future actions is independent of the past actions given the current world state and the current memory value:

$$(A_i)_{i>n} \perp\!\!\!\perp (A_i)_{i\leq n} \,|W_n, C_{n+1} \text{ for any } n \in \mathbb{N} \tag{1.58}$$

Another consequence is that the world process is independent from the memory process given all actions and all sensor values:

$$(W_i)_{i\in\mathbb{N}_0} \perp\!\!\!\perp (C_i)_{i\in\mathbb{N}_0} \,\big|(S_i)_{i\in\mathbb{N}_0}, (A_i)_{i\in\mathbb{N}} \tag{1.59}$$

We will prove a partial converse of Theorem 1.3.1, namely

Theorem 1.3.2 - (**Causal model from appropriate conditional independence properties**)

Let $G = (V, E)$ be a recursively constructible graph and let $(\mathbf{S}_v, \mathcal{F}_v)$ (where $v \in V$) be a family of Borel spaces. Furthermore let P be a probability measure on $\otimes_{v\in V}\mathcal{F}_v$ satisfying the following conditional independence property:
For every disjoint sets $A, B, S \subseteq V$ such that S d-separates A and B we have

$$\mathcal{F}_A \perp\!\!\!\perp \mathcal{F}_B \,|\mathcal{F}_S \tag{1.60}$$

Then there exists a measure $p \in M_1(\otimes_{v\in V_0}\mathcal{F}_v)$ with independent marginals and a collection of kernels $\mathfrak{T}_v \in \Lambda^{(\mathbf{S}_v,\mathcal{F}_v)}_{(\prod_{w\in\mathrm{Par}(v)}\mathbf{S}_w,\otimes_{w\in\mathrm{Par}(v)}\mathcal{F}_w)}$ for every $v \in V \setminus V_0$ such that $C := (G, (\mathbf{S}_v)_{v\in V}, \mathfrak{T})$ is a causal model and the measure p induces the law P on C.

The proof relies on a fundamental theorem on the existence of regular versions of conditional distributions for Borel spaces (a proof can be found in Kallenberg [99], Bauer [22], König [105] for example):

Lemma 1.3.5 - (**Borel spaces and regular versions of conditional distributions**)

Let X and Y be random variables over some probability space $(\Omega, \mathcal{F}, P)$. Assume that X has values in the Borel space $(\mathbf{X}, \mathcal{F}_X)$ and that Y has values in some measurable space

$(\mathbf{Y}, \mathcal{F}_Y)$. Then there exists a kernel $K \in \Lambda^{(X,\mathcal{F}_X)}_{(Y,\mathcal{F}_Y)}$ such that for every $A \in \mathcal{F}_X$

$$P\left[X \in A \,|Y\right] = K(Y, A) \text{ a.s.} \tag{1.61}$$

the kernel K will be called regular version of the conditional probability distribution of X given Y.

Proof of Theorem 1.3.2. For every $v \in V \backslash V_0$ let $\mathfrak{T}_v$ be a regular version of the conditional distribution of π_v given $(\pi_w)_{w \in \mathrm{Par}(v)}$. Define $p := \left((\pi_v)_{v \in V_0}\right)_* P$ to be the distribution of $(\pi_w)_{w \in V_0}$. Note that arbitrary finite, disjoint sets $A, B \in V_0$ are d-separated by the empty set. Therefore $(\pi_v)_{v \in A}$ and $(\pi_v)_{v \in B}$ are independent for every pair of disjoint sets $A, B \subseteq V_0$, implying that p has independent marginals.

Let $v \in V_n$. Since $\mathrm{Par}\,(v)$ d-separates $\{v\}$ and $V_{<n} \setminus \mathrm{Par}(v)$:

$$P\left[\pi_v \in C\,|\mathcal{F}_{V_{<n}}\right] = \mathfrak{T}_v\left((\pi_w)_{w \in \mathrm{Par}(v)}, C\right) \text{ a.s. , for all } C \in \mathcal{F}_v \tag{1.62}$$

Every pair of disjoint sets $A, B \subseteq V_n$ is d-separated by $V_{<n}$, such that

$$(\pi_v)_{v \in A} \perp\!\!\!\perp (\pi_v)_{v \in B}\,|\mathcal{F}_{<n} \tag{1.63}$$

for every pair of disjoint sets $A, B \subseteq V_n$. Hence Eq. 1.62 implies:

$$P\left[(\pi_v)_{v \in V_n} \in C\,|\mathcal{F}_{V_{<n}}\right] = \mathfrak{T}_{V_n}\left((\pi_w)_{w \in V_{<n}}, C\right) \text{ a.s. , for all } C \in \otimes_{v \in V_n}\mathcal{F}_v \tag{1.64}$$

By the uniqueness statement about the law, $\hat{P}$ compatible with the measure $p \in M_1\left(\otimes_{v \in V_0}\mathcal{F}_v\right)$ and the causal structure (see Theorem 1.3.1), we have:

$$P = \hat{P} \tag{1.65}$$

■

Note that the restriction to Borel spaces is sufficient but not necessary. The proof of Theorem 1.3.2 shows that a sufficient (and obviously also necessary) requirement is that π_v possesses a regular conditional distribution given $(\pi_w)_{w \in \mathrm{Par}(v)}$ for every $v \in V \setminus V_0$.

Remark 1.3.2 - (**Comment on Theorem 1.3.2**)

There is a subtle point connected to Theorem 1.3.2. In a certain sense a causal model contains far more informative then a single probability distribution that is compatible with the causal structure of the underlying graph. Consider for example two Borel spaces $(\mathbf{X}, \mathcal{F}_X)$ and $(\mathbf{Y}, \mathcal{F}_Y)$ and a probability measure $P \in M_1\left(\mathcal{F}_X, \mathcal{F}_Y\right)$. By Theorem 1.3.2 P is compatible with the graph

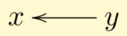
$$x \longrightarrow y$$

Caus. mod. 12 *- Example: Causal model from probability distribution*

i.e. there exists an initial distribution $p_x \in M_1\left(\mathcal{F}_X\right)$ and a kernel $K_1 \in \Lambda^{\mathbf{Y}}_{\mathbf{X}}$ such that for every $A \in \mathcal{F}_X \otimes \mathcal{F}_Y$:

$$P(A) = \int \mathbb{1}_A\left[(x, y)\right] p_x(dx) K_1\left[x, dy\right]$$

However P is also compatible with the graph



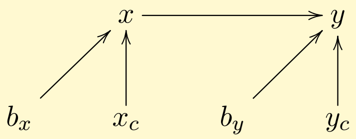
Caus. mod. 13 - Example: Causal model from probability distribution

i.e. there exists an initial distribution $p_y \in M_1(\mathcal{F}_Y)$ and a kernel $K_2 \in \Lambda_{\mathbf{Y}}^{\mathbf{X}}$ such that for every $A \in \mathcal{F}_X \otimes \mathcal{F}_Y$:

$$P(A) = \int \mathbb{1}_A[(x,y)]\, p_y(dy) K_2[y, dx]$$

A causal model on the other hand does not describe a single probability distribution but rather describes the stochastic dynamics for any initial measure (which can be considered as process independent exterior input).

In order to allow for manipulation of some (or all) variables for Caus. mod. 12, it is reasonable to add four vertices b_x, b_y, x_c and y_c and change the underlying graph to the following one:



Caus. mod. 14 - Example: Causal model with controllable x-and y-variable

Here the b-variables indicate whether the variable is controlled or not, and x_c, y_c denote the input in case of manual control. Having this idea in mind attach the state space $(\mathbf{X}, \mathcal{F}_X)$ to x_c, the state space $(\mathbf{Y}, \mathcal{F}_Y)$ to y_c and the state space $\left(\{0,1\}, 2^{\{0,1\}}\right)$ to the vertices b_x and b_y. The intuitive meaning of b_x, b_y, x_c and y_c can be formalized as follows:

$$\mathfrak{T}_x\left[(b, x'), A\right] := \begin{cases} p_x[A] & \text{if } b = 0 \\ \mathbb{1}_A(x') & \text{else} \end{cases} \tag{1.66}$$

where $x' \in \mathbf{X}$, $b \in \{0,1\}$ and $A \in \mathcal{F}_X$ and

$$\mathfrak{T}_y\left[(x', b, y'), A\right] := \begin{cases} K_1[x', A] & \text{if } b = 0 \\ \mathbb{1}_A(y') & \text{else} \end{cases} \tag{1.67}$$

where $y' \in \mathbf{Y}$, $x' \in \mathbf{X}$, $b \in \{0,1\}$ and $A \in \mathcal{F}_Y$. The construction of manipulable variables, shown in the two vertex-example above provides a general strategy to formalize Pearl's do-calculus (compare Pearl [140] and Huang and Valtorta [90]) as a special causal statistical model over a causal model ("do(x')" means calculation with respect to the law generated by $b_1 = 1$ and $\pi_{x_c} = x'$ a.s., or in the statistical model language $p_{b_x} = \delta_{\{1\}}$ and $p_{x_c} = \delta_{\{x'\}}$).

The introduction of manipulable variables in the sense above allows an investigation of causal effects using statistical methods. Whereas Caus. mod. 12 and Caus. mod. 13 are stochastically indistinguishable as long as only one probability distribution is involved, the model Caus. mod. 14 clearly breaks this symmetry: a fixed value $x' \in \mathbf{X}$ usually changes the distribution of π_y whereas a fixed value of $y' \in \mathbf{Y}$ has no influence on the distribution of π_x. This is exactly what is commonly understood as a causal influence of x on y.

The discussion shows that a proper causality measures cannot be defined for a single probability distribution on the fundamental variables (by this we mean the variables whose causal relation is supposed to be investigated, in the example above these variables are given by x and y) only. Any proper causality measure must include interactions with the system. Note however that the dynamic of the original model with interaction is a model itself, and therefore should be made explicit. Even though the construction above is very natural if not to say obvious, it relies on the underlying assumptions as any model does. The knowledge of

b_x being the "switch" for x for example is essential for the conclusion to be valid. In other words: the definition of a causality measure requires a causal model for the consequences of input manipulations. In any way a measure defined for a single probability distribution of the fundamental variables, like the transfer entropy (Schreiber [166]), an extension of Granger causality (compare Granger [76]), is no proper causality measures in this understanding. The transfer entropy measures the average improvement of predictability of some vertex random variable(s) by inspecting another vertex random variable (or a collection of vertex random variables) but fails to quantify proper causal influence.

A more elaborate discussion of proper causality measures necessarily involves the consideration of more complicated underlying graphs. Within the causal model framework correlations between two vertex random variables at x and y might originate from either causal influence (i.e. $x \in \mathrm{An}(y)$ and every path from an element $v \in \mathrm{An}(x)$ to y has to pass vertex x or the same holds true with vertex x and y interchanged), from a common cause (i.e. non of the two vertices lies in the ancestral set generated by the other) or by a mixtures of these two possibilities (i.e. $x \in \mathrm{An}(y)$ and there exists a path from an ancestor of x to y that does not pass through x or the same holds true with x and y interchanged). For a more detailed discussion and a suggestion of a proper causality measure see for example Ay [9], Ay and Zahedi [14] and Ay and Polani [13].

1.4 Strong conditional independence in causal models

In this section we will prove a generalization of the strong Markov property of Markov processes (compare Kallenberg [99], Bauer [22], König [105]) to causal models. This generalization allows an extension of the conditional independence results from states on deterministically chosen sets $A, B, S \subseteq V$ (compare Theorem 1.3.1) to certain randomly chosen sets. We will use this results frequently for our analysis of the sensorimotor loop in Chapter 4. We will provide plenty of examples to motivate and illustrate the concepts.

Loosely speaking the strong Markov properties for Markov processes can be understood as the independence of events in the future of a given stopping time, τ, and events in the past of τ given the value of τ and the process value at τ, X_τ. For general causal models it is straight forward to replace "past" and "future" by arbitrary random sets, A and B whose position depends on another "random set", τ, (generalizing the stopping time) and the vertex values at τ, $(\pi_v)_{v\in\tau}$ (more generally we will allow τ to be an N-tuples of random sets such that the positions of A and B can be specified with respect to multiple random sets). To define random sets as set-valued random variables a σ-algebra on the power set 2^V is needed. For our purpose we simply choose the entire power set:

$$\mathcal{G}_{2^V}{}^{(N)} := 2^{2^{V^N}} \tag{1.68}$$

This choice implies that measurability of a 2^{V^N}-valued function on some probability space which is a very strong requirement, but also implies that any map $f : 2^{V^N} \to 2^{V^M}$ is $\mathcal{G}_{2^V}{}^{(N)}$/$\mathcal{G}_{2^V}{}^{(M)}$-measurable. Since we are only interested in countably-valued random sets, the latter advantage overcompensates the former disadvantage.

Definition 1.4.1 - (**Random sets**)

► **Definition 1.4.1.1:** *Let $C := ((V, E), \mathfrak{S}, \mathfrak{T})$ be a causal model. A random set is a $\mathcal{F}_V$/$\mathcal{G}_{2^V}{}^{(1)}$-measurable map:*

$$\tau : \mathfrak{S} \to 2^V \tag{1.69}$$

► **Definition 1.4.1.2:** *Let $C := ((V,E),\mathfrak{S},\mathfrak{T})$ be a causal model and let $N \in \mathbb{N}$. An N-dimensional random set on C is a $\mathcal{F}_V/\mathcal{G}_{2^V}{}^{(N)}$-measurable map:*

$$\tau : \mathfrak{S} \to \left(2^V\right)^N \tag{1.70}$$

all $N \geq 1$ dimensional random sets will be summarized under the name multi-dimensional random sets.

In addition to a random set generalizing the stopping time, the formulation of the strong Markov property requires the specification of sets relative to this random set.

Definition 1.4.2 - (**Random sets relative to another random set**)

Let $C := ((V,E),\mathfrak{S},\mathfrak{T})$ be a causal model and let τ be an N-dimensional random set on C. A random set relative to τ is a map:

$$I : \mathrm{Range}(\tau) \to 2^V$$

Let $C := ((V,E),\mathfrak{S},\mathfrak{T})$ be a causal model and let τ be an N-dimensional random set on C. By the comment following Eq. 1.68 a random set I relative to τ naturally defines a random set, $I' := I(\tau)$. On the event $\{\tau = t\}$ the random variable $I'(\tau)$ takes on the value $I(t)$. This is the intuitive idea behind the concept of "random set relative to another one".
So far we have specified the vertex sets relative to a given random set, τ. We also need some notion of events that can be inferred from looking at a given vertex set once the value of some discrete random set is known. The associated σ-algebra is naturally defined as follows:

Definition 1.4.3 - (**Inference σ-algebras**)

Let $C := ((V,E),\mathfrak{S},\mathfrak{T})$ be a causal model, let τ be an N-dimensional discrete random set on C, and let I be a random set relative to τ. The following collection of events:

$$\mathcal{F}_I := \left\{ B \in \mathcal{F} \,\middle|\, B \cap \{\tau = W\} \in \mathcal{F}_{I(W)} \cap \{\tau = W\} \ \text{for every } W \in \mathrm{Range}(\tau) \right\}, \tag{1.71}$$

will be called inference σ-algebra of I relative to τ.

Remark 1.4.1 - (**Comment on Definition 1.4.3**)

► **Remark 1.4.1.1:** *The inference σ-algebra of I relative to τ is indeed a σ-algebra. To see this fix a countable collection $B_i \in \mathcal{F}_I$ (where $i \in \mathbb{N}$) and some $W \in \mathrm{Range}(\tau)$. For every $i \in \mathbb{N}$ there exists $\tilde{B}_i \in \mathcal{F}_{I(W)}$ such that $B_i \cap \{\tau = W\} = \tilde{B}_i \cap \{\tau = W\}$. Then*

$$\cap_{i\in\mathbb{N}} B_i \cap \{\tau = W\} = \cap_{i\in\mathbb{N}} \tilde{B}_i \cap \{\tau = W\} \in \mathcal{F}_{I(W)} \cap \{\tau = W\}\,.$$

Since $\mathcal{F}_{I(W)}$ is closed with respect to countable intersections, $\mathcal{F}_I$ is also closed with respect to countable intersections. Trivially $\Omega \in \mathcal{F}_I$. So it remains to show that $\mathcal{F}_I$ is also closed with respect to complementation. This follows from the observation that for every $B \in \mathcal{F}_I$ we have $B^C \cap \{\tau = W\} = \{\tau = W\} \cap (B \cap \{\tau = W\})^C$ since $\mathcal{F}_{I(W)}$ is closed with respect to complementation.

► **Remark 1.4.1.2:** *Inference σ-algebras are monotonous in the following sense: For two random sets I_A and I_B relative to some (multidimensional) random set, τ satisfying $I_A(r) \subseteq I_B(r)$ for every $r \in \mathrm{Range}(\tau)$ we have: $\mathcal{F}_{I_A} \subseteq \mathcal{F}_{I_B}$.*

We are only interested in scenarios where τ is countably-valued. In this case there exists another characterization of the inference σ-algebra:

Lemma 1.4.1 - (**Inference σ-algebras and countably-valued random sets**)

Let $C := ((V,E), \mathfrak{S}, \mathfrak{T})$ be a causal model, let τ be an N-dimensional, countably-valued random set on C and let

$$I : \mathrm{Range}(\tau) \to 2^V \tag{1.72}$$

be a random set relative to τ. Then

$$\mathcal{F}_I = \sigma\left(\{\tau = W\} \cap B_W \,\middle|\, W \in \mathrm{Range}(\tau), B_W \in \mathcal{F}_{I(W)}\right) \tag{1.73}$$

Proof of Lemma 1.4.1. We have $\{\tau = W\} \cap B_W \in \mathcal{F}_I$ for every $W \in \mathrm{Range}(\tau)$ and $B_W \in \mathcal{F}_{I(W)}$ by definition of $\mathcal{F}_I$. Therefore

$$\sigma\left(\{\tau = W\} \cap B_W \,\middle|\, W \in \mathrm{Range}(\tau), B_W \in \mathcal{F}_{I(W)}\right) \subseteq \mathcal{F}_I.$$

On the other hand by definition for every $B \in \mathcal{F}_I$ and $W \in \mathrm{Range}(\tau)$ there exist $B_W \in \mathcal{F}_{I(W)}$ such that $B \cap \{\tau = W\} = B_W \cap \{\tau = W\}$ and therefore by the countability assumption:

$$\begin{aligned} B = \cup_{W\in\mathrm{Range}(\tau)} \{\tau = W\} \cap B &= \cup_{W\in\mathrm{Range}(\tau)} \{\tau = W\} \cap B_W \\ &\in \sigma\left(\{\tau = W\} \cap B_W \,\middle|\, W \in \mathrm{Range}(\tau), B_W \in \mathcal{F}_{I(W)}\right), \end{aligned}$$

proving the second inclusion. ■

The concepts developed so far will be illustrated in a simple example now:

Example 1.4.1 - (**Stopping times and their inference maps**)

Let $C := ((\mathbb{N}_0, E), \mathfrak{S}, \mathfrak{T})$ be a causal model where $(a,b) \in E$ if and only if $a < b$ (this choice corresponds to a general process that can be constructed recursively by kernels. This choice of E does not make any further conditional independence assumptions; another canonical choice could be $(a,b) \in E$ if and only if $a + 1 = b$ corresponding to a discrete time Markov process). To simplify the situation we assume identical state spaces, i.e. $\mathfrak{S}_i := \mathbf{X}$ for some measurable space $(\mathbf{X}, \mathcal{F}_X)$. As before we fix an initial measure, $p \in M_1(\mathcal{F}_X)$,

inducing the law $\hat{P} \in M_1(\otimes_{v\in\mathbb{N}_0}\mathcal{F}_v)$ *and consider the process* $(\pi_v)_{v\in\mathbb{N}_0}$ *on the probability space* $\left(\mathfrak{S}, \otimes_{v\in\mathbb{N}_0}\mathcal{F}_v, \hat{P}\right)$.
Let

$$\tau : \mathfrak{S} \to \mathbb{N}_0$$

be a random time. As already mentioned we identify τ *with the random set* $\tilde{\tau} := \{\tau\}$. *Consider the "past of* τ*", i.e. the random set relative to* $\tilde{\tau}$ *given by:*

$$I : \{n\} \mapsto \{k \in \mathbb{N}\,|0 \le k \le n\} \tag{1.74}$$

The associated inference σ*-algebra* $\mathcal{F}_I$ *includes the event*

$$B := \cup_{0\le k\le\tau}\{X_k \in A\}\,. \tag{1.75}$$

for example. If τ *happens to be a stopping time with respect to the standard filtration,* $\mathbb{F}_n := \mathcal{F}_{\{0\le k\le n\}}$*, i.e.* $\{\tau \le n\} \in \mathbb{F}_n$*, then the associated stopping* σ*-algebra (compare Kallenberg [99], Bauer [22], König [105], Liptser and Shiryayev [119], Barndorff-Nielsen and Shiryaev [21]) is defined to be*

$$\mathcal{F}_\tau := \{A \in \mathcal{F}\,|A \cap \{\tau \le n\} \in \mathbb{F}_n\} \tag{1.76}$$

In this case obviously

$$\mathcal{F}_\tau = \mathcal{F}_I \tag{1.77}$$

Using the terminology introduced above, it is very simple to formalize all events that occur after τ *for example. They lie in the inference* σ*-algebra generated by:*

$$I_2 : \{n\} \mapsto \{k \in \mathbb{N}_0\,|k > n\} \tag{1.78}$$

As another example consider all events that occur after τ *and have an even label. The corresponding inference* σ*-algebra is generated by:*

$$I_3 : \{n\} \mapsto \{k \in \mathbb{N}_0\,|k > n, k \equiv 0 \mod 2\}\,. \tag{1.79}$$

A special random set relative to a given random set τ is the union of all entries of τ. Alluding to the usual stopping time scenario we call it present random set associated to τ:

Definition 1.4.4 - (**The present random set and the present σ-algebra associated to a discrete random set**)

Let $C := ((V, E), \mathfrak{S}, \mathfrak{T})$ *be a causal model and let* $\tau = (\tau_i)_{1\le i\le N}$ *be an* N*-dimensional discrete random set. The present random set associated to* τ *is a special random set relative to* τ*, namely:*

$$I_\tau : \mathrm{Range}\,(\tau) \to 2^V; W \mapsto \cup_{1\le i\le N} W_i \tag{1.80}$$

and the present σ*-algebra of* τ *is the associated inference* σ*-algebra:*

$$\mathcal{F}_{\mathrm{pr},\tau} := \mathcal{F}_{I_\tau} \tag{1.81}$$

Before continuing we define the concept of local equality of σ-algebras (compare Kallenberg [99] for example)

Definition 1.4.5 - (**Local equality of σ-algebras**)

Let $\mathcal{G}$ and $\mathcal{F}$ be two σ-algebras over some set, Ω, and let $A \in \mathcal{F} \cap \mathcal{G}$. Then $\mathcal{F}$ is equal to $\mathcal{G}$ on A, written as

$$\mathcal{F} = \mathcal{G} \text{ on } A$$

if $B \in \mathcal{F}$ implies $B \cap A \in \mathcal{G}$ and $B \in \mathcal{G}$ implies $B \cap A \in \mathcal{F}$.

The σ-algebra $\mathcal{F}_{\text{pr},\tau}$ captures all events that are completely specified by the value(s) of τ and the vertex values of vertices in $\tau(\omega)$. This is the essential message behind the following lemma:

Lemma 1.4.2 - (**Local properties of the present σ-algebras**)

Let $C := ((V, E), \mathfrak{S}, \mathfrak{T})$ be a causal model and let τ be an N-dimensional random set on C then for every $W \in \left(2^V\right)^N$:

$$\mathcal{F}_{\text{pr},\tau} = \sigma\left(\{B \cap \{\tau = W\}\}_{B \in \mathcal{F}_{W_1 \cup \ldots \cup W_N}}\right) \text{ on } \{\tau = W\} \tag{1.82}$$

Proof. To shorten notation we set $\mathcal{G} := \sigma\left(\{B \cap \{\tau = W\}\}_{B \in \mathcal{F}_{W_1 \cup \ldots \cup W_N}}\right)$ for this proof. First consider an arbitrary set $A \in \mathcal{F}_{\text{pr},\tau}$ with $A \subseteq \{\tau = W\}$. By definition of $\mathcal{F}_{\text{pr},\tau}$ there exists $\tilde{A} \in \mathcal{F}_{W_1 \cup \ldots \cup W_N}$ such that:

$$A \cap \{\tau = W\} = \tilde{A} \cap \{\tau = W\} \in \mathcal{G} \tag{1.83}$$

Hence $\mathcal{F}_{\text{pr},\tau} \subseteq \mathcal{G}$ on $\{\tau = W\}$.
Now let $A \in \mathcal{F}_{\cup_{1 \leq i \leq N} W_i}$. Then:

$$A \cap \{\tau = W\} \cap \{\tau = W'\} = \begin{cases} A \cap \{\tau = W'\} & \text{if } W = W' \\ \emptyset & \text{else} \end{cases} \tag{1.84}$$

and therefore $A \cap \{\tau = W\} \in \mathcal{F}_{\text{pr},\tau}$. Since sets of the form $A \cap \{\tau = W\}$ generate $\mathcal{G}$ the result follows. ■

Using the notation introduced above, the strong Markov property for causal models is captured by the following theorem:

Theorem 1.4.1 - (**Strong Markov property of causal models**)

Let $C := ((V, E), \mathfrak{S}, \mathfrak{T})$ *be a causal model with law* P *generated by some admissible initial measure* $p \in M_1(\mathcal{F}_{V_0})$*, let* $\tau = (\tau_i)_{1\le i\le N}$ *be an* N*-dimensional, countably-valued random set. Let* I_A *and* I_B *be random sets relative to* τ*. Assume that the following separation property holds:*
For every $W \in \mathrm{Range}(\tau)$ *there exist* $W_A, W_B \subseteq V$*,* $C_A \in \mathcal{F}_{W_A}$ *and* $C_B \in \mathcal{F}_{W_B}$ *such that*

- $\{\tau = W\} = C_A \cap C_B$
- *The set* $S := \cup_{1\le i\le N} W_i$ *d-separates*

$$\tilde{A} := [I_A(W) \cup W_A] \setminus S \text{ and } \tilde{B} := [I_B(W) \cup W_B] \setminus S$$

Then

$$\mathcal{F}_{I_A} \perp\!\!\!\perp \mathcal{F}_{I_B} \left| \mathcal{F}_{\mathrm{pr},\tau} \right. \tag{1.85}$$

The proof uses a technical lemma, that can be found in Kallenberg [99], p.105

Lemma 1.4.3 - (**Locality of the conditional expectation**)

Let $(\Omega, \mathcal{F}, P)$ *be a probability space, let* $\mathcal{G}_1, \mathcal{G}_2 \subseteq \mathcal{F}$ *be sub* σ*-algebras, let* $f_1, f_2 : \Omega \to \mathbb{R}$ *be a* $\mathcal{F}/\mathcal{B}_{\mathbb{R}}$*-measurable functions and let* $A \in \mathcal{G}_1 \cap \mathcal{G}_2$*. Assume that*

- $f_1 = f_2$ *almost surely on* A
- $\mathcal{G}_1 = \mathcal{G}_2$ *on* A

Then:

$$E[f_1 | \mathcal{G}_1] = E[f_2 | \mathcal{G}_2] \text{ almost surely on } A \tag{1.86}$$

Proof of Theorem 1.4.1. Fix $W \in \mathrm{Range}(\tau)$ and set

$$S_W := \cup_{1\le i\le N} W_i.$$

By the separation property there exists $C_A \in \mathcal{F}_{W_A}$, $C_B \in \mathcal{F}_{W_B}$ where $W_A, W_B \subseteq V$, such that

$$\{\tau = W\} = C_A \cap C_B,$$

S_W d-separates

$$I_A(W) \cup W_A \setminus S_W$$

and

$$I_B(W) \cup W_B \setminus S_W$$

Therefore by Theorem 1.3.1

$$\mathcal{F}_{I_A(W)\cup W_A \setminus S_W} \perp\!\!\!\perp \mathcal{F}_{I_B(W)\cup W_B \setminus S_W} | \mathcal{F}_{S_W}, \tag{1.87}$$

which implies

$$\mathcal{F}_{I_A(W)\cup W_A} \perp\!\!\!\perp \mathcal{F}_{I_B(W)\cup W_B} | \mathcal{F}_{S_W} \tag{1.88}$$

Moreover by assumption:

$$\mathcal{G}_{A,W} := \{\emptyset, \Omega, C_A, {C_A}^c\} \subseteq \mathcal{F}_{W_A}$$

and

$$\mathcal{G}_{B,W} := \{\emptyset, \Omega, C_B, {C_B}^c\} \subseteq \mathcal{F}_{W_B}$$

such that Eq. 1.88 implies

$$\left(\mathcal{F}_{I_A(W)}, \mathcal{G}_{A,W}\right) \perp\!\!\!\perp \left(\mathcal{F}_{I_B(W)}, \mathcal{G}_{B,W}\right) |\mathcal{F}_{S_W} \tag{1.89}$$

By Lemma 1.3.1:

$$\mathcal{F}_{I_A(W)} \perp\!\!\!\perp \mathcal{F}_{I_B(W)} \,|\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W} \tag{1.90}$$

Now we will show that

$$\mathcal{F}_{\text{pr},\tau} \;=\; \sigma\left(\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W}\right) \text{ on } \{\tau = W\} = C_A \cap C_B \tag{1.91}$$

To see this note that by Lemma 1.4.2 and the definition of C_A and C_B:

$$\begin{aligned} \mathcal{F}_{\text{pr},\tau} &= \sigma\left(\{D \cap \{\tau = W\}\}_{D \in \mathcal{F}_{S_W}}\right) && (1.92)\\ &= \sigma\left(\{D \cap C_A \cap C_B\}_{D \in \mathcal{F}_{S_W}}\right) \text{ on } \{\tau = W\} = C_A \cap C_B. \end{aligned}$$

Clearly $\sigma\left(\{D \cap C_A \cap C_B\}_{C \in \mathcal{F}_{S_W}}\right) \subseteq \sigma\left(\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W}\right)$. On the other hand every $D \in \sigma\left(\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W}\right)$ has the form

$$\begin{aligned} &D = && (1.93)\\ &(D_{1,1} \cap C_A \cap C_B) \cup (D_{0,1} \cap {C_A}^c \cap C_B) \cup (D_{1,0} \cap C_A \cap {C_B}^c) \cup (D_{1,1} \cap {C_A}^c \cap {C_B}^c) \end{aligned}$$

where $D_{i,j} \in \mathcal{F}_{S_W}$. This follows from the observation that all sets compatible with the right-hand side of 1.93 lie in $\sigma\left(\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W}\right)$, form a σ-algebra and contain any element of $\mathcal{F}_{S_W} \cup \mathcal{G}_{A,W} \cup \mathcal{G}_{B,W}$. Therefore elements $D \in \sigma\left(\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W}\right)$ with the further property $D \subseteq C_A \cap C_B$ can be written as:

$$D = D_{1,1} \cap C_A \cap C_B \in \sigma\left(\{D \cap C_A \cap C_B\}_{D \in \mathcal{F}_{S_W}}\right), \tag{1.94}$$

showing that $\sigma\left(\{D \cap C_A \cap C_B\}\Big|_{D \in \mathcal{F}_{S_W}}\right) = \sigma\left(\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W}\right)$ on $\{\tau = W\}$ and proving Eq. 1.91.
Now let $A' \in \mathcal{F}_{I_A}$ and $B' \in \mathcal{F}_{I_B}$. Then almost surely

$$P\left[A' \cap B' \,|\mathcal{F}_{\text{pr},\tau}\right] = \sum_{W \in \text{Range}(\tau)} P\left[A' \cap B' \cap \{\tau = W\} \,|\mathcal{F}_{\text{pr},\tau}\right] \tag{1.95}$$

Consider one particular summand for fixed $W \in \text{Range}(\tau)$. By definition of $\mathcal{F}_{I_A}$ and $\mathcal{F}_{I_B}$ there exist $A_W \in \mathcal{F}_{I_A(W)}$ and $B_W \in \mathcal{F}_{I_B(W)}$ such that

$$A' \cap \{\tau = W\} = A_W \cap \{\tau = W\} \text{ and } B' \cap \{\tau = W\} = B_W \cap \{\tau = W\}.$$

A short calculation gives:

$$\begin{aligned} & P\left[A' \cap B' \cap \{\tau = W\} \,|\mathcal{F}_{\text{pr},\tau}\right] \\ =\; & \mathbb{1}_{\{\tau = W\}} P\left[A_W \cap B_W \,|\sigma\left(\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W}\right)\right] \\ =\; & \mathbb{1}_{\{\tau = W\}} P\left[A_W \,|\sigma\left(\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W}\right)\right] \cdot P\left[B_W \,|\sigma\left(\mathcal{F}_{S_W}, \mathcal{G}_{A,W}, \mathcal{G}_{B,W}\right)\right] \\ =\; & \mathbb{1}_{\{\tau = W\}} P\left[A' \,|\mathcal{F}_{\text{pr},\tau}\right] \cdot P\left[B' \,|\mathcal{F}_{\text{pr},\tau}\right], \end{aligned}$$

where we used Lemma 1.4.3 together with Eq. 1.91 for the step from line 1 to line 2 and from line 3 to line 4, and the conditional independence, Eq. 1.90, for the step from line 2 to line 3. Inserting this into the right-hand side of Eq. 1.95 and carrying out the summation yields:

$$P\left[A' \cap B' \,|\mathcal{F}_{\text{pr},\tau}\right] = P\left[A' \,|\mathcal{F}_{\text{pr},\tau}\right] \cdot P\left[B' \,|\mathcal{F}_{\text{pr},\tau}\right] \tag{1.96}$$

Since this holds true for any $A' \in \mathcal{F}_{I_A}$ and $B' \in \mathcal{F}_{I_B}$ the validity of the theorem follows. ■

We will use this theorem in Chapter 4 when we investigate learning algorithms on the sensorimotor loop. As a general comment note that the Markov property for time discrete Markov processes is really a special case of the strong Markov property for causal models.

Example 1.4.2 - (**Strong Markov properties of discrete time Markov processes**)

► **Example 1.4.2.1:** *Define* $G_{\text{Markov}} := (\mathbb{N}_0, E_{\text{Markov}})$ *where* $(a, b) \in E_{\text{Markov}}$ *if and only if* $a + 1 = b$. *Let* $C := (G_{\text{Markov}}, \mathfrak{S}, \mathfrak{T})$ *be a causal model and let* τ *be a stopping time. For simplicity we assume that the vertex state spaces are identical:* $\mathfrak{S}_v = \mathbf{X}$ *where* $(\mathbf{X}, \mathcal{F}_X)$ *is some measurable space. Write* $\tilde{\tau} := \{\tau\}$ *for the corresponding random set. Define*

$$I_{<\tau} : \{n\} \mapsto \{0, \ldots, n-1\} \ ; I_{>\tau} : \{n\} \mapsto \{k \in \mathbb{N}_0 \,|k > n\} \tag{1.97}$$

Since τ *is a stopping time, we have:*

$$\{\tilde{\tau} = \{n\}\} \in \mathcal{F}_{\{0,\ldots,n\}} \tag{1.98}$$

Moreover for every $n \in \mathbb{N}_0$ *the set* $\{n\}$ *d-separates*

$$\tilde{A} := I_{<\tau}(\{n\}) \cup \{0, \ldots, n\} \setminus \{n\} = \{0, \ldots, n-1\} \tag{1.99}$$

and

$$\tilde{B} := I_{>\tau}(\{n\}) \setminus \{n\} = \{k \in \mathbb{N}_0 \,|k > n\} \tag{1.100}$$

therefore

$$\mathcal{F}_{I_{<\tau}} \perp\!\!\!\perp \mathcal{F}_{I_{>\tau}} \,|\mathcal{F}_{\text{pr},\tilde{\tau}} \tag{1.101}$$

by Theorem 1.4.1. This is the strong Markov property for discrete time Markov processes.
► **Example 1.4.2.2:** *As a slightly more involved example that is covered by the theorem consider the same causal model and two stopping times* τ_1 *and* τ_2. *The two stopping times naturally defines the two-dimensional random set:*

$$\tau' : \omega \mapsto (\{\tau_1(\omega)\}, \{\tau_2(\omega)\}) \tag{1.102}$$

Assume that $\tau_2 > \tau_1 + 1$ *almost surely. By the former findings we have*

$$\pi_{\tau_1 - 1} \perp\!\!\!\perp \pi_{\tau_1 + 1} \,|\mathcal{F}_{\text{pr},\tilde{\tau}_1}$$

Does this imply

$$\pi_{\tau_1 - 1} \perp\!\!\!\perp \pi_{\tau_1 + 1} \,|\mathcal{F}_{\text{pr},\tau'}\,?$$

The answer to this question is negative. Consider an IID sequence of random variables with values in $\{0, 1\}$ *each occurring with probability one half. Define*

$$\tau_1(s) = 3; \qquad \tau_2(s) = \begin{cases} 5 \text{ if } s_2 = s_4 \\ 6 \text{ else} \end{cases}$$

It is easy to see that τ_1 *and* τ_2 *are stopping times. We have*

$$P\left[\{\pi_2 = a\} \cap \{\pi_4 = b\} \,|\mathcal{F}_{\text{pr},\tau'}\right] = \begin{cases} \frac{1}{2}\delta_{a,b} & \text{if } \tau_2 - \tau_1 = 2 \\ \frac{1}{2}(1 - \delta_{a,b}) & \text{else} \end{cases}$$

but

$$P\left[\{\pi_2 = a\}\left|\mathcal{F}_{\text{pr},\tau'}\right.\right] = P\left[\{\pi_4 = b\}\left|\mathcal{F}_{\text{pr},\tau'}\right.\right] = \frac{1}{2}$$

It is very illustrative to analyze which assumptions of Theorem 1.4.1 fail and under which further requirements on τ_1 *and* τ_2 *the conditional independence result*

$$\mathcal{F}_{I_{<\tau_1}} \perp\!\!\!\perp \mathcal{F}_{I_{>\tau_1}}\left|\mathcal{F}_{\text{pr},\tau'}\right. \tag{1.103}$$

is guaranteed to be true nevertheless, where we set:

$$I_{<\tau_1} : (\{n\}, \{m\}) \mapsto \{0, \ldots, n\} \tag{1.104}$$

and

$$I_{>\tau_1} : (\{n\}, \{m\}) \mapsto \{k \in \mathbb{N}\,|k > n\}\,. \tag{1.105}$$

► **Example 1.4.2.3:** *Since* τ_1 *and* τ_2 *are stopping times*

$$\tau'(\{n\}, \{m\}) = C_{1,n} \cap C_{2,m} \tag{1.106}$$

where

$$C_{1,n} \in \mathcal{F}_{\{0,\ldots,n\}} \text{ and } C_{2,m} \in \mathcal{F}_{\{0,\ldots,m\}}$$

For fixed $m > n$*:*

$$S := \{n, m\}$$

clearly d-separates

$$\tilde{A} := I_{<\tau_1}(\{n\}, \{m\}) \setminus S = \{k \in \mathbb{N}\,|0 \leq k < n\}$$

and

$$\tilde{B} := I_{>\tau_1}(\{n\}, \{m\}) \setminus S = \{k \in \mathbb{N}\,|k > n, k \neq m\}\,.$$

However whenever $m > n$ *it is not possible to append the sets* $\{0, \ldots, n\}$ *and* $\{0, \ldots, m\}$*, to* A *or* B *in a way that preserves the separation property.*

However under further assumptions on τ_1 *and* $\tau_2 \geq \tau_1$ *the conditional independence result might still hold. One (admittedly rather uninteresting) possibility is that the second stopping time does not "look at values that occur after the first one". A trivial example is* $\tau_2 = \tau_1 + K$ *for some fixed integer constant* $K > 1$*. Then*

$$\tau'((\{n\}, \{m\})) \in \mathcal{F}_{\{0,\ldots,n\}} \tag{1.107}$$

and, as in the single stopping time example,

$$S = \{n, m\} \tag{1.108}$$

d-separates

$$\tilde{A} := I_{<\tau_1}((\{n\}, \{m\})) \cup \{0, \ldots, n\} \setminus S = \{0, \ldots, n-1\} \tag{1.109}$$

and

$$\tilde{B} := I_{>\tau_1}((\{n\}, \{m\})) \setminus S = \{k \in \mathbb{N}\,|k > n, k \neq m\} \tag{1.110}$$

such that Eq. 1.103 is satisfied by Theorem 1.4.1.

► **Example 1.4.2.4:** *A more interesting case is that* τ_2 *depends on values after* τ_1 *only. This happens for example if* τ_1 *and* τ_2 *are the first and second hitting time of a set* $A \in \mathcal{F}_X$*. Then*

$$\{\tau' = (\{n\}, \{m\})\} = C_{1,n} \cap C_{2,n,m} \tag{1.111}$$

where

$$C_{1,n} = \cap_{0 \leq k < n}\{\pi_k \notin A\} \cap \{\pi_n \in A\} \in \mathcal{F}_{\{0,\ldots,n\}} \tag{1.112}$$

and

$$C_{2,n,m} = \cap_{n<k<m} \{\pi_k \notin A\} \cap \{\pi_m \in A\} \in \mathcal{F}_{\{n+1,\dots,m\}} \tag{1.113}$$

Then

$$S = \{n, m\}$$

d-separates

$$\tilde{A} := I_{<\tau_1}\left((\{n\}, \{m\})\right) \cup \{0,\dots,n\} \setminus S = \{0,\dots,n-1\} \tag{1.114}$$

and

$$\tilde{B} := I_{>\tau_1}\left((\{n\}, \{m\})\right) \cup \{n+1,\dots,m\} \setminus S = \{k \in \mathbb{N}_0 \,|k > n, k \neq m\} \tag{1.115}$$

Therefore

$$\mathcal{F}_{I_{<\tau_1}} \perp\!\!\!\perp \mathcal{F}_{I_{>\tau_1}} \left| \mathcal{F}_{\text{pr},\tau'} \right. \tag{1.116}$$

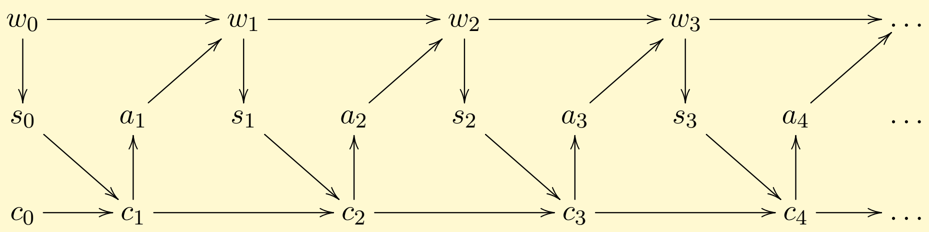
by Theorem 1.4.1.

Now we will give some examples for more general causal models:

Example 1.4.3 - (**Graph separation and conditional independence in Causal models**)

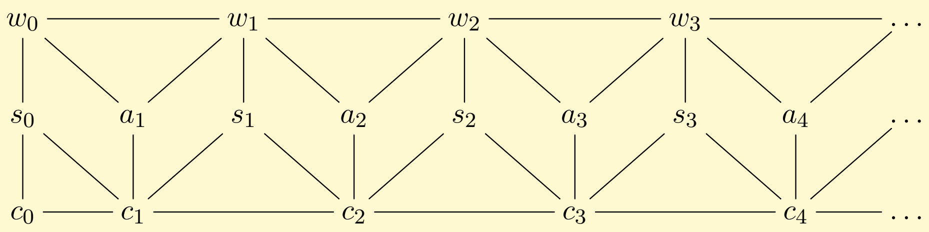
Let $C = (G, \mathfrak{S}, \mathfrak{T})$ be a causal model. We use the convention from Example 1.3.1 again: vertices of the graph will be expressed by lowercase letters and the corresponding vertex random variables, π_v, will be denoted by upper case letters.

Consider an agent interacting with the environment by controlling some motor value. This can be modelled by the following causal model (where w_i denotes the i-th state of the world, s_i denotes the i-th sensor value, a_i denotes the i-th action and c_i denotes the i-th state of the agent's memory):



Caus. mod. 15 *- Example: Agent interacting with the world*

The corresponding moral graph is:



Caus. mod. 16 *- Example: Moral graph of Caus. mod. 15*

Let τ_1 be a stopping time with respect to the filtration

$$\mathbb{G}_{\text{all},n} := \sigma\left(\{S_i\}_{0\le i\le n} \cup \{A_i\}_{1\le i\le n} \cup \{C_i\}_{0\le i\le n+1} \cup \{W_i\}_{0\le i\le n}\right).$$

The events in $\mathbb{G}_{\text{all},n}$ can be inferred from looking at any collection of vertices prior to c_{n+1}. The strong Markov property implies that the process of actions after τ_1 is independent of the actions before τ_1 given the world state at τ_1 and the memory value at τ_1:

$$(A_i)_{i>\tau_1} \perp\!\!\!\perp (A_i)_{1\leq i\leq\tau_1} \left|(W_{\tau_1}, C_{\tau_1+1}, \tau_1)\right. \tag{1.117}$$

Now let τ_2 be a stopping time with respect to the filtration

$$\mathbb{G}_{\text{agent},n} := \sigma\left(\{S_i\}_{0\leq i\leq n} \cup \{A_i\}_{1\leq i\leq n} \cup \{C_i\}_{0\leq i\leq n+1}\right).$$

The events in $\mathbb{G}_{\text{agent},n}$ can be inferred from looking at any collection of memory values, actions and sensor values (i.e. quantities directly accessible to the agent) prior to c_{n+1}. By the strong Markov property:

$$(W_i)_{0\leq i\leq\tau_2} \perp\!\!\!\perp (C_i)_{0\leq i\leq\tau_2+1} \Big|(S_i)_{0\leq i\leq\tau_2}, (A_i)_{1\leq i\leq\tau_2}, \tau_2 \tag{1.118}$$

After these examples we outline some further consequences of Theorem 1.4.1. The first immediate corollary is essentially a reformulation that holds true if there exist certain regular versions of the conditional probability distribution:

Corrolary 1.4.1 - (**Reformulation of strong Markov property**)

Let $C := ((V,E), \mathfrak{S}, \mathfrak{T})$ be a causal model and let $\tau = (\tau_i)_{1\leq i\leq N}$ be an N-dimensional, countably-valued random set. Let I_A and I_B be random sets relative to τ and assume that the separation property of Theorem 1.4.1 is satisfied. Assume that for every $W \in \operatorname{Range}(\tau)$ there exists a regular version K_W of $\mathcal{F}_{I_B(W)}$ given $\mathcal{F}_{\cup_{1\leq i\leq N}W_i}$, i.e.:

$$P\left[C\left|\mathcal{F}_{\cup_{1\leq i\leq N}W_i}\right.\right] = K_W\left(\pi_{\cup_{1\leq i\leq N}W_i}, C\right) \text{ for every } C \in \mathcal{F}_{I_B(W)} \tag{1.119}$$

Let $\tilde{B} \in \mathcal{F}_{I_B}$. By definition for every $W \in \operatorname{Range}(\tau)$ there exist $B_W \in \mathcal{F}_{I_B(W)}$ such that $\tilde{B} \cap \{\tau = W\} = B_W \cap \{\tau = W\}$. Then:

$$P\left[\tilde{B}\left|\mathcal{F}_{\text{pr},\tau}, \mathcal{F}_{I_A}\right.\right] = \sum_{W\in\operatorname{Range}(\tau)} \mathbb{1}_{\{\tau=W\}} K_W\left(\pi_{\cup_{1\leq i\leq N}W_i}, B_W\right) \tag{1.120}$$

We restricted the investigation to N-tuples of random sets but the results are also true for sequences of random sets, $(\tau_i)_{i\in\mathbb{N}}$. This is straight forward and we omit a proof.

In this work we are mainly interested in directed models. However the concepts developed so far (like random sets, random sets relative to another random set and the associated σ-algebras for example) are only related to the vertex set and make sense for undirected graphical models as well. Moreover a closer look on the proof of the strong Markov property shows that the directedness entered only very indirectly, namely at the point where we used the relation between graph separation and conditional independence for deterministically chosen sets. A proof of a strong Markov property for undirected models is essentially identical and we only cite the result:

Remark 1.4.2 - (**Strong Markov properties for undirected models**)

Let $G = (V,E)$ be an undirected graph and let $(\mathfrak{S}_v, \mathcal{F}_v)$ where $v \in V$ be measurable spaces. Define the measurable space of total configurations:

$$\mathfrak{S} := \prod_{v\in V} \mathfrak{S}_v \;;\; \mathcal{F} := \otimes_{v\in V}\mathcal{F}_v \tag{1.121}$$

As before let $\pi_v : \mathfrak{S} \to \mathfrak{S}_v$ *denote the projection onto the* v*-th factor and set* $\mathcal{F}_W := \sigma\left(\{\pi_v\}_{v \in W}\right)$ *for any* $W \subseteq V$. *Let* $P \in M_1(\mathcal{F})$ *be a probability law on the space of configurations equipped with product* σ*-algebra satisfying the global Markov property with respect to G, i.e.:*

$$\mathcal{F}_A \perp\!\!\!\perp \mathcal{F}_B \,|\mathcal{F}_S \tag{1.122}$$

whenever $S, A, B \subseteq V$ *such that* S *separates* A *and* B.
Let

$$\tau : \mathfrak{S} \to \left(2^V\right)^N \tag{1.123}$$

be $\mathcal{F}/\mathcal{G}_{2^V}{}^{(N)}$*-measurable. Assume* τ *to be countably-valued and let* I_A, I_B *be random sets relative to* τ*. Assume that the following separation property holds:*
For every $W \in \mathrm{Range}(\tau)$ *there exist* $W_A, W_B \in 2^V$, $C_A \in \mathcal{F}_{W_A}$ *and* $C_B \in \mathcal{F}_{W_B}$ *such that*

- $$\{\tau = W\} = C_A \cap C_B \tag{1.124}$$

- *The set*
$$S := \cup_{1 \leq k \leq N} W_k$$
separates
$$\tilde{A} := I_A(W) \cup W_A \setminus S \text{ and } \tilde{B} := I_B(W) \cup W_B \setminus S$$

then

$$\mathcal{F}_{I_A} \perp\!\!\!\perp \mathcal{F}_{I_B} \,|\mathcal{F}_{\mathrm{pr},\tau} \tag{1.125}$$

Chapter 2

Projected stochastic gradient algorithms

In the current chapter we will provide a collection of mathematical tools including concepts from optimization theory, set-valued analysis and projected differential inclusions. We will prove a theorem on the asymptotic behavior of a collection of stochastic approximation algorithms (see and Theorem 2.4.1 and Theorem 2.4.3). Our result partially extends results from Kushner and Yin [110] and Kushner and Clark [109]. Stochastic approximation sequences are a frequently used method in machine learning and optimization theory. Most of the background knowledge on stochastic optimization theory provided in this chapter is based in the excellent books Kushner and Clark [109], Kushner and Yin [110], Bertsekas and Tsitsiklis [24] and Borkar [38]

The main result of this chapter is a theorem about the asymptotic behavior of a projected stochastic approximation algorithm and a convergence theorem for a stochastic general-metric gradient ascent algorithm (see and Theorem 2.4.1 and Theorem 2.4.3). For the theorems we keep the regularity requirements on the metric for the gradient algorithm and the objective funcion as small as possible. We require the set of critical values to be "sufficiently small" (to be precised later on). Our requirement is true whenever the stationary points are isolated but holds true for a much broader class of functions. Furthermore we do not restrict the gradient ascent algorithms to be with respect to continuous metrices but allow for certain discontinuities or more generally for certain selections from an entire range of metrics. For the projection onto the desired constraint set we use general quasi-projectors (see Aubin [7] and Definition 2.3.1). The use of quasi-projectors different from othorgonal projection with respect to Euclidean metric is advantageous if the latter is computationally expensive or even intractable. These general quasi-projectors usually introduce spurious stationary points on the boundary of the domain as we will explain in detail. We circumvent this problem by doing the gradient ascent with respect to a metric that is specially adapted to the specific quasi-projector in use. This idea is essentially new to our knowledge. Therefore the definition of the preimage cone in Definition 2.3.1, the associated compatibility condition between metric and quasi-projector and the successive examples can also be considered to be a central result of the current chapter.

We will provide many examples to illustrate the concepts and theorems. As examples for quasi-projectors we consider the orhorgonal projection onto the ϵ-simplex in $\mathbb{R}^n$ and the canonical retraction onto the unit ball in the vector space of square matrices equipped with operator norm, i.e.:

$$A \mapsto \begin{cases} \frac{A}{\|A\|_{\mathrm{Op}}} & \text{if } \|A\| > 1 \\ A & \text{else} \end{cases}$$

For each of these quasi-projectors we explicitely provide a compatible metric. Both examples aim at the second part of this thesis where we use a projected stochastic gradient algorithm to solve certain learning problems in the sensorimotor loop.

2.1 First order optimality in constrained optimization problems

In this section we will review some concepts frequently used in mathematical optimization. Many of the concepts posses very natural extensions to more general constraint sets, infinite dimensional vector spaces, functions satisfying milder differentiability requirements (compare for example Clarke [50], Jahn [93], Aubin [7], Aubin and Frankowska [8], Anger, Aubin, and Cellina [6], Troutman [185] and Troutman [185]). To keep this section short we will try to find a compromise between generality and technical simplicity. We first try to formulate first order optimality conditions for the following optimization problem:

Problem 2.1.1 - (**General constrained optimization problem**)

Let $\phi \in C_1(\mathbb{R}^n, \mathbb{R})$*,* $K \subseteq \mathbb{R}^n$ *and assume that*

$$M := \max\{\phi(x)\,|x \in K\} \tag{2.1}$$

exists. In this case: determine M *and find the maximizers:*

$$C := \{x \in K\,|M = \phi(x)\} \tag{2.2}$$

In order to formulate first-order optimality conditions some notion of achievable, infinitesimal changes at elements $x \in K$ is needed. This idea is captured by the (Bouligand) tangent cone (compare Clarke [50], Aubin and Frankowska [8], Jahn [93]) [1]:

Definition 2.1.1 - (**Tangent cone and normal cone**)

Let $K \subseteq \mathbb{R}^n$ *and* $x \in K$*. Then the (Bouligand) tangent cone of* K *at* x *is:*

$$T_K(x) = \left\{ v \in \mathbb{R}^n \,\middle|\, \liminf_{t \searrow 0} \frac{d_K(x+tv)}{t} = 0 \right\} \tag{2.3}$$

where $d_K(x) := \inf\{y \in K\,|\|x-y\|_2\}$ *and* $\|\cdot\|_2$ *is the Euclidean distance. The normal cone at* x *is the polar cone of* $T_K(x)$*:*

$$N_K(x) = \{v \in \mathbb{R}^{n*}\,|v\,[w] \leq 0 \text{ for every } w \in T_K(x)\} \tag{2.4}$$

Before continuing the excursion we will present some important tangent and cotangent cones. As a first example consider the closed R-ball, $\overline{B_R(0)} := \{x \in \mathbb{R}^n\,|\|x\|_2 \leq R\}$ of $\mathbb{R}^n$ equipped with Euclidean norm:

---

[1] There are actually plenty of ways to define tangent cones to sets but we restrict to the Bouligand tangent cone, since it is the right concept for the validity of the viability theorem that we will need later on. For a a good overview of this topic and an illustration of different concepts of tangency see Aubin and Frankowska [8]

Example 2.1.1 - (**Tangent and normal cone for a closed ball in** $\mathbb{R}^n$)

Let $\overline{B_R(0)} := \{x \in \mathbb{R}^n \,|\, \|x\|_2 \leq R\}$, *then:*

$$T_{B_R(0)}(x) = \begin{cases} \mathbb{R}^n & \text{for } \|x\| < R \\ \{v \in \mathbb{R}^n \,|\langle v, x\rangle \leq 0\} & \text{for } \|x\| = R \end{cases} \tag{2.5}$$

and

$$N_{B_R(0)}(x) = \begin{cases} 0 & \text{for } \|x\| < R \\ \{\alpha \langle x, \cdot\rangle \,|\alpha \in \mathbb{R}_{\geq 0}\} & \text{for } \|x\| = R \end{cases} \tag{2.6}$$

Another constraint set (that will play an important role in Chapter 4) is the $\epsilon-$simplex (with $\mathbf{S}$ being a finite set and $0 \leq \epsilon \leq \frac{1}{|\mathbf{S}|}$):

Example 2.1.2 - (**Tangent and normal cone for the** ϵ**-simplex**)

Let

$$\Delta_{\epsilon;\mathbf{S}} := \left\{ v \in \mathbb{R}^{\mathbf{S}} \,\middle|\, \sum_{s\in\mathbf{S}} v_s = 1; v_s \geq \epsilon \text{ for every } s \in \mathbf{S} \right\} \tag{2.7}$$

be the ϵ*-simplex over the finite set,* $\mathbf{S}$. *Then*

$$T_{\Delta_{\epsilon;\mathbf{S}}}(x) = \left\{ c \in \mathbb{R}^{\mathbf{S}} \,\middle|\, \sum_{s\in\mathbf{S}} c_s = 0; c_s \geq 0 \text{ whenever } x_s = \epsilon \right\} \tag{2.8}$$

and

$$\begin{aligned} N_{\Delta_{\epsilon;\mathbf{S}}}(x) &= \left\{ \lambda \underline{1}^T - \sum_{s\in I(x)} \lambda_s e_{s,*} \,|\lambda \in \mathbb{R}; \lambda_s \in \mathbb{R}_{\geq 0} \right\} \\ &= \left\{ \lambda \underline{1}^T + \sum_{s\in I(x)} \lambda_s \left(\frac{1}{|\mathbf{S}|}\underline{1}^T - e_{s,*} \right) |\lambda \in \mathbb{R}; \lambda_s \in \mathbb{R}_{\geq 0} \right\} \end{aligned} \tag{2.9}$$

where $I(x) := \{s \in \mathbf{S} \,|x_s = \epsilon\}$ *and* $\underline{1}^T$ *acts on vectors via matrix multiplication, i.e.*

$$\underline{1}v := \sum_{s\in\mathbf{S}} v_s$$

Later on we will also need the tangent cone of the closed unit ball $\overline{B_1(0)} \subseteq \mathbb{R}^{\mathbf{S}\times\mathbf{S}}$ where $\mathbb{R}^{\mathbf{S}\times\mathbf{S}}$ is equipped with some matrix operator norm:

Lemma 2.1.1 - (**Tangent and normal cone for the unit ball in** $\mathbb{R}^{\mathbf{S}\times\mathbf{S}}$ **equipped with operator norm**)

Let $\overline{B_1(0)}$ *be the closed unit ball of* $\mathbb{R}^{\mathbf{S}\times\mathbf{S}}$ *equipped with operator norm*

$$\|A\|_{\mathrm{Op},p} := \sup \left\{ \|Ax\|_p \,\middle|\, x \in \mathbb{R}^{\mathbf{S}}; \|x\|_p = 1 \right\} \tag{2.10}$$

where $\|\cdot\|_p$ *(with* $1 \leq p \leq \infty$*) is the p-norm:*

$$\|v\|_p := \left(\sum_{s\in\mathbf{S}} |v_s|^p \right)^{\frac{1}{p}} \text{ for } p < \infty \text{ and } \|v\|_\infty := \max\{|v_s| \,|s \in \mathbf{S}\}$$

for $A \in \overline{B_1(0)}$ *with* $\|A\|_{\mathrm{Op},p} = 1$ *set:*

$$K_A := \left\{(\lambda, v) \in \mathbb{R}^{\mathbf{S}*} \times \mathbb{R}^{\mathbf{S}} \,\middle|\, \|v\|_p = \|\lambda\|_{p*} = 1 \,;\, \lambda\,(Av) = \|A\|_{\mathrm{Op},p} = 1\right\} \tag{2.11}$$

where we wrote $\|\cdot\|_{p*}$ *for the dual norm of* $\|\cdot\|_p$ *(by a standard result from functional analysis this norm is equal to the q-norm where* $\frac{1}{p} + \frac{1}{q} = 1$*). Then*

$$N_{B_1(0)}\,(A) = \begin{cases} \overline{\operatorname{convcone}\left(\{X \mapsto \lambda\,(Xv)\,|(\lambda, v) \in K_A\}\right)} & \text{if } \|A\|_{\mathrm{Op},p} = 1 \\ 0 & \text{else} \end{cases}, \tag{2.12}$$

where $\operatorname{convcone}(A)$ *is the convex cone generated by the set A, and*

$$T_{B_1(0)}\,(A) := \begin{cases} \left\{X \in \mathbb{R}^{\mathbf{S}\times\mathbf{S}}\,|\lambda\,(Xv) \leq 0 \text{ for all } (\lambda, v) \in K_A\right\} & \text{if } \|A\|_{\mathrm{Op},p} = 1 \\ \mathbb{R}^{\mathbf{S}\times\mathbf{S}} & \text{else} \end{cases} \tag{2.13}$$

Proof of Lemma 2.1.1. First of all note that $B_1(0)$ is a convex set. This implies that the tangent cone, $T_{B_1(0)}$, is convex, closed and the polar cone of $N_{B_1(0)}$ (compare Aubin and Frankowska [8]). If $\|A\|_{\mathrm{Op},p} < 1$ then for every $X \in \mathbb{R}^{\mathbf{S}\times\mathbf{S}}$ we have $A + tX \in B_1(0)$ for sufficiently small $t \in \mathbb{R}_{>0}$ such that

$$T_{B_1(0)}\,(A) = \mathbb{R}^{\mathbf{S}\times\mathbf{S}}$$

which proves the statement for $\|A\|_{\mathrm{Op},p} < 1$.

Let $B_{\mathbb{R}^{\mathbf{S}},1}(0)$ denote the unit ball in $(\mathbb{R}^{\mathbf{S}}, \|\cdot\|_p)$ and let $B_{\mathbb{R}^{\mathbf{S}*},1}(0)$ denote the unit ball in $(\mathbb{R}^{\mathbf{S}*}, \|\cdot\|_{p*})$. Whenever $\lambda \in B_{\mathbb{R}^{\mathbf{S}*},1}(0)$ and $v \in B_{\mathbb{R}^{\mathbf{S}},1}(0)$ then

$$\lambda\,(Bv) \leq 1$$

for every $B \in \overline{B_1(0)}$ such that for every $(\lambda, v) \in K_A$

$$\inf\left\{\frac{\lambda\left((A + tX)\,v\right) - \lambda(Bv)}{t}\,\middle|\, B \in \overline{B_1(0)}; t > 0\right\} \geq \lambda(Xv) \tag{2.14}$$

which implies $\liminf_{t\searrow 0} \frac{d\left(A+tX, \overline{B_1(0)}\right)}{t} > 0$ whenever $\lambda\,(Xv) > 0$. Hence

$$T_{B_1(0)}\,(A) \subseteq \left\{X \in \mathbb{R}^{\mathbf{S}\times\mathbf{S}}\,|\lambda\,(Xv) \leq 0 \text{ for every } (\lambda, v) \in K_A\right\} \tag{2.15}$$

For $\epsilon > 0$ define

$$\tilde{T}_\epsilon\,(A) := \left\{X \in \overline{B_1(0)}\,|\lambda\,(Xv) \leq -\epsilon \text{ for all } (\lambda, v) \in K_A\right\}, \tag{2.16}$$

and set

$$K_{A,\epsilon} := \left\{(\lambda, v) \in \overline{B_{\mathbb{R}^{\mathbf{S}*},1}(0)} \times \overline{B_{\mathbb{R}^{\mathbf{S}},1}(0)}\,\middle|\,\lambda\,(Xv) < -\frac{\epsilon}{2} \text{ for all } X \in \tilde{T}_\epsilon\,(A)\right\} \tag{2.17}$$

By compactness of $\overline{B_1(0)}$ the set $K_{A,\epsilon}$ is an open subset of the compact set $\overline{B_{\mathbb{R}^{\mathbf{S}*},1}(0)} \times \overline{B_{\mathbb{R}^{\mathbf{S}},1}(0)}$. This together with $K_A \subseteq K_{A,\epsilon}$ implies

$$M_\epsilon := \sup\left\{\lambda\,(Av)\,\middle|\,(\lambda, v) \in \overline{B_{\mathbb{R}^{\mathbf{S}},1}(0)} \times \overline{B_{\mathbb{R}^{\mathbf{S}*},1}(0)} \setminus K_{A,\epsilon}\right\} < 1 \tag{2.18}$$

Now fix an arbitrary $X \in \tilde{T}_\epsilon\,(A)$ and $(\lambda, v) \in \overline{B_{\mathbb{R}^{\mathbf{S}*},1}(0)} \times \overline{B_{\mathbb{R}^{\mathbf{S}},1}(0)}$. There are two possibilities:

- $(\lambda, v) \in K_{A,\epsilon}$, implying that $\lambda\,(Av) \leq 1$ and $\lambda\,(Xv) \leq -\frac{\epsilon}{2}$

- $(\lambda, v) \notin K_{A,\epsilon}$, implying that $\lambda(Av) \leq M_\epsilon$ and $\lambda(Xv) \leq 1$

Therefore whenever $X \in \tilde{T}_\epsilon(A)$ and $0 \leq t \leq 1 - M_\epsilon$ then:

$$\lambda((A+tX)v) \leq 1 \text{ for every } (\lambda, v) \in \overline{B_{\mathbb{R}^{\mathbf{S}},1}(0)} \times \overline{B_{\mathbb{R}^{\mathbf{S}*},1}(0)}$$

such that by the Hahn-Banach theorem:

$$\|A+tX\|_{\mathrm{Op},p} \leq 1, \text{ i.e. } A+tX \in \overline{B_1(0)} \tag{2.19}$$

since $T_{B_1(0)}(A)$ is a convex cone this implies

$$\operatorname{convcone}\left(\tilde{T}_\epsilon(A)\right) \subseteq T_{B_1(0)}(A) \tag{2.20}$$

Since the tangent cone is closed this implies

$$\begin{aligned} &\left\{X \in \mathbb{R}^{\mathbf{S}\times\mathbf{S}} \,|\lambda(Xv) \leq 0 \text{ for every } (\lambda, v) \in K_A\right\} \\ = &\ \overline{\cup_{\epsilon>0} \operatorname{convcone}\left(\tilde{T}_\epsilon\right)} \subseteq T_{B_1(0)}(A), \end{aligned}$$

which proves $T_{B_1(0)}(A) = \left\{X \in \mathbb{R}^{\mathbf{S}\times\mathbf{S}} \,|\lambda(Xv) \leq 0 \text{ for every } (\lambda, v) \in K_A\right\}$. Since the normal cone, as we defined it, is always a closed, convex cone (compare Aubin and Frankowska [8]), this also proves the claim for the normal cone. ■

Most generally the first order optimality condition for the constrained optimization problem can be formulated as follows:

Theorem 2.1.1 - (First order optimality condition - abstract version)

Consider Problem 2.1.1. Any $x_* \in C$ satisfies:

$$D\phi(x_*)[v] \leq 0 \text{ for all } v \in T_K(x_*), \tag{2.21}$$

i.e. $D\phi(x_*) \in N_K(x_*)$.

Frequently the set K in Problem 2.1.1 is specified by equality or inequality constraints. The following theorem specifies the tangent cone in this situation (compare Clarke [50] and Jahn [93]):

Theorem 2.1.2 - (Sets defined by equality and inequality constraints)

▶ **Theorem 2.1.2.1:** *Let $f_i \in C_1(\mathbb{R}^n, \mathbb{R})$ where $1 \leq i \leq k$. Let*

$$K := \{x \in \mathbb{R}^n \,|f_i(x) \leq 0 \text{ for all } 0 \leq i \leq k\} \tag{2.22}$$

Let $x \in K$ and set $I(x) := \{i \in \{1, \ldots, k\} \,|f_i(x) = 0\}$. Assume that the collection $\{Df_i(x)\}_{i\in I(x)}$ is positively linearly independent, i.e.

$$\sum_{i\in I(x)} \lambda_i Df_i(x) = 0 \text{ and } \lambda_i \geq 0 \text{ implies } \lambda_i = 0 \text{ for all } i \in I(x) \tag{2.23}$$

Then:

$$T_K(x) = \{v \in \mathbb{R}^n \,|Df_i(x)[v] \leq 0 \text{ for all } i \in I(x)\} \tag{2.24}$$

and

$$N_K(x) = \left\{\sum_{i\in I(x)} \lambda_i Df_i(x) \,|\lambda_i \in \mathbb{R}_{\geq 0}\right\} \tag{2.25}$$

▶ **Theorem 2.1.2.2:** *Let $f \in C_1(\mathbb{R}^n, \mathbb{R}^m)$ and let*

$$x \in K := \{x \in \mathbb{R}^n \,|f(x) = 0\} \tag{2.26}$$

Assume that $Df(x)$ has full rank. Then:

$$T_K(x) = \{v \in \mathbb{R}^n \,|Df_i(x)\,[v] = 0\} \tag{2.27}$$

and

$$N_K(x) = \left\{ \sum_{1 \le i \le m} \lambda_i Df_i(x) \,|\lambda_i \in \mathbb{R} \right\} \tag{2.28}$$

Consider the following special case of Problem 2.1.1

Problem 2.1.2 - (**Constrained optimization problem with equality and inequality constraint**)

Let $f \in C_1(\mathbb{R}^n, \mathbb{R}^m)$, $g_i \in C_1(\mathbb{R}^n, \mathbb{R})$ for $1 \le i \le k$ and let $K \subseteq \mathbb{R}^n$ be closed and convex. Define

$$M := \max\{\phi(x)\,|x \in K, f(x) = 0, g_i(x) \le 0\}, \tag{2.29}$$

where $\phi \in C_1(\mathbb{R}^n, \mathbb{R})$, whenever this maximum exists. In this case also find the maximizers:

$$C := \left\{x \in K \cap f^{-1}(\{0\}) \cap \cap_{1 \le i \le k} {g_i}^{-1}\left((-\infty, 0\ ]\right) |M = \phi(x)\right\} \tag{2.30}$$

A first order optimality condition for this problem can be formulated using Lagrange-multipliers (it is a special case of the multiplier rule on page 221 of Clarke [50] adapted to Problem 2.1.2):

Theorem 2.1.3 - (**Lagrange multipliers**)

Consider Problem 2.1.2 and assume that $x_* \in C$ then there exist $\eta \in \mathbb{R}$, $\gamma \in \mathbb{R}^k$ and $\lambda \in \mathbb{R}^m$ such that the following conditions hold true:

- *Nontriviality, i.e. at least one of the multipliers η, γ, λ is not equal to zero.*
- *Stationarity, i.e.*

$$D\left(\eta f - \sum_{1 \le i \le k} \gamma_i g_i - \sum_{1 \le j \le m} \lambda_j f_j\right)(x_*) \in N_K\left(x_*\right) \tag{2.31}$$

- *Positivity, i.e.*

$$\eta = 0 \text{ or } \eta = 1\ ;\ \gamma_i \ge 0 \text{ for all } 1 \le i \le k \tag{2.32}$$

- *Complementary slackness, i.e.*

$$\sum_{1 \le i \le k} \gamma_i g_i\left(x_*\right) = 0 \tag{2.33}$$

2.2 Set-valued analysis

For many problems in optimization theory, optimal control and related fields it is useful to consider set-valued maps and to equip the power-set of the target space with an appropriate topology. This is particularly useful if one wants to state and prove continuity of the solutions to an optimization problem with respect to the initial data (which is the constraint set and possibly some additional parameters).
Another potential application is a switch from ordinary differential equations to differential inclusions. The latter ones are an appropriate tool to define a mathematical convenient theory of ODEs with non-continuous right-hand side and projected differential equations. We are only interested in set-valued maps with values in the collection of closed sets of $\mathbb{R}^n$, so we restrict our definition to this case:

Definition 2.2.1 - (**Set valued maps**)

A set valued map from $\mathbb{R}^n$ *to* $\mathbb{R}^m$ *is a map*

$$f : \mathbb{R}^n \to 2^{\mathbb{R}^m}$$

We will always assume that $f(x)$ *is a closed subset of* $\mathbb{R}^m$*.*

The restriction to maps with closed values is reasonable, since most topologies defined on the power set do not distinguish a set from its closure. The collection of closed subsets of $\mathbb{R}^m$, denoted by $\mathrm{Cl}(\mathbb{R}^m)$, is also often called a hyperspace and topologies on the hyperspace are usually called hyperspace topologies. $\mathrm{Cl}(\mathbb{R}^m)$ equipped with the ordering relation $\subseteq$, is a complete lattice. For a collection of elements $A_i \in \mathrm{Cl}(\mathbb{R}^m)$ where i lies in an some index set I of arbitrary cardinality. The supremum of this collection is given by:

$$\sup_{i\in I} A_i = \overline{\cup_{i\in I} A_i} \tag{2.34}$$

and the infimum is given by

$$\inf_{i\in I} A_i = \cap_{i\in I} A_i \tag{2.35}$$

The lattice structure establishes a notion of monotonous convergence on $\mathrm{Cl}\left(\mathbb{R}^m\right)$ and every hypertopology is supposed to be compatible with this structure. The probably most common hypertopology in the literature on set-valued analysis is the Vietoris topology:

Definition 2.2.2 - (**The Vietoris topology**)

▶ **Definition 2.2.2.1:** *For a finite collection of open sets* $\mathbf{U} = (U_i)_{i\in I}$ *where* $U_i \subseteq \mathbb{R}^m$ *and* $|I| < \infty$ *define the hit-set of* (U_i)*:*

$$B_{\mathbf{U}} = \{A \in \mathrm{Cl}\left(\mathbb{R}^m\right) | A \cap U_i \neq \emptyset \text{ for all } i \in I\} \tag{2.36}$$

The sets $B_{\mathbf{U}}$ *where* $\mathbf{U}$ *is any finite collection of open sets that is stable with respect to intersection, therefore they form the base of a topology on the hyperspace, called the lower Vietoris topology. We will denote this topology by* $\tau_{\downarrow}\left(\mathbb{R}^m\right)$*. A canonical subbase of* $\tau_{\downarrow}\left(\mathbb{R}^m\right)$ *is given by the hitting sets of a single open set.*

▶ **Definition 2.2.2.2:** *A set valued map* $f : \mathbb{R}^n \to \mathrm{Cl}\left(\mathbb{R}^m\right)$ *is called lower semi-continuous if it is continuous with respect to the standard topology on* $\mathbb{R}^n$ *and the topology* $\tau_{\downarrow}\left(\mathbb{R}^m\right)$ *on* $\mathrm{Cl}\left(\mathbb{R}^m\right)$*.*

▶ **Definition 2.2.2.3:** *For a closed set $U \subseteq \mathbb{R}^m$ define the miss-set of U:*

$$B^U = \{A \in \mathrm{Cl}(\mathbb{R}^m) \,|A \cap U = \emptyset\} \tag{2.37}$$

The family of sets $B^{\mathbf{U}}$ where U is a closed set is stable with respect to finite intersections and therefore forms the base of a topology on the hyperspace. This topology is called upper Vietoris topology. We will denote it by by $\tau_\uparrow(\mathbb{R}^m)$.

▶ **Definition 2.2.2.4:** *A set valued map $f : \mathbb{R}^n \to \mathrm{Cl}(\mathbb{R}^m)$ is called upper semi-continuous if it is continuous with respect to the standard topology on $\mathbb{R}^n$ and the topology $\tau_\uparrow(\mathbb{R}^m)$ on $\mathrm{Cl}(\mathbb{R}^m)$.*

▶ **Definition 2.2.2.5:** *A set valued map $f : \mathbb{R}^n \to \mathrm{Cl}(\mathbb{R}^m)$ is called continuous if it is continuous with respect to both, $\tau_\uparrow(\mathbb{R}^m)$ and $\tau_\downarrow(\mathbb{R}^m)$. Equivalently it is continuous with respect to the Vietoris topology $\tau_{\uparrow\downarrow}(\mathbb{R}^m) := \tau(\tau_\uparrow(\mathbb{R}^m), \tau_\downarrow(\mathbb{R}^m))$.*

The notation is due to the following observation. Whenever $A \in U$ where $U \in \tau_\downarrow(\mathbb{R}^m)$ and $B \supseteq A$ then $B \in U$ and hence $\tau_\downarrow(\mathbb{R}^m)$. Equivalently if $A \in U$ where $U \in \tau_\uparrow(\mathbb{R}^m)$ and $B \subseteq A$ then $B \in U$. The Vietoris topology is a Hausdorff topology and a canonical base is given by the sets:

$$B^V_{\mathbf{U}} = \{A \in \mathrm{Cl}(\mathbb{R}^m) \,|A \cap V = \emptyset; A \cap U_i \neq \emptyset\} \tag{2.38}$$

where $V \subseteq \mathbb{R}^m$ is a closed set and $\mathbf{U} = (U_i)_{i \in I}$ is a finite family of open subsets $U_i \subseteq \mathbb{R}^m$. In this aspect $\tau_{\uparrow\downarrow}(\mathbb{R}^m)$ is a hit and miss topology [2].
An important theorem for optimization is the maximum theorem (compare Aubin [7] for example). In case of compact constraints it allows to conclude the continuity of the solution of a constrained optimization problem in the initial data:

Theorem 2.2.1 - (Maximum theorem)

Let $(\mathbf{X}, d_X)$ and $(\mathbf{Y}, d_Y)$ be metric spaces, let $\phi : \mathbf{X} \times \mathbf{Y} \to \mathbb{R}$ be a function and let $F : \mathbf{Y} \to 2^{\mathbf{X}}$ be set-valued map. If ϕ and F are lower semicontinuous, the marginal function

$$y \mapsto M_y := \sup\{\phi(x, y) \,|x \in F(y)\}$$

is lower semicontinuous. If ϕ and F are upper semicontinuous and F has compact values then the marginal function is upper semicontinuous. Moreover if ϕ is continuous and F is continuous with compact values, then the set-valued map

$$y \mapsto \{x \in F(y) \,|\phi(x, y) = M_y\}$$

has non-empty values and is upper semicontinuous.

Example 2.2.1 - (Application to invariant distributions of Markov chains)

As a consequence the set-valued map which maps a given stochastic matrix to its invariant distributions is upper semicontinuous. To see this fix a finite set $\mathbf{S}$ *and set*

$$\phi : \quad M_1\left(2^{\mathbf{S}}\right) \times \Lambda^{\mathbf{S}}_{\mathbf{S}} \to \mathbb{R} \tag{2.39}$$

$$(\mu, T) \mapsto -\|\mu T - \mu\|_1 \tag{2.40}$$

---

[2] There are many other hypertopologies that can be expressed as hit-and miss topologies for an appropriate collection of sets, another very well-known one is the Hausdorff topology (see for example Lucchetti and Pasquale [123])

where $\|\cdot\|_1$ *is the total variation norm. In Theorem 2.2.1 set* $\mathbf{X} := M_1\left(2^{\mathbf{S}}\right)$, $\mathbf{Y} := \Lambda_{\mathbf{S}}^{\mathbf{S}}$ *and* $F \equiv \mathbf{X}$*. By Theorem 6.0.2 any maximizer,* μ_**, of* $\phi\left(\cdot, K\right)$ *satisfies* $\mu_* = \mu_* K$*, i.e.* $\phi\left(\mu_*, K\right) = 0$ *and therefore Theorem 2.2.1 gives the desired result.*

2.3 Differential inclusions

An important application of set-valued maps arises in the context of differential inclusions. Let $F : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ be a set-valued map. A differential inclusion is an equation of the form

$$\frac{dx}{dt}(t) \in F\left(x(t)\right) \tag{2.41}$$

a local solution on some interval $I := (t_1, t_2) \subseteq \mathbb{R}$, is an absolutely continuous function $x : I \to \mathbb{R}^n$ that satisfies 2.41 almost everywhere. The standard existence proofs usually require F to have compact, convex values but there exist several extensions to slightly more general cases (see Anger, Aubin, and Cellina [6], Aubin [7], Aubin and Frankowska [8] for example). The proof proceeds by constructing appropriate approximate solutions and using the Arzela-Ascoli theorem to extract a converging subsequence. Since existence is an immediate consequence of the stochastic version to be considered in the next section, we will not write it down here. We only mention a very powerful approximation result, that we will also need for our proof later on. The theorem originates from Aubin and Frankowska [8], pp.67-68:

Lemma 2.3.1 - (**Approximation for differential inclusions**)

Let I be some interval and let $F : \mathbb{R}^n \to 2^{\mathbb{R}^n}$ *be upper semicontinuous with closed, convex images. Let* $x_m, y_m : I \to \mathbb{R}^n$ *be measurable functions and assume* $y \in L_1(d\nu_{\text{Leb}})$*. Assume that for every* $\epsilon > 0$ *and almost every* $t \in I$ *there exists* $M > 0$ *such that for all* $m > M$*:*

$$\text{dist}\left[\left(x_m(t), y_m(t)\right), \text{Graph}\left(F\right)\right] < \epsilon, \tag{2.42}$$

where $\text{dist}\left[\cdot,\cdot\right]$ *denotes the distance from a set and* $\text{Graph}\left(F\right)$ *is the graph of* F*. Assume further that*

- x_m *converges almost everywhere to some function* x
- y_m *converges weakly in* $L_1(d\nu_{\text{Leb}})$ *to some function* $y \in L_1(d\nu_{\text{Leb}})$*,*

then

$$y(t) \in F\left(x(t)\right) \text{ for almost every } t \in I \tag{2.43}$$

The learning algorithms in Chapter 4 will be based on discretized projected gradient schemes of the form

$$x_0 := x_s \in K; x_{n+1} := \hat{P}\left[x_n + h_n \nabla_g \phi\right] \tag{2.44}$$

where $(h_n)_{n\in\mathbb{N}}$ is the step size at step n (often called learning rate), ∇_g is the gradient with respect to some metric g on the constraint set $K \subseteq \mathbb{R}^n$, ϕ is the function to be maximized and $\hat{P}$ is a suitable projector that forces the solution to stay in the constraint set, K. Our only assumption on $\hat{P}$ is that it is a quasi-projector (compare Aubin [7] for example) that is Lipschitz continuous in a neighborhood of K.

Definition 2.3.1 - (**Quasi-projectors**)

► **Definition 2.3.1.1:** *A map* $\hat{P} : \mathbb{R}^n \to \mathbb{R}^n$ *will be called essentially Lipschitz continuous quasi-projector onto* K*, if* $\mathrm{Range}\left(\hat{P}\right) = K$*,* $\hat{P}^2 = \hat{P}$ *and if there exists some* $\epsilon > 0$ *and some constant* $L \in \mathbb{R}_{\geq 0}$ *such that:*

$$\left\|\hat{P}(x) - \hat{P}(y)\right\| \leq L \|x - y\| \tag{2.45}$$

whenever $\mathrm{dist}(x, K), \mathrm{dist}(y, K) < \epsilon$*.*

► **Definition 2.3.1.2:** *Let* $\hat{P}$ *be an essentially Lipschitz continuous quasi-projector onto some set* $K \subseteq \mathbb{R}^n$*. For every* $x \in K$ *define the preimage cone of* $\hat{P}$ *at* x*:*

$$C_{\hat{P}}(x) := \cap_{\epsilon>0} \overline{\mathrm{convcone}\left[\cup_{y \in B_\epsilon(x) \cap K}\left(\hat{P}^{-1}(y) - y\right)\right]} \tag{2.46}$$

where as before $\mathrm{convcone}\,[X]$ *stands for the convex cone generated by the set* $X \subseteq \mathbb{R}^n$

► **Definition 2.3.1.3:** *Let* $\hat{P}$ *be an essentially Lipschitz-continuous quasi-projector onto some set* $K \subseteq \mathbb{R}^n$*. A metric* g *on* K *will be called* $\hat{P}$*-compatible if for all* $x \in K$ *and every* $v \in C_{\hat{P}}(x)$*:*

$$g_x(v, \cdot) \in N_K(x) \tag{2.47}$$

The definition of the preimage cone and the compatibility condition between a metric and a quasi-projector are central definitions for the current chapter. Now we will illustrate the preimage cone, $C_{\hat{P}}(x)$, and motivate our interest in this cone.

Consider the iterative sequence specified by Eq. 2.44. Assume that $h_n > 0$, $h_n \to 0$ and $\sum_{n\in\mathbb{N}} h_n = \infty$. Then this sequence can be considered as a sequence of shrinking time steps. For every $t \in \mathbb{R}_{\geq 0}$ we set $h_{-1} := 0$ and set

$$\lfloor t \rfloor := \sup\left\{\sum_{k=-1}^{n} h_k \,\middle|\, \sum_{k=-1}^{n} h_k < t\right\} \text{ and } \lceil t \rceil := \inf\left\{\sum_{k=-1}^{n} h_k \,\middle|\, \sum_{k=-1}^{n} h_k \geq t\right\} \tag{2.48}$$

and

$$m(t) := \max\left\{n \,\middle|\, \sum_{k=-1}^{n-1} h_k \leq \lfloor t \rfloor\right\} \tag{2.49}$$

Then the iterated sequence 2.44 naturally embeds into $C\left([0, \infty), \mathbb{R}^n\right)$ via linear interpolation

$$x(t) := x_{\lfloor t \rfloor} + \frac{t - \lfloor t \rfloor}{\lceil t \rceil - \lfloor t \rfloor}\left(x_{m(t)+1} - x_{m(t)}\right) \tag{2.50}$$

It can be guessed (and will be made more precise later on) that for large times x "nearly satisfies" the differential inclusion

$$\frac{d}{dt}x(t) \in \nabla_g \phi(x) - C_{\hat{P}}(x) \text{ and } x(t) \in K \text{ a.e.} \tag{2.51}$$

A point $z \in K$ is a stationary point of this differential inclusion if and only if:

$$\nabla_g \phi(z) \in C_{\hat{P}}(z) \tag{2.52}$$

if g is compatible with $\hat{P}$ then this implies:

$$D\phi(z) = g_z\left(\nabla_g \phi(z), \cdot\right) \in N_K(x) \tag{2.53}$$

such that z satisfies the first order optimality condition for the optimization problem:
Find $\max\left\{\phi(x)\,|x \in K\right\}$ (compare Theorem 2.1.1).

In Chapter 4 we will present some learning algorithms, corresponding to the following two cases:

- K is the $\epsilon-$simplex and $\hat{P}$ is the best-approximation in Euclidean distance:

$$\hat{P}(x) = y \text{ iff } \|y-x\|_2 = \min\left\{\left\|y'-x\right\|_2 \middle| y' \in K\right\} \tag{2.54}$$

- K is the unit ball in $\mathbb{R}^{\mathbf{S}\times\mathbf{S}}$ in operator norm and $\hat{P}$ is the canonical retraction onto the unit ball:

$$\hat{P}(x) = \begin{cases} \frac{x}{\|x\|_{\mathrm{Op},p}} & \text{if } \|x\|_{\mathrm{Op},p} > 1 \\ x & \text{else} \end{cases} \tag{2.55}$$

we will also consider the case

- K is the closed R-ball in $(\mathbb{R}, \|\cdot\|_2)$, denoted by $\overline{B_1(0)}$ again and $\hat{P}$ is the best approximation.

since we think that this is a rather important choice in applications. As the former illustration (hopefully) motivates it is essential to find $\hat{P}$-compatible metrics for these cases.

Example 2.3.1 - (**A compatible metric for the best projection onto some convex set**)

As a first example let $K \subseteq \mathbb{R}^n$ be a compact, convex set and let $\hat{P}(x)$ be the best approximation of a point $x \in \mathbb{R}^n$ onto K, i.e.

$$\hat{P}(x) = y \text{ if and only if } y \in K \text{ and } \|x-y\|_2 = \min\left\{\|x-z\|_2 \,|\, z \in K\right\} \tag{2.56}$$

As a special property of the best approximation $\left\langle x - \hat{P}(x), y - \hat{P}(x)\right\rangle \leq 0$ for every $y \in K$. Hence $\left\langle x - \hat{P}(x), \cdot\right\rangle \in N_K(\hat{P}(x))$ for every x. Moreover since the set K is compact and convex, $C_{\hat{P}}(x) = \hat{P}^{-1}(x)$ for every $x \in K$ such that the Euclidean scalar product or any (possibly point dependent) multiple is compatible with $\hat{P}$.

Remark 2.3.1 - (**Comment on working with non-compatible metrics**)

Let K be a convex set and assume that K is the closure of its interior. If there are reasons to perform the gradient ascent algorithm with a metric, that is not compatible with the best approximation, $\hat{P}$, it is also possible to distort the desired metric slightly, such that it satisfies the compatibility requirement at the boundary. If the desired metric is g and $\delta > 0$ is some constant, a continuous $\hat{P}$-compatible metric is

$$g'_x := \frac{\max\left(\delta - \operatorname{dist}(x, \partial K), 0\right)}{\delta}\langle\cdot,\cdot\rangle + \frac{\min\left(\operatorname{dist}(x, \partial K), 1\right)}{\delta} g_x \tag{2.57}$$

Alternatively the discontinuous metric

$$g''_x := \begin{cases} g_x & \text{if } \operatorname{dist}(x, \partial K) \geq \delta \\ \langle\cdot,\cdot\rangle & \text{else} \end{cases} \tag{2.58}$$

can be used.

If the underlying norm on $\mathbb{R}^n$ is not Euclidean, a back-projection via the canonical retraction is always possible and easy to calculate. However the construction of a compatible metric is usually more involved. We give an example for $\mathbb{R}^{\mathbf{S}\times\mathbf{S}}$ for some finite set, $\mathbf{S}$, equipped with operator norm.

Example 2.3.2 - (**A compatible metric for the canonical retraction onto the unit ball of $\left(\mathbb{R}^{\mathbf{S}\times\mathbf{S}}, \|\cdot\|_{\mathrm{Op},p}\right)$**)

Consider the unit ball, $\overline{B_1(0)}$, in $\left(\mathbb{R}^{\mathbf{S}\times\mathbf{S}}, \|\cdot\|_{\mathrm{Op},p}\right)$ and let $\hat{P}$ be the canonical retractions:

$$X \mapsto \begin{cases} \frac{X}{\|X\|_{\mathrm{Op},p}} & \text{if } \|X\|_{\mathrm{Op},p} > 1 \\ X & \text{else} \end{cases} \tag{2.59}$$

For a point $A \in \mathbb{R}^{\mathbf{S}\times\mathbf{S}}$ with $\|A\|_{\mathrm{Op},p} \neq 0$. Fix $\lambda \in \mathbb{R}^{\mathbf{S},*}$ and $v \in \mathbb{R}^{\mathbf{S}}$ with $\|\lambda\|_{p,*} = \|v\|_p = 1$ and $\lambda(Av) = \|A\|_{\mathrm{Op},p}$. Define

$$g^{(\lambda,v)}{}_A(X,Y) := \lambda(Xv)\,\lambda(Yv) + \mathrm{Tr}\left[\left(X - \frac{\mathrm{Tr}\left[X^T A\right]}{\mathrm{Tr}\left(A^T A\right)}A\right)^T \left(Y - \frac{\mathrm{Tr}\left[Y^T A\right]}{\mathrm{Tr}\left(A^T A\right)}A\right)\right] \tag{2.60}$$

Then $g^{(\lambda,v)}{}_A$ is a scalar product. Unfortunately it is impossible to chose the vectors $\lambda \in \mathbb{R}^{\mathbf{S},*}$ and $v \in \mathbb{R}^{\mathbf{S}}$ continuously in A (discontinuities necessarily occur at points with several choices for λ and v). However this is no essential problem for the algorithms as will be shown later. The maximum theorem ensures that the set-valued map:

$$A \mapsto \left\{ g^{(\lambda,v)}{}_A \,\middle|\, \lambda \in \mathbb{R}^{\mathbf{S},*}; v \in \mathbb{R}^{\mathbf{S}}; \|\lambda\| = \|v\| = 1; \lambda(Av) = \|A\|_{\mathrm{Op}} \right\} \tag{2.61}$$

is upper semi-continuous on $\overline{B_1(0)} \setminus \{0\}$ and compatible with the retraction.

In the next section we will prove the essential theorems on stochastic approximation algorithms.

2.4 Stochastic approximation

The upcoming section about stochastic approximation is a partial extension of some classical results on stochastic approximation theory (compare Borkar [38], Bharath and Borkar [25], Kushner and Clark [109], Kushner and Yin [110]). All methods used in our proof can be found in these books and in the literature on set-valued analysis already presented. Our approach is very closed to the line of reasoning followed by Kushner, Clark and Yin. In contrast to their work we allow a broader class of projectors, consider more general classes of vector fields on the right-hand side and need less restrictions on the constraint set on the cost of stronger restrictions on the noise. Unfortunately the proof presented in the (in other aspects fantastic) book of Yin and Kushner is very incomplete and even contains some essential mistakes. However the original proofs by Kushner and Clark are solid and carefully written down. Still the authors prove the theorem by considering a special case (the constraint set being the closure of an open rectangle) and mention that it transfers to a more general situation and it is not easy to extract in a clear manner where the assumptions on the underlying constraint set (namely a specification by finite number of of differential constraints and the assumption that the constraint set is the closure of its interior points) really enter and how they can possibly be relaxed. We later want to analyze projected gradient ascent algorithms on a finite product of simplices over finite sets, which is a closed subset of $\mathbb{R}^{m\times n}$ and therefore violates the underlying assumptions. We are also interested in gradient ascents with respect to non-continuous metrics. This is why we insert an own proof of a theorem on stochastic approximation. .

Let $(\Omega, \mathcal{F}, P, \mathbb{F})$ be a filtered probability space, let M_n and β_n be $\mathbb{R}^n$-valued random variables on Ω (they can be thought of as noise variables). Let K (the constraint set) be a compact set, let $\hat{P}$ be an essentially Lipschitz continuous quasi-projector onto K, let $F : K \to 2^{\mathbb{R}^n}$ be a set-valued map (indicating the possible directions to go) and let

$g_n : \Omega \times K \to \mathbb{R}^n$ be a collection of random selections of F (i.e. $g_n(\omega, x) \in F(x)$ for almost every $\omega \in \Omega$ and every $x \in K$). Let $(h_n)_{n\in\mathbb{N}}$ be a sequence of positive random variables, the step size process. We are interested in the following iterative sequence:

$$x_0 := x_s \,;\, x_{n+1} := \hat{P}\left[x_n + h_n\left(g_n(x_n) + M_n + \beta_n\right)\right] \tag{2.62}$$

We impose the following assumptions on the underlying data:

Assumption 2.4.1 - (**Technical Restrictions on approximation sequence**)

We assume that:

► **Assumption 2.4.1.1:** *M_n is a martingale difference sequence (with respect to $\mathbb{F}$), satisfying*

$$\sup\left\{E\left[{M_n}^2\,|\mathbb{F}_{n-1}\right]|n \in \mathbb{N}_0\right\} < C \text{ a.s.}$$

for some constant $C \in \mathbb{R}_{\geq 0}$.

► **Assumption 2.4.1.2:** *The process $(\beta_n)_{n\in\mathbb{N}_0}$ is $\mathbb{F}$-adapted with*

$$\lim_{n\to\infty} |\beta_n| = 0 \text{ a.s.}$$

► **Assumption 2.4.1.3:** *The step-size process $(h_n)_{n\in\mathbb{N}_0}$ is previsible, i.e. g_n is $\mathbb{F}_{n-1}$-measurable, and satisfies*

$$h_n > 0 \sum_{n\in\mathbb{N}_0} h_n = \infty \,;\, h_n \to 0 \text{ a.s.}$$

and

$$\sum_{n\in\mathbb{N}_0} E\left[{h_n}^2\right] < \infty$$

► **Assumption 2.4.1.4:** *F is upper semicontinuous with compact, convex values.*

► **Assumption 2.4.1.5:** *The random selections $(g_n)_{n\in\mathbb{N}_0}$ are progressively previsible, by this we mean that h_n is $(\mathbb{F}_{n-1} \otimes \mathcal{B}_K)/\mathcal{B}_{\mathbb{R}^n}$ measurable.*

As before we set $h_{-1} := 0$ and define

$$\lfloor t \rfloor := \sup\left\{\sum_{k=-1}^{n} h_k \,\middle|\, \sum_{k=-1}^{n} h_k < t\right\} \text{ and } \lceil t \rceil := \inf\left\{\sum_{k=-1}^{n} h_k \,\middle|\, \sum_{k=-1}^{n} h_k \geq t\right\}, \tag{2.63}$$

$$m(t) := \max\left\{n \,\middle|\, \sum_{k=-1}^{n-1} h_k \leq \lfloor t \rfloor\right\} \tag{2.64}$$

and define the piecewise linear interpolation

$$x(t) := \begin{cases} x_s & \text{if } t \leq 0 \\ x_{\lfloor t \rfloor} + \frac{t-\lfloor t \rfloor}{\lceil t \rceil - \lfloor t \rfloor}\left(x_{m(t)+1} - x_{m(t)}\right) & \text{else} \end{cases} \tag{2.65}$$

Then we have the following statement:

Theorem 2.4.1 - (**Limiting behavior of the iterative stochastic sequence**)

Consider the stochastic sequence $(x_n)_{n\in\mathbb{N}}$ *defined by Eq 2.62 and assume that Assumption 2.4.1 holds. Let* x *denote the piecewise linear interpolation as defined by Eq. 2.65. For almost all* $\omega \in \Omega$ *the family of functions*

$$\phi_s(t) := x(s+t)\ ;\ s \in \mathbb{R} \tag{2.66}$$

is sequentially precompact in $C\left(\mathbb{R}, \overline{\mathrm{co}(K)}\right)$ *equipped with the (locally convex) topology of uniform convergence on compact subsets.*

Moreover any limit point, ϕ_**, of some sequence* $(\phi_{s_n})_{n\in\mathbb{N}}$*, where* $(s_n)_{n\in\mathbb{N}}$ *is a sequence of real numbers with* $s_n \to \infty$ *is absolutely continuous and there exists some (sequence independent!) constant* R *such that* ϕ_* *satisfies*

$$\frac{d}{dt}\phi_*(t) \in \big(F(\phi_*(t)) - C_{\hat{P}}(\phi_*(t))\big) \cap B_R(0) \text{ and } \phi_*(t) \in K \text{ for a.e. } t \in \mathbb{R} \tag{2.67}$$

where $C_{\hat{P}}(x)$ *is the convex cone defined in Definition 2.3.1.*

For the proof we need three fundamental lemmas, the first one is the well-known Arzela-Ascoli theorem, a proof of which can be found in any good introductory text book on functional analysis (like Rudin [160], Triebel [184], Werner [192], Aliprantis and Border [1]):

Lemma 2.4.1 - (**Arzela-Ascoli theorem**)

Let $\mathbf{X}$ *be a compact metric space and let* $\Lambda \subseteq C(\mathbf{X}, \mathbb{R}^n)$*. Then* Λ *is sequentially precompact in the Banach space* $(C(\mathbf{X}, \mathbb{R}^n), \|\cdot\|_\infty)$ *if and only if* $\{\lambda(x)\,|\lambda \in \Lambda\}$ *is precompact for every* $x \in \mathbf{X}$ *and* Λ *is equicontinuous, i.e. for every* $\epsilon > 0$ *there exists* $\delta > 0$ *such that* $d(x,y) < \delta$ *implies* $|\lambda(x) - \lambda(y)| < \epsilon$ *for every* $\lambda \in \Lambda$.

And we need a separation theorem that is a consequence of the Hahn-Banach theorem (see for example Rudin [160], Triebel [184], Werner [192], Aliprantis and Border [1]):

Lemma 2.4.2 - (**Separation theorem**)

Let $K \subseteq \mathbb{R}^n$ *be a closed convex cone with negative polar cone*

$$K^* := \{\lambda \in \mathbb{R}^{n,*}\,|\lambda(x) \leq 0 \text{ for every } x \in K\}.$$

Then

$$K = \{x \in \mathbb{R}^n\,|\lambda(x) \leq 0 \text{ for all } \lambda \in K^*\} \tag{2.68}$$

Moreover $\mathrm{dist}(y, K) \leq \epsilon$ *if and only if*

$$\lambda(y) \leq \epsilon \text{ for every } \lambda \in K^* \cap {B^*}_1(0) \tag{2.69}$$

The next lemma, the Banach-Alaoglu theorem, is a consequence of Tychonoff's theorem and is also a classical result in linear functional analysis:

Lemma 2.4.3 - (**Banach-Alaoglu theorem**)

Let $(\mathbf{B}, \|\cdot\|)$ *be a Banach space then the closed unit* $\overline{B_0(1)} \subset \mathbf{B}^*$ *is compact in the* weak* *topology, i.e. the coarsest topology that renders the maps* $\lambda \mapsto \lambda(v)$ *continuous, for every* $v \in \mathbf{B}$.

Proof of Theorem 2.4.1. We start with a proof of sequential precompactness. Since $\mathrm{Range}\,(x) = \overline{\mathrm{co}(K)}$ is compact we only have to show that x is almost surely uniformly continous. Then the Arzela-Ascoli theorem and a diagonal argument will give the desired result. More explicit - given that x is uniformly continuous- consider an arbitrary sequence $(f_k)_{\in \mathbb{N}}$ with $f_n \in \{\phi_s \,| s \in \mathbb{R}\}$. Then by the Arzela-Ascoli theorem there exist subsequences $y_{L,k}$ such that $(y_{L+1,k})_{k \in \mathbb{N}}$ is a subsequence of $(y_{L,k})_{k \in \mathbb{N}}$ and $y_{L,k}$ converges uniformly on $[-L, L]$. Then the diagonal sequence $y_{k,k}$ converges to some limit uniformly on compact subsets.

Since continuous functions on compact sets are always uniformly continuous, it is enough to show that for almost every $\omega \in \Omega$ and for every $\epsilon > 0$ there exists some $\delta > 0$ and some $T > 0$ such that:

$$|x(s) - x(t)| < \epsilon \text{ whenever } |s - t| < \delta \text{ and } s, t > T \tag{2.70}$$

By our assumptions x_{n+1} is $\mathbb{F}_n$ measurable and by our definition of essentially Lipschitz continuous quasi-projectors there exists some $\epsilon > 0$ such that

$$\left\|\hat{P}(x) - \hat{P}(y)\right\| < L \|x - y\| \tag{2.71}$$

whenever $\operatorname{dist}(x, K), \operatorname{dist}(y, K) < \kappa$ for some $\kappa > 0$. Since F is upper semicontinuous with compact values, $F(x_n)$ is uniformly bounded. This can be seen easily from the following argument: By upper semicontinuity of F it is possible to fix some some δ_y for every $y \in K$ such that $|x - y| < \delta_y$ implies

$$F(y) \subseteq U(F(x), 1) := \{v \in \mathbb{R}^n \,|\operatorname{dist}(v, F(x)) < 1\}.$$

By compactness of K there exists a finite set $K' \subseteq K$ such that the balls $B_{\delta_y}(y)$ with $y \in K'$ cover K. Then $F(K)$ is contained in the compact set $\cup_{y \in K'} \overline{U(F(y), 1)}$. So there exists some $R' \in \mathbb{R}_{\geq 0}$ such that

$$\sup \{v \,| v \in F(x); x \in K\} < R' \tag{2.72}$$

Note that M_n is square integrable, such that the martingale

$$B_N := \sum_{k=0}^{N} h_n M_n$$

converges almost surely and in L_2 by the martingale convergence theorem. Therefore:

$$\lim_{n \to \infty} h_n (\|g(x_n)\| + \|\beta_n\| + \|M_n\|) = 0 \text{ a.s.} \tag{2.73}$$

Define the set

$$A_{\kappa,n} := \left\{h_n (\|g(x_n)\| + \|M_n\|) \leq \frac{\kappa}{2}\right\} \tag{2.74}$$

The validity of Eq. 2.73 implies that almost surely $\omega \in A_{\kappa,n}$ for all but finitely many n, i.e.

$$P\left[\cup_{n \in \mathbb{N}} \cap_{k > n} A_{\kappa k}\right] = 1 \tag{2.75}$$

The defining sequence Eq. 2.62 can be rewritten in the following way:

$$x_{n+1} = \mathbb{1}_{A_{\kappa,n}} x_n + Y_n + M'_n + \beta'_n \tag{2.76}$$

where

$$Y_n := E\left[\left(\hat{P}\left[x_n + h_n\left(g(x_n) + M_n\right)\right] - x_n\right) \mathbb{1}_{A_{\kappa,n}} \,|\mathbb{F}_{n-1}\right], \tag{2.77}$$

$$\beta'_n := \hat{P}\left[x_n + h_n\left(g(x_n) + M_n + \beta_n\right)\right] - \mathbb{1}_{A_{\kappa,n}} \hat{P}\left[x_n + h_n\left(g(x_n) + M_n\right)\right] \tag{2.78}$$

and

$$M'_n := \left(\hat{P}\left[x_n + h_n\left(g(x_n) + M_n\right)\right] - x_n\right) \mathbb{1}_{A_{\kappa,n}} - Y_n \tag{2.79}$$

Since we assumed $\hat{P}$ to be essentially Lipschitz continuous Eq. 2.73 implies:

$$\left|{\beta'}_n\right| \leq L h_n \left|\beta_n\right| \tag{2.80}$$

almost surely for sufficiently large n.

By assumption

$$E\left[\left(M_n\right)^2 |\mathcal{F}_{n-1}\right] \leq C \tag{2.81}$$

For some constant C such that Eq 2.72 implies

$$\left|Y_n(\omega)\right| \leq h_n L\left(R' + \sqrt{C}\right) \tag{2.82}$$

almost surely. The sequence M'_n is a martingale difference sequence and $\hat{P}(x_n) = x_n$ (since $x_n \in K$) and therefore the Minkowski inequality gives:

$$E\left[\left(\hat{P}\left[x_n + h_n\left(\phi(x_n) + M_n\right)\right] - x_n\right)^2 \mathbb{1}_{A_{\kappa,n}}\right] \leq L^2 E\left[{h_n}^2\right]\left(R' + \sqrt{C}\right)^2$$

Therefore our assumptions on $(h_n)_{n\in\mathbb{N}}$ and the martingale convergence theorem ensure that

$$\tilde{M}_n := \sum_{0\leq k\leq n} M'_k \text{ converges a.s. and in } L_2(dP) \tag{2.83}$$

Now set $C_2 := R' + \sqrt{C}$, fix some version of the conditional expectation whenever it occurs and some null-set $\mathcal{N} \in \mathcal{F}$ such that Eq. 2.80, Eq. 2.83, Eq. 2.73 and Eq. 2.82 are satisfied for every $\omega \in \mathcal{F} \setminus \mathcal{N}$. For every $\omega \in \Omega \setminus \mathcal{N}$ there exists some $N_0 \in \mathbb{N}$ such that for every $n, m \in \mathbb{N}_0$ with $n, m \geq N_0$:

$$\begin{aligned} & \left|x_{n+m} - x_n\right| \\ \leq\ & C_2\left(\sum_{k=n}^{n+m-1} h_k(\omega)\right) + L\sum_{k=n}^{n+m-1} h_k\left|\beta_k(\omega)\right| + \left|\tilde{M}_{n+m-1}(\omega) - \tilde{M}_{n-1}(\omega)\right| \end{aligned} \tag{2.84}$$

For given $\epsilon > 0$ fix $N \geq N_0$ such that for every $n \geq N$:

$$h_n < \frac{\epsilon}{8C_2}; \left|\beta_n(\omega)\right| < \frac{C_2}{2L}; \left|M_n(\omega) - \lim_{k\to\infty} M_k(\omega)\right| < \frac{\epsilon}{8} \tag{2.85}$$

Then for every $s \leq t \in \mathbb{R}_{\geq 0}$ with $m(s) \geq N$ and $|t - s| < \frac{\epsilon}{4C_2}$

$$\begin{aligned} & \left|x(t) - x(s)\right| \qquad (2.86) \\ \leq\ & \max\left\{\left|x(\lfloor t\rfloor) - x(\lfloor s\rfloor)\right|, \left|x(\lceil t\rceil) - x(\lfloor s\rfloor)\right|, \left|x(\lfloor t\rfloor) - x(\lceil s\rceil)\right|, \left|x(\lceil t\rceil) - x(\lceil s\rceil)\right|\right\} \\ \leq\ & \left(\lceil t\rceil - \lfloor s\rfloor\right)\left(C_2 + \frac{C_2}{2}\right) + 2\cdot\frac{\epsilon}{8} \\ \leq\ & \left(t - s + 2\cdot\frac{\epsilon}{8C_2}\right)\frac{3C_2}{2} + \frac{\epsilon}{4} \leq \epsilon \end{aligned} \tag{2.87}$$

Therefore the family $\{\phi_s\}_{s\in\mathbb{R}}$ is equicontinuous and hence precompact by the Arzela-Ascoli theorem.

To characterize the limit of a convergent subsequence define the piece-wise constant function, Y, via:

$$Y(t) := \begin{cases} 0 & \text{for } t \leq 0 \\ \frac{Y_{m(t)}}{h_{m(t)}} & \text{else} \end{cases} \tag{2.88}$$

For $\omega \notin \cap_{k\geq 0} A_{\kappa,n}$ the identity Eq 2.76 can be rewritten in the following form: For all $t \in \mathbb{R}$

$$\begin{aligned} & x(t) - x_s - \int_0^t Y(t')dt' \qquad (2.89) \\ = & \sum_{k=0}^{m(t)-1} \left(M'_k + \beta'_k\right) + \frac{t - \lfloor t \rfloor}{\lceil t \rceil - \lfloor t \rfloor} \left(M'_{m(t)+1} - M'_{m(t)} + \beta'_{m(t)+1} - \beta'_{m(t)}\right) \end{aligned}$$

Now let $(t_k)_{k\in\mathbb{N}}$ be a sequence of real number with $\lim_{k\to\infty} t_k = \infty$ such that ϕ_{t_k} converges to some limit ϕ_* uniformly on compact intervals. Defining $\psi_s := Y(t+s)$ the identities Eq. 2.80, Eq. 2.83 and Eq. 2.75 imply that for every $s, t \in \mathbb{R}$ and almost all $\omega \in \Omega$:

$$\phi_*(t) - \phi_*(s) = \lim_{k\to\infty} \int_s^t \psi_{t_k}(t')dt'$$

Now by 2.82 the function $Y\big|_{[-T,T]}$ is bounded by $R := L\left(R' + \sqrt{C}\right)$ therefore for every $T > 0$ the family $\psi_s\big|_{[-T,T]}$ is precompact in $L_\infty\left([-T,T], \mathbb{R}^n; d\nu_{\text{Leb}}\right)$ equipped with the weak* topology as a consequence of the Banach-Alaoglu theorem [3]. Hence it is possible to extract a weak$^*-$ convergent subsequence $\left(\psi_{t'_k}\right)$ of ψ_{t_k} with some limit ψ_* that necessarily satisfies $\|\psi_*\|_\infty \leq R$ by compactness of the norm closure of the unit ball in weak* topology. Since the Lebesgue measure restricted to a compact interval is finite we have $L_\infty\left([-T,T], \mathbb{R}^n; d\nu_{\text{Leb}}\right) \subseteq L_1\left([-T,T], \mathbb{R}^n; d\nu_{\text{Leb}}\right)$ and the embedding is actually weak*-weak continuous [4]. Therefore the sequence $\left(\psi_{t_k}\right)_{k\in\mathbb{N}}$ converges weakly in $L_1\left([-T,T], \mathbb{R}^n; d\nu_{\text{Leb}}\right)$ and by a diagonal argument again it is possible to extract a subsequence that converges weakly to some integrable function Y_* (without loss of generality we assume convergence along (t_k) already). Necessarily:

$$\phi_*(t) - \phi_*(s) = \int_s^t Y_*(t')dt'$$

such that ϕ_* is absolutely continuous with derivative Y_*.
In order to conclude the desired result from Lemma 2.3.1 it remains to show that

$$\text{dist}\left[\left(\phi_{t_k}, \psi_{t_k}\right), \text{Graph}(F - C_{\hat{P}})\right]$$

converges to zero almost everywhere. To show this fix some linear functional $\lambda \in \mathbb{R}^{n,*}$ with $\|\lambda\|_* = 1$ and some constant $N \in \mathbb{N}$, set $A_{N,n} := \{\|M_n + g(x_n)\| \leq N\}$. Then since M_n is a martingale difference sequence:

$$\lambda\left(\frac{Y_n}{h_n} - g\left(x_n\right)\right) = I_1 + I_2 + I_3 \qquad (2.90)$$

where

$$I_1 := E\left[\lambda\left(\frac{\hat{P}\left[x_n + h_n\left(g(x_n) + M_n\right)\right] - \left(x_n + h_n\left(g(x_n) + M_n\right)\right)}{h_n}\right)\mathbb{1}_{A_{\kappa,n}\cap A_{N,n}} |\mathbb{F}_{n-1}\right],$$

$$I_2 := E\left[\lambda\left(\frac{\hat{P}\left[x_n + h_n\left(g(x_n) + M_n\right)\right] - \left(x_n + h_n\left(g(x_n) + M_n\right)\right)}{h_n}\right)\mathbb{1}_{A_{\kappa,n}\cap A_{N,n}{}^c} |\mathbb{F}_{n-1}\right],$$

and

$$I_3 := -E\left[\lambda\left(M_n + g\left(x_n\right)\right)\mathbb{1}_{A_{\kappa,n}{}^c} |\mathbb{F}_{n-1}\right] \qquad (2.91)$$

---

[3] Being a consequence of Tychonoff's theorem the Banach-Alaoglu theorem does not imply sequential compactness in general. Since $L_1(\mathbb{R}, \mathbb{R}^n, d\nu_{\text{Leb}})$ is separable the sequential compactness follows nevertheless.

[4] This is clear since in this case every linear functional on L_1 is necessarily also continuous on L_∞

For a bound on the third integral note that $h_n \frac{\|M_n\|+\|g(x_n)\|}{\kappa} > 1$ on ${A_{\kappa,n}}^c$ such that:

$$|I_3| \leq \frac{h_n}{\kappa} E\left[\left(\|M_n\| + \|g\left(x_n\right)\|\right)^2 |\mathbb{F}_{n-1} \right] \leq \frac{h_n \left(R' + \sqrt{C}\right)^2}{\kappa} \tag{2.92}$$

For a bound on the second integral we use a similar argument. On $A_{\kappa,n}$ the Lipschitz estimate holds true and on ${A_{N,n}}^c$ we have $\frac{\|M_n+g(x_n)\|}{N} > 1$ such that:

$$|I_2| \leq \frac{L+1}{N} \left(R' + \sqrt{C}\right)^2 \tag{2.93}$$

The first summand, I_1, is actually the most interesting one. On $A_{\kappa,n} \cap A_{N,n}$ we have $\|x_{n+1} - x_n\| \leq L h_n N$. Therefore the norm of the integrand is bounded by $(L+1)\,N$. Moreover

$$\left(\frac{\hat{P}\left[x_n + h_n\left(g(x_n) + M_n\right)\right] - \left(x_n + h_n\left(g(x_n) + M_n\right)\right)}{h_n} \right)$$

lies in $C_{\hat{P}}\left(x_{n+1}\right)$ by definition such that:

$$I_1 \leq \sup\left\{ \lambda(v) \left| v \in C_{\hat{P}}(x') \cap B_{(L+1)N}\left(0\right); \left\|x' - x_n\right\| \leq L h_n N \right. \right\} \tag{2.94}$$

This result holds true almost surely for a given functional λ and a given $N \in \mathbb{N}$. Since the unit ball in $\mathbb{R}^{n,*}$ is separable there exists some null-set $\mathcal{N}'$ such that whenever $\omega \in \Omega \setminus \mathcal{N}'$ then

$$\begin{aligned} & \lambda\left(\frac{Y_n}{h_n} - g\left(x_n\right)\right) \\ \leq\; & \sup\left\{ \lambda(v) \left| v \in C_{\hat{P}}(x') \cap B_{(L+1)N}; \left\|x' - x_n\right\| \leq L h_n N \right. \right\} + \\ & + \left(\frac{L+1}{N} + \frac{h_n}{\kappa}\right)\left(\sqrt{C} + R'\right)^2 \end{aligned}$$

holds true for all linear functionals of unit length and for all $N \in \mathbb{N}$.
Now let $\omega \in \Omega \setminus \left(\mathcal{N} \cup \mathcal{N}'\right)$ and consider the sequence ϕ_{t_k} converging to ϕ_* uniformly on compact subsets and the sequence ψ_{t_k} converging to Y_* weakly in L_1. Fix any $\epsilon > 0$ and any t, that does not lie in the Lebesgue null set where one of the functions ψ_{n_k} is not continuous. First fix $N > 0$ such that $\left(\frac{L+1}{N} + \frac{h_n}{\kappa}\right)\left(\sqrt{C} + R\right)^2 < \frac{\epsilon}{3}$ whenever $n > N$. Then using the upper semi-continuity of $C_{\hat{P}} \cap \overline{B_1(0)}$ fix δ such that $\operatorname{dist}\left(v, C_{\hat{P}} \cap \overline{B_1(0)}(\phi_*(t))\right) \leq \frac{\epsilon}{3((L+1)N)}$ for every $v \in \overline{B_1(0)} \cap \cup_{x \in B_\delta(\phi_*(t))} C_{\hat{P}}(x)$. Then fix N_2 sufficiently large such that

- For all $n \geq N_2$ we have $h_n L N \leq \frac{\delta}{2}$
- Whenever $m(t_k) \geq N_2$ then $\|\phi_{t_k}(t) - \phi_*(t)\| < \frac{\epsilon}{3}$
- Whenever $m(t_k) \geq N_2$ then $\|x(\lfloor t + t_k \rfloor) - \phi_*(t)\|, \|x(\lceil t + t_k \rceil) - \phi_*(t)\| < \frac{\delta}{2}$ (this can be reached since the convergence of ϕ_{t_k} is uniform on compacts. The construction involves the usual ϵ-third argument - in this case a $\delta/6$ argument to be precise.)

Then whenever $m(t_k{+}t) \geq \max\left(N_1, N_2\right)$ we have for any unit functional $\lambda \in C_{\hat{P}}\left(\psi_*(t)\right)^*$, the negative polar cone of $C_{\hat{P}}\left(\psi_*(t)\right)$, that

$$\lambda\left(\psi_{t_k}(t) - g\left(x_{t_k}(t)\right)\right) \leq \frac{2\epsilon}{3} \tag{2.95}$$

such that, in virtue of Lemma 2.4.2, $\operatorname{dist}\left[\left(\phi_{t_k}(t), \psi_{t_k}(t)\right), \operatorname{Graph}\left(F - C_{\hat{P}}\right)\right] < \epsilon$ for sufficiently large k. This proves the second statement of the theorem. ■

We would like to give a typical application of this theorem.

Example 2.4.1 - (**Application of the stochastic approximation theorem**)

Let (Y_i) *be a sequence of random variables and let* λ *be a* $\mathbb{R}^m$*-valued parameter of the distribution of this sequence. Assume that there exists a sequence of consistent estimators* $\left(\hat{\lambda}_n\right)_{n\in\mathbb{N}}$*, i.e. we assume that* $\hat{\lambda}_n$ *is* $\sigma\left(\{Y_i\}_{1\le i\le n}\right)$ *measurable and that* $\hat{\lambda}_n \to \lambda$ *a.s.* *Let* $(Z_n)_{n\in\mathbb{N}}$ *be a sequence of* $\mathbb{R}^p$*-valued IID random variable, independent of* $(Y_n)_{n\in\mathbb{N}}$*. We assume that the distribution of* Z_1 *has finite p-th moment for some* $p \ge 1$*, i.e.* $E\left[\|Z_1\|^p\right] < \infty$*. Let* $K \subseteq \mathbb{R}^n$ *be a compact set, let* $\hat{P}$ *be a quasi-projector onto this set, let* $\frac{1}{2} < \alpha \le 1$ *be some decay coefficient and let* $g : K \times \mathbb{R}^m \times \mathbb{R}^p \to \mathbb{R}^n$ *we assume that*

- *g is continuous and the continuity in the second coordinate is uniformly for fixed first and second coordinates ,uniformly in the first and third coordinate. By this we mean that for every* $\epsilon > 0$ *there exists* $\delta > 0$ *such that for all* $x \in K$ *and* $z \in Z_n$*:*

$$\left|g(x,\lambda,z) - g(x,\lambda',z)\right| < \epsilon \text{ whenever } \left\|\lambda - \lambda'\right\| < \delta \tag{2.96}$$

- *g satisfies a certain growth condition in the third coordinate. Namely we assume that there exists a continuous function* $h : K \times \mathbb{R}^m \to \mathbb{R}_{\ge 0}$ *such that*

$$|g(x,\lambda,z)| \le h(x,\lambda)\left(1 + \|z\|^{\frac{p}{2}}\right)$$

Consider the stochastic approximation algorithm

$$x_{n+1} := \hat{P}\left[x_n + \frac{1}{n^\alpha} g\left(x_n, \hat{\lambda}_n, Z_n\right)\right] \tag{2.97}$$

We rewrite it in the following form:

$$x_{n+1} = \hat{P}\left[x_n + \frac{1}{n^\alpha}\left(\tilde{g}\left(x_n,\lambda\right) + \beta\left(x_n, \hat{\lambda}_n, Z_n\right) + M_n\right)\right] \tag{2.98}$$

where

$$\tilde{g}(x,\lambda) := \int P_{Z_1}(dz) g\left(x,\lambda,z\right), \tag{2.99}$$

$$\beta(x,\lambda',z) := g\left(x,\lambda',z\right) - g\left(x,\lambda,z\right) \tag{2.100}$$

and

$$M_n := g\left(x_n,\lambda,Z_n\right) - \tilde{g}\left(x_n,\lambda\right) \tag{2.101}$$

Since $(Z_n)_{n\in\mathbb{N}}$ *is independent of* $(Y_n)_{n\in\mathbb{N}}$ *we have*

$$\tilde{g}\left(x_n,\lambda\right) = E\left[g\left(x_n,\lambda,Z_n\right)|\mathbb{F}_{n-1}\right],$$

where

$$\mathbb{F}_{\mathrm{n}} := \sigma\left(\{Z_i, Y_i\}_{1\le i<n}\right)$$

By the growth condition on g and the compactness of K*,* $\tilde{g}$ *is dominated by the integrable random variable* $R\left(1 + \|Z_1\|^{\frac{p}{2}}\right)$ *where* $R \in \mathbb{R}_{\ge 0}$ *is an appropriate constant. Therefore by dominated convergence* $\tilde{g}$ *is continuous. By the same argument, the martingale difference,* M_n*, is square integrable with*

$$E\left[{M_n}^2 \left|\mathbb{F}_{n-1}\right.\right] \le R^2\left(1 + \sqrt{E\left[|Z_1|^p\right]}\right)^2 \tag{2.102}$$

and by the uniform continuity assumptions on g:

$$\beta\left(x_n, \hat{\lambda}_n, Z_n\right) \to 0 \text{ a.s.} \tag{2.103}$$

Hence Theorem 2.4.1 ensures that the piecewise linear interpolation of $x(t)$ almost surely "asymptotically satisfies" the differential inclusion problem:

$$\frac{d}{dt}x(t) \in \tilde{g}\left(x(t),\lambda\right) - C_{\hat{P}}\left(x(t)\right)\ ;\ x(t) \in K \text{ a.e.} \tag{2.104}$$

A problem of the form

$$\frac{d}{dt}x(t) \in F(x)\ ;\ x(t) \in K \text{ a.e.} \tag{2.105}$$

is known as a viability problem (compare Aubin [7]). Any solution x necessarily satisfies $x(t) \in T_K(x(t))$ for a.e. t (this is the first part of the well-known viability theorem for differential inclusion, a proof of which can be found in Aubin [7] for example. The other direction is the existence of a solution if $F(x)$ is upper semicontinuous with compact, convex values, if K is locally compact and if $F(x) \cap T_K(x) \neq \emptyset$). Before proving some consequences of Theorem 2.4.1 we need the following definition, common in the theory of dynamical systems (see for example Jost [97], Perko [141] and Marx and Vogt [125]) and a definition of "well-behavior" for the critical values of a function:

Definition 2.4.1 - (**Attractors and asymptotical stability**)

► **Definition 2.4.1.1:** *Let*

$$x(t) \in F(x(t))\ ;\ x(t) \in K \text{ a.e.} \tag{2.106}$$

be a differential inclusion. A subset $A \subseteq K$ will be called positively Lyapunov stable, if for every $\epsilon > 0$ there exists some $\delta > 0$ such that for any solution x of Eq. 2.106:

$$\text{dist}\left(x(0), A\right) < \delta \text{ implies } \text{dist}\left(x(t), A\right) < \epsilon \text{ for a.e. } t \geq 0 \tag{2.107}$$

► **Definition 2.4.1.2:** *Consider the problem Eq. 2.106. A set $A \subseteq K$ will be called attractor if it is positively Lyapunov stable and if there exists some neighborhood U of A such that every solution, x, to Eq. 2.106 with $x(0) \in U$ converges to A, i.e.*

$$\lim_{t\to\infty} \text{dist}\left(x(t), A\right) = 0 \tag{2.108}$$

► **Definition 2.4.1.3:** *A continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ will be called Lyapunov function for problem Eq. 2.106. If for any $x \in K$ we have:*

$$\sup\left\{DV(x)\left[w\right] | w \in F(x) \cap T_K(x)\right\} \leq 0 \tag{2.109}$$

► **Definition 2.4.1.4:** *A continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ will be said to have a negligible set of generalized critical values for problem Eq. 2.106, if the set*

$$S_L := \left\{V(x) \in \mathbb{R} \,|\, x \in K \text{ and } \sup\left\{DV(x)\left[w\right] | w \in F(x) \cap T_K(x)\right\} = 0\right\}$$

has isolated accumulation points only (by accumulation point we mean an element $y \in \mathbb{R}$ such that there exists a sequence $(y_n)_{n\in\mathbb{N}} \in S_L$ with $y_n \neq y$ and $y_n \to y$.

Remark 2.4.1 - (**Comment on Definition 2.4.1**)

First of all note that S_L *is closed. To see this define*

$$\phi(x) := \max\{-|DV(x)[v]|\,|v \in F(x) \cap T_K(x) \cap B_1(x)\} \tag{2.110}$$

By the maximum theorem ϕ *is upper semicontinuous and therefore*

$$S_L := \{x \in K\,|V(x) \geq 0\}$$

is closed. The assumption that V *has a negligible set of critical values is essential for our proof technique for the upcoming theorem. It is clearly satisfied if* V *is convex or concave. More generally it holds true whenever the set of generalized stationary points is discrete, which again holds true whenever* V *is twice continuously differentiable and if the second derivative is strictly positive or strictly negative at all stationary points. It is also trivially true if the set of generalized stationary points can be splited into finitely many connected components, on which* V *is constant. However it is a maybe surprising but well-known insight that the set of critical values of* V *can be much more complicated - even for differentiable functions. In 1935 Whitney presented a function* $f : \mathbb{R}^n \to \mathbb{R}$ *that is* $n-1$ *times continuously differentiable but is non-constant on a connected component of critical points (compare Whitney [193]). Whenever* f *is* n *times continuously differentiable this behavior is excluded by Sard's theorem, that states that in this case the set of critical values has Lebesgue measure zero (compare Spivak [174, 173] and Lang [112] for proofs of partial results and the original paper Sard [165] for a complete proof). A latter version even shows, that the Hausdorff-dimension is zero in this case (compare Sard [164]). However this is still much weaker then having a negligible set of critical values as we defined it. As a current result of Bolte, Danilidis, Lewis and Shiota (Bolte et al. [33], Bolte, Daniilidis, and Lewis [32]) our assumption is also satisfied by analytical functions. In this case the set of critical values is locally finite.*

A Lyapunov function is always non-increasing along a given trajectory. To see this let V be a Lyapunov function for problem Eq. 2.106. Then by the viability theorem:

$$\frac{dV \circ x}{dt}(t) = DV\left(x(t)\right)\left[\frac{dx}{dt}(t)\right] \in \{DV\left(x(t)\right)[w]\,|w \in F(x) \cap T_K(x)\} \text{ a.e.}$$

define $\phi_{\text{max}}(t) := \max\{DV\left(x(t)\right)[w]\,|w \in F(x(t))\}$. Then ϕ_{max} is bounded from above by zero and upper semicontinuous whenever F is upper semicontinuous with compact values (by the maximum theorem). In any case for $t \geq s$:

$$V \circ x(t) - V \circ x(s) \leq \int_s^t \phi_{\text{max}}(s')ds' \leq 0$$

We need the following theorem (the proof of the first part is equal to the proof in Kushner and Clark [109] pp. 39-43, adapted to our theorem, also see Bertsekas and Tsitsiklis [24])

Theorem 2.4.2 - (**Asymptotics of stochastic approximation sequence**)

Consider the stochastic sequence $(x_n)_{n\in\mathbb{N}}$ *defined by Eq 2.62. Assume that Assumption 2.4.1 holds. Let* A *be an attractor for the mean differential inclusion*

$$\frac{dx}{dt}(t) \in \big(F(x(t)) - C_{\hat{P}}(x(t))\big) \cap B_R(0) \text{ and } x(t) \in K \text{ for a.e. } t \in \mathbb{R} \tag{2.111}$$

and let V *be compact set in the domain of attraction of* A*. Then*

$$\lim_{n\to\infty} \operatorname{dist}(x_n, A) = 0 \text{ a.s. on } \{x_n \in A \text{ infinitely often }\} \tag{2.112}$$

Moreover if V is a Lyapunov function for Eq. 2.111 then the sequence $(x_n)_{n\in\mathbb{N}}$ has a subsequence that converges to the set

$$S := \{x \in K \,|\phi_{\text{max}}(x) = 0\} \tag{2.113}$$

where

$$\phi_{\text{max}}(x) := \max\left\{DV(x)\,[w]\,\middle|w \in \big(F(x) - C_{\hat{P}}(x)\big) \cap T_K(x) \cap B_R(0)\right\}$$

If V has negligible critical values (compare Definition 2.4.1), then x_n converges almost surely to S.

Proof of Theorem 2.4.2. To prove the first statement fix some $\epsilon > 0$ such that

$$U_\epsilon(A) := \{x \in K \,|\text{dist}\,(x, A) < \epsilon\} \subseteq V$$

and fix some $\delta > 0$ such that any solution of Eq. 2.111 with $\text{dist}\,(x(0), A) < \delta$ remains in $U_{\epsilon/2}(A)$. Since the linear interpolation of the sequence, $x(t)$, visits V infinitely often and since V is compact, there exists a sequence t_k such that the family of functions $\phi_k(s) := x\,(t_k + s)$ converges to some solution, x_*, of Eq. 2.111 uniformly on compact intervals by Theorem 2.4.1. Since A is an attractor, there exists some $T > 0$ such that $\text{dist}\,(x_*(t), A) < \delta$ for every $t > T$. Hence $(x_n)_{n\in\mathbb{N}}$ visits $U_\delta(A)$ infinitely often.

If the sequence x_n would leave $U_\epsilon(A)$ infinitely often, then there existed a sequence of real numbers $(s_k)\,k \in \mathbb{N}$ and $(T_k)_{k\in\mathbb{N}}$ such that $x(s_k) \in \partial U_\delta(A)$, $x(s_k + t) \notin U_\delta(A)$ for $0 < t \leq T_k$ and $\text{dist}(x(s_k + T_k), A) = \epsilon$. By Theorem 2.4.1 again there exists a convergent subsequence for the family $\phi'_k(t) := x(s_k + t)$ and we assume without loss of generality that ϕ'_k is already convergent. Since A is asymptotically stable, the limit x'_* satisfies $\text{dist}\,(x_*(t), A) \leq \frac{\epsilon}{2}$ for all t and since A is an attractor $\text{dist}(x(t), A) < \frac{\delta}{2}$ for sufficiently large $t \geq T$.

By uniform convergence on $[0, T]$ we have

$$\sup\left\{\text{dist}\,\big(\phi'_k(t), A\big)\,|0 \leq t \leq T\right\} \leq \frac{3\epsilon}{4}$$

and

$$\text{dist}\,\big(\phi'_k(T), A\big) \leq \frac{3\delta}{4}$$

for sufficiently large k. This contradicts the construction of ϕ'_k. Therefore the assumption is wrong and x_k cannot leave $U_\epsilon(A)$ infinitely often and since ϵ can be chosen arbitrary small, the convergence to A follows.

Now let V be a Lyaponov function for Eq. 2.111. We first argue that ϕ_{max} is bounded from below. To see this consider the function

$$\begin{aligned}\phi_{\text{min}} &:= \min\left\{DV(x)\,[w]\,\middle|w \in \big(F(x) - C_{\hat{P}}(x)\big) \cap T_K(x) \cap B_R(0)\right\} \\ &= -\max\left\{-DV(x)\,[w]\,\middle|w \in \big(F(x) - C_{\hat{P}}(x)\big) \cap T_K(x) \cap B_R(0)\right\}.\end{aligned}$$

It is lower semicontinuous and therefore takes on its finite minimum on K. Therefore there exists some constant $R' \geq 0$ such that

$$\phi_{\text{max}} \geq \phi_{\text{min}} \geq -R'. \tag{2.114}$$

Therefore

$$\liminf_{t\to\infty} V\,(x(t)) := C_{\text{min}} > -\infty \tag{2.115}$$

We will show that any subsequence $x\,(t_k)$ with $\lim_{k\to\infty} V(x(t_k)) = C_{\text{min}}$ converges to S. Assume that this is not true. By taking a further subsequence if necessary we can assume that

$$\liminf_{k\to\infty} \text{dist}\,(x\,(t_k)\,, S) \geq \epsilon > 0 \tag{2.116}$$

Then by upper semicontinuity of $\phi_{\max}$ and by compactness of K we have

$$\sup\left\{\phi_{\max}\left(x\right)|x \in K \setminus U_{\epsilon}\left(S\right)\right\} =: -\delta < 0 \tag{2.117}$$

Taking another subsequence if necessary and employing Theorem 2.4.1 again we can assume that $\phi_k(s) := x(t_k + s)$ converges to some solution x_* of the mean ODE. By construction this solution necessarily has the following properties:

$$V(x_*(0)) = C_{\min} \,;\, \phi_{\min}(x_*(0)) \leq -\delta \tag{2.118}$$

By upper semicontinuity there exists some $\delta' > 0$ such that $\phi_{\max}(y) < -\delta/2$ whenever $|y - x_*(0)| < \delta'$.
We have $\|x_*(t) - x_*(0)\| \leq Rt$ and therefore for $t' := \frac{\delta'}{R}$:

$$V\left(x_*(t')\right) - V\left(x_*(0)\right) \leq \int_0^{t'} \frac{-\delta}{2} = -\frac{\delta\delta'}{2R} \tag{2.119}$$

In this case $x(t'_k)$ would be smaller then $C_{\min} - \frac{\delta\delta'}{4R}$ along some sequence t'_k with $\lim_{k\to\infty} t'_k = \infty$, which is a contradiction since $C_{\min}$ is the limit inferior of $x(t)$.
The set $V_\alpha := \{x \in K\,|V(x) \leq \alpha\}$ is positively Lyapunov stable. To see this fix $\epsilon > 0$ and define:

$$U_{\epsilon,\alpha} := \{x \in K\,|\mathrm{dist}(x, V_\alpha) < \epsilon\} \tag{2.120}$$

$U_{\epsilon,\alpha}$ is an open subset of K and by construction:

$$M := \min\left\{V(x)\,|x \in K \setminus U_{\epsilon,\alpha}\right\} > \alpha \tag{2.121}$$

Then

$$U' := \left\{x \in K \,\middle|\, V(x) < \alpha + \frac{M-\alpha}{2}\right\}$$

is an open neighborhood of V that never leaves $V_{\alpha+\frac{M-\alpha}{2}} \subseteq U_{\epsilon,\alpha}$. If V has negligible critical values then there remain two possibilities:

1. $C_{\min}$ is an isolated critical value. In this case there exists some constant $\epsilon_{\max}$ such that $(C_{\min}, C_{\min} + \epsilon_{\max}]$ does not contain a critical value. Fix $0 < \epsilon < \epsilon_{\max}$. Then there exists some $\delta' > 0$

   $$\sup\left\{\phi_{\max}(x)\,|V(x) \in [-\epsilon, \epsilon_{\max}]\right\} \leq -\delta'$$

   and an argument similar to the previous one shows that any solution to the mean differential inclusion that starts in $V_{C_{\min}+\epsilon_{\max}}$ reaches the set $V_{C_{\min}+\epsilon}$ after a time T bounded from above by $\frac{\epsilon_{\max}-\epsilon}{\delta'}$. Consequently $V_{C_{\min}}$ is an attractor and the first part of the proof gives the desired result.

2. $C_{\min}$ is an isolated limit point in the set of critical values. Then there exist sequences $(\epsilon_n)_{n\in\mathbb{N}}$ and $(\epsilon'_n)_{n\in\mathbb{N}}$ with $0 < \epsilon_n < \epsilon'_n$ with $\epsilon_n \to 0$ such that the interval $(\epsilon_n, \epsilon'_n]$ does not contain any generalized critical values. By the former argument this implies that $V_{C_{\min}+\epsilon_n}$ is an attractor and since ϵ_n tends to zero this together with the first part of the proof gives the desired result.

■

Our main intention is an application of these theorems to gradients with respect to a discontinuous metric (or with respect set-valued metrics). First of all we equip all bilinear forms on $\mathbb{R}^n \times \mathbb{R}^n$, from now on denoted by $\mathrm{bil}\left(\mathbb{R}^n\right)$, with the following norm:

$$\|g\| := \sup\left\{g(v,w)\,|\|v\|_2 = \|w\|_2 = 1\right\} \tag{2.122}$$

where $\|\cdot\|_2$ is the usual Euclidean distance. The following definitions are very natural (the last part is an extensions of Definition 2.3.1):

Definition 2.4.2 - (**Set-valued metrics**)

► **Definition 2.4.2.1:** *A set-valued map*

$$G : K \to 2^{\mathrm{bil}(\mathbb{R}^n)}$$

will be called set-valued metric on K, if for every $x \in K$, $g \in G(x)$ we have that g is symmetric (i.e. $g(v,w) = g(w,v)$ for every $v, w \in \mathbb{R}^n$) and positive definite (i.e. $g(v,v) > 0$ for every $v \in \mathbb{R}^n \setminus 0$).

► **Definition 2.4.2.2:** *Let $G : K \to 2^{\mathrm{bil}(\mathbb{R}^n)}$ be a set-valued metric on $K \subseteq \mathbb{R}^n$ and let $V : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable. A vector $v \in \mathbb{R}^n$ will be called gradient of G at x if there exists $g \in G(x)$ such that for every $w \in \mathbb{R}^n$:*

$$DV(x)\,[w] = g(v,w) \tag{2.123}$$

The set valued map

$$\nabla_G V : K \to 2^{\mathbb{R}^n}\ ; x \mapsto \{v \in \mathbb{R}^n\,|v \text{ is gradient of } V \text{ at } x\} \tag{2.124}$$

will be called set-valued gradient of V with respect to G.

► **Definition 2.4.2.3:** *Let G be a set-valued metric on some compact set $K \subseteq \mathbb{R}^n$ and let $\hat{P}$ be a quasi-projector onto K. Then G will be called $\hat{P}$-adapted if for every $x \in K$, every $v \in C_{\hat{P}}(x)$ and every $g \in G(x)$:*

$$g(v,\cdot) \in N_K(x) \tag{2.125}$$

Note that upper semicontinuity, convexity and compactness of values of G carry over to the gradient map $\nabla_G V$. A discontinuous metric, $\tilde{g}$, on $\mathbb{R}^n$ restricted to K naturally defines an upper semicontinuous map with closed convex images, the Fillipov regularization (compare Anger, Aubin, and Cellina [6], Aubin [7], Aubin and Frankowska [8] and Filippov [72]):

$$G : x \mapsto \cap_{\epsilon>0}\overline{\mathrm{co}\left[\cup_{y\in B_\epsilon(x)}\{g(x)\}\right]} \tag{2.126}$$

Note that G is a set-valued metric if and only if $\tilde{g}$ is lower bounded in operator norm and has compact values if and only if $\tilde{g}$ is upper bounded in operator norm. We need to impose some regularity condition on the objective function, (compare Definition 2.4.1) V:

Definition 2.4.3 - (**Functions with negligible set of critical values**)

A continuously differentiable function $V : \mathbb{R}^n \to \mathbb{R}$ has a negligible set of critical values over some set K if the set of first order optimal values

$$S := \{y\,|\text{ there exists } x \in K \text{ s.t. } y = V(x) \text{ and } DV(x) \in N_K(x)\} \tag{2.127}$$

has isolated accumulation points only.

After this preparation we can formulate and prove the main theorem of this section:

Theorem 2.4.3 - (**Set-valued stochastic gradient ascent**)

► **Theorem 2.4.3.1:** *Let $G : K \to 2^{\mathrm{bil}\,\mathbb{R}^n}$ be an upper semicontinuous set-valued metric with convex and compact values. Let $V : \mathbb{R}^n \to \mathbb{R}$ be a continuously differentiable function. Let*

$$h : \Omega \times K \to \mathbb{R}^n \tag{2.128}$$

be a random selection of $\nabla_G V$. *Consider the iterative sequence Eq. 2.62 (with* $g := h$*) and assume that Assumption 2.4.1 holds. Assume further that the set-valued metric* G *is adapted to the quasi-projector* $\hat{P}$. *Then* x_n *is infinitely often in a neighborhood of first-order optimal points:*

$$S := \{x \in K \,|DV(x) \in N_K(x)\}\,. \tag{2.129}$$

► **Theorem 2.4.3.2:** *Assume the same situation as in Theorem 2.4.3 but with* V *having negligible critical values on* K *(according to Definition 2.4.3). Then* x_n *converges to* S.

Proof of Theorem 2.4.3. We show that $-V$ is a Lyapunov function for the system. For every vector $w \in \mathbb{R}^n$ and every metric $g \in G(x)$ we have:

$$DV(x)\,[v] = g\left(\nabla_g V(x), v\right) \tag{2.130}$$

hence for the special case that $v = \nabla_g V(x) - n$ with $n \in C_{\hat{P}}(x)$:

$$-\,DV(x)\,[v] = -g\left(\nabla_g V(x), \nabla_g V(x) - n\right) \tag{2.131}$$

for almost every $t \in \mathbb{R}$ we have $\frac{d}{dt}(x)(t) \in T_K(x) \cap -T_K(x)$ by the viability theorem (the second claim follows from differentiability almost everywhere.). Therefore, since G is $\hat{P}$-adapted, $g\left(n, \frac{d}{dt}x(t)\right) = 0$ for almost every $t \in \mathbb{R}$ and hence:

$$-\,DV(x)\,[v] \in \left\{-g\left(\nabla_g V(x) - n, \nabla_g V(x) - n\right) \big| g \in G(x), n \in C_{\hat{P}}(x)\right\} \subseteq (-\infty, 0] \tag{2.132}$$

Moreover $DV(x)\left[\frac{d}{dt}x(t)\right] = 0$ if and only if there exists $g \in G(x)$ such that $\nabla_g V(x) \in C_{\hat{P}}(x)$ and therefore $DV(x) = g\left(\nabla_g V(x), \cdot\right) \in N_K(x)$. Therefore the claim follows from Theorem 2.4.2. ■

Part II

Learning in the sensorimotor loop

Chapter 3

Markov decision processes, learning algorithms and policy functionals

In this chapter we define and motivate the class of learning problems that we will investigate in Chapter 4. We start with a mathematical model describing the interaction between the agent and the world (see Definition 3.1.1 and Remark 3.1.1). The dynamic originates from the well-known Markov decision problem (compare for example Eugene and Feinberg [71], Dynkin and Yushkevich [68] and Bertsekas and Shreve [23]).

Very often the terms "Markov decision problem" and "Markov decision process" are used synonymously. We however will carefully separate the dynamical model of the agent-environment interaction from the optimization problem. The name, Markov decision process (MDP), is reserved for the pure dynamical model within this thesis. This model is a special instance of a causal model over a recursively constructible graph, as defined in Chapter 1. The language developed in that chapter allows a mathamtical rigorous definition of a learning algorithm over an MDP. It can be formulated an appropriate controller extension of the original causal model (compare Definition 1.2.5). We provide the corresponding definitions and emphasize some immediate consequences from the general theory in Chapter 1.

In the usual formulation of a Markov decision problem the dynamical part is not separated from the optimization problem, which includes the maximization of the expected reward of some reward function. There are two reasons why we distinguish strictly between the dynamics of the agent world interaction on the one hand and the optimization problem on the other hand.

First of all, we want to consider real learning problems. By this we mean that not all details of the system are known at the beginning. We need a clear statement of an entire class of possible transition models that are *a priori* considered to be possible. On a technical level this is done by considering a collection of probability laws on the causal model describing the agent-environment interaction (compare definition of causal statistical models over causal models in Definition 1.2.2 of Chapter 1). The fundamental assumption on these laws, their properties and relations are crucial for the formulations of the learning problems and later proofs. We therefore prefer a rigorous, well-motivated treatment of these purely dynamical questions leaving aside the optimization problem at first.

The second reason for a separate treatment of the optimization problem and the dynamical laws governing the model lies in the fact that we would like to generalize the common objective functionals, the expected reward of some reward function. Different instances of reward maximization have been discussed extensively in the literature, for example optimization over a finite time horizon, the expected discounted reward, the mean time average reward etc. (a good overview is given in Eugene and Feinberg [71]). We are however also interested in a formal treatment of situations that are not covered by reward maximization, like:

- a risk averse agent (who might try to decrease the variance of some random variable of later sensor values). This behavior is also a desirable goal for tracking problems or stabilization algorithms,
- an agent optimizing certain information measures (to be discussed and clarified within this chapter),
- an agent trying to control the ergodic properties of the sensor process,

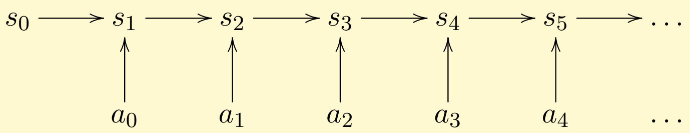
to mention only a few examples. The interest in these functionals especially arises from recent goals in artificial intelligence to make robots more "curious" or to let them explore the environment in an efficient way that is restricted to and/or guided by their embodiment. A general applicable example of an "behavior optimizing algorithm" is a gradient ascent algorithms maximizing the predictive information (compare Ay et al. [17], Zahedi, Ay, and Der [196] and Ay et al. [16]). The predictive information is high if the stationary distribution of the sensor values has high entropy while still allowing a prediction of the next sensor value with high accuracy. Therefore maximization of the predictive information is expected to lead to an exploration of large parts of the state space in a highly coordinated way. Simulations of physical robots following these algorithms show very interesting behavior that often reminds to playing or fighting machines (see for example Der and Martius [63] and the website: http://www.playfulmachines.com/).

The main results from the current chapter are:

- a rigorous definition of the mathematical models of the agent-world dynamic for a non-learning and a learning agent (see Definition 3.1.1 and Remark 3.1.1),
- a clear definition of a class of learning/optimization problems that can describe the scenarios listed in the enumeration above (see Problem 3.2.1 and the special instances Problem 3.2.2 and Problem 3.2.3) ,
- a collection of relevant examples like the expected discounted reward, the long time average reward, the variance of the expected discounted reward, the predictive information and other discounted and ergodic information measures of the sensor process of an MDP (see section 3.3),
- a collection of gradient formulas for several sensor process functionals for finite state space MDPs (see Remark 3.2.1 and Remark 3.2.2 for two very general formulas and section 3.3 for applications) and
- a theorem about the relation between discounted sensor functionals and their ergodic counterparts in the case of finite state and action spaces. The theorem proves and generalizes the insight that the optimization of a discounted reward is "somehow similar" to the optimization of the reward in the stationary distribution in the limit of the discount factor approaching 1 from below. (see Theorem 3.2.1)

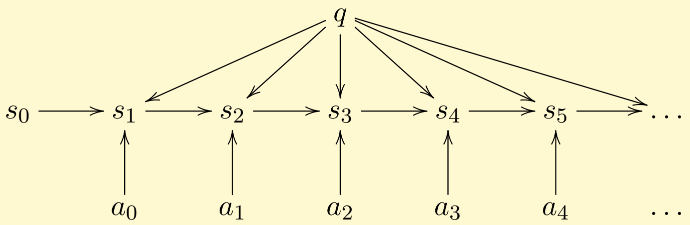
3.1 Markov decision processes and learning algorithms over Markov decision processes

Consider an agent interacting with the environment. We assume that the agent performs a sequence of actions, a_n, and receives a sequence of sensor values, s_n. We assume that the next sensor value, s_{n+1}, depends probabilistically on the old sensor value, s_n, and the action performed by the agent, a_n. At this stage we assume that there is no control, i.e. the actions of the agent are independent of the sensor process. This situation is described by the following recursively constructible graph (see Definition 1.2.1):



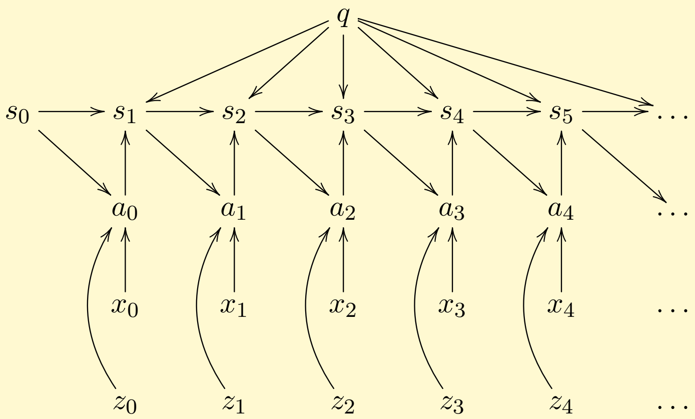
Caus. mod. 17 - *Causal structure uncontrolled, unparameterized MDP*

We are interested in scenarios where the agent has only partial information about the world dynamics. By this we mean that there are several *a priori* possible transition laws from a given state-action pair to the new state. We assume that this transition law is the same at every instance of time. This situation can be modeled by introducing a parameter vertex q into the previous causal graph. The next sensor value then depends on the former sensor value, the previous action and the parameter vertex:



Caus. mod. 18 - *Causal structure uncontrolled, parameterized MDP*

For the final graph we introduce a feedback loop such that the new action depends on the former sensor value. For applications it is frequently very convenient to realize the transition from an old state- action pair to a new sensor value as a stochastically perturbed, function. We therefore introduce some randomization variables, x_n, and some policy parameters, z_n. This results in the following recursively constructible graph:



Caus. mod. 19 - *Causal structure controlled, parameterized MDP*

In principle the randomization variables, x_i, are superfluous for the description but have some technical adventages. The idea behind the construction is the following:

- The agent observes some sensor value $s \in \mathbf{S}$ and is supposed to generate the next action with some probability distribution $P_{s,z}$ where z is the policy parameter.

- To do so, some $\mathbf{X}$-valued sample, x, is drawn from a known distribution, p. The next action is chosen to be $a' := \Pi(s, x, z)$ where Π is a (deterministic) function. This procedure produces a random variable with the desired distribution, $P_{s,z}$, if and only if $\Pi(s, \cdot, z)_* p = P_{s,z}$.

In spite of being mathematical equivalent in many purposes (see Theorem 1.2.2) a realization of random transtions with randomized functions is sometimes more convenient than a direct specification via transition kernels. Many models (for example autoregressive models) specify transitions in the former way. This also often simplifies a direct computer-based implementation.

The dynamical model of the agent-environment interaction is completely specified by a causal model over the recursively constructible graph, Caus. mod. 19 (for the definition of a causal model see Definition 1.2.1). The complete definition of any causal model involves a specification of all state spaces, all transition kernels and a statistical model over the causal model (compare Definition 1.2.2). Since we are interested in time homogeneous situations only (i.e. time-independent state and action spaces, time independent transition rules etc.), the most general definition of an MDP is the following:

Definition 3.1.1 - (**Markov decision process and the associated causal model**)

► **Definition 3.1.1.1:** *A Markov decision process is an eight-tuple* $(\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *where*

- $\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}$ *are measurable spaces. If needed the σ-algebras will be denoted by $\mathcal{F}_Q, \mathcal{F}_Z$ and so on.*
- $\mathbf{Q}$ *denotes the set of parameters for the world transition kernel,* $\mathbf{Z}$ *denotes the set of policy parameters.*
- $\mathbf{S}$ *is the space for the sensor values,* $\mathbf{A}$ *denotes the set of actions.*
- $\mathbf{X}$ *denotes the space of possible noise outcomes for the policy transition and sensor transition respectively. Usually we will take* $\mathbf{X} := [0, 1]$ *(this choice is sufficient for most practical purposes, as Theorem 1.2.2 shows).*
- p_x *is a probability measure on* $\mathcal{F}_X$ *- the distribution of the policy randomization variable. Usually we will choose* p_x *to be the uniform distribution on* $[0, 1]$.
- $T \in \Lambda^{\mathbf{S}}_{\mathbf{S}\times\mathbf{A}\times\mathbf{Q}}$ *is a transition kernel, describing the sensor updating dynamic.*
- $\Pi : \mathbf{S} \times \mathbf{X} \times \mathbf{Z} \to \mathbf{S}$ *is a measurable map (encoding the policy transition function).*

► **Definition 3.1.1.2:** *Let* $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a Markov decision process and let* (V, E) *denote the recursively constructible graph Caus. mod. 19 . The causal model* $C' := ((V, E), \mathfrak{S}, \mathfrak{T})$ *(compare Definition 1.2.1) associated with* C *is a causal model over* (V, E) *defined by*

$$\mathfrak{S}_v = \begin{cases} \mathbf{S} \text{ iff } v = s_i \\ \mathbf{A} \text{ iff } v = a_i \\ \mathbf{Q} \text{ iff } v = q \\ \mathbf{Z} \text{ iff } v = z_i \\ \mathbf{X} \text{ iff } v = x_i \end{cases}$$

and

$$\mathfrak{T}_v = \begin{cases} T \text{ iff } v = s_i \\ \Pi \text{ iff } v = a_i \end{cases}$$

Remark 3.1.1 - (**Comment on Definition 3.1.1 and definition of the statistical model over an MDP**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a Markov decision process and let* $C' := ((V, E), \mathfrak{S}, \mathfrak{T})$ *be the associated causal model. We will follow the general construction and notation introduced in Chapter 1. The set of initial vertices of the MDP graph is:*

$$V_0 = \{s_0, q\} \cup_{i \in \mathbb{N}_0} \{z_i, x_i\}$$

By Theorem 1.2.1 there exists a unique probability kernel

$$\hat{K} \in \Lambda^{(\mathfrak{S}, \mathcal{F})}_{\left(\mathfrak{S}_{V_0}, \otimes_{v \in V_0} \mathcal{F}_v\right)}$$

with the property that $\hat{K}(\mathbf{v}, \cdot) \in M_1(\mathcal{F}_V)$ *is compatible with the causal structure and with the initial configuration* $\mathbf{v} \in \mathfrak{S}_{V_0}$*, i.e. the conditional probability distributions coincide with the corresponding kernels and for every* $B \in \otimes_{v \in V_0} \mathcal{F}_v$ *we have*

$$\hat{K}\left(\mathbf{v}, \left\{(\pi_v)_{v \in V_0} \in B\right\}\right) = \delta_{\mathbf{v}}(B),$$

where again $\pi_v : \mathfrak{S} \to \mathfrak{S}_v$ *denotes the projection of* $\mathfrak{S}$ *onto the factor* $\mathfrak{S}_v$*. The distribution of each noise variable is required to coincide with* p_x*, i.e.* π_{x_i} *is assumed to have distribution* p_x*. Hence by Theorem 1.2.1 for any given initial values,* $q' \in \mathbf{Q}$*,* $s' \in \mathbf{S}$ *and* $\mathbf{z}' \in \mathbf{Z}$ *the law of the process is given by:*

$$P_{\text{MDP},q',s',\mathbf{z}'}[A] = \int_{\mathbf{X}^{\mathbb{N}_0}} \hat{K}\left(\left(q', s', \mathbf{z}', \mathbf{x}'\right), A\right) \left(\otimes_{v \in \{x_i | i \in \mathbb{N}_0\}} p_x\right) \left(d\mathbf{x}'\right) \tag{3.1}$$

As the MDP is a specific example of a causal model over a recursively constructible graph, this law satisfies the independence properties of Theorem 1.3.1 and Theorem 1.4.1. For every $v \in V$ *the canonical projections* $\pi_v : \mathfrak{S} \to \mathfrak{S}_v$ *are random variables on the probability space* $\left(\mathfrak{S}, \mathcal{F}_V, P_{\text{MDP},q',s',\mathbf{z}'}\right)$*. To improve readability we will frequently denote these random variables by the same letter as the corresponding vertex but with an upper-case letter. So we will write* S_i *for* π_{s_i} *for example. The collection of all process laws forms a statistical model over the probability space* $(\mathfrak{S}, \mathcal{F}_V)$*.*

The set $\{s_n, q\}$ d-separates $\{s_j \,|0 \leq j < i\}$ and $\{s_j \,|j > i\}$ in the recursively constructible graph of the MDP (see Caus. mod. 19). Therefore Theorem 1.3.1 implies

$$S_{n+1} \perp\!\!\!\perp (S_k)_{k<n} \,|S_n, Q \tag{3.2}$$

Moreover by compatibility with the causal structure and the fact that $Q = q'$ almost surely (w.r.t. $P_{\text{MDP},q',s',\mathbf{z}'}$):

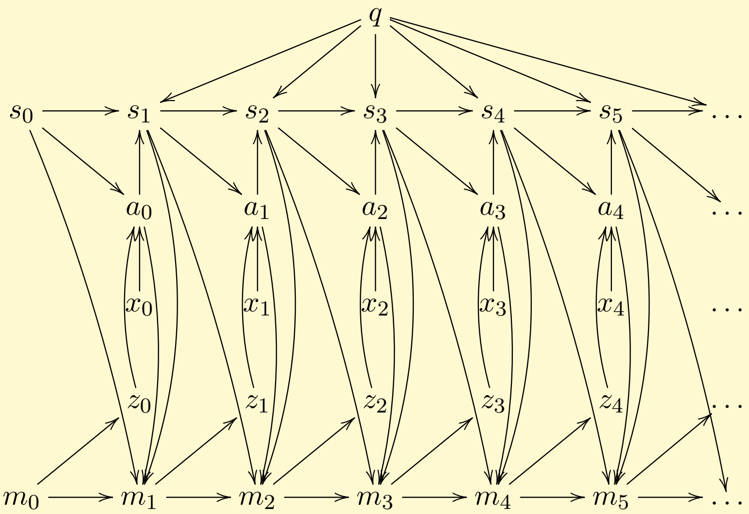
$$\begin{aligned} & P_{\text{MDP},q',s',\mathbf{z}'}\left[S_{n+1} \in A \,\middle|\, S_n = s, Q = q''\right] \\ = & \int_{\mathbf{X}} T\left[\left(s, \Pi(s, x, \mathbf{z}'_i), q'\right), A\right] p_x(dx) =: K_{q',\mathbf{z}'_i}(s, A) \text{ a.s.} \end{aligned} \tag{3.3}$$

is a regular version of the conditional distribution of S_{n+1} given $(S_i)_{i \leq n}$ and Q. Since $K_{q',\mathbf{z}'_i}$ does not depend on q'' this implies:

$$P_{\text{MDP},q',s',\mathbf{z}'}\left[S_{n+1} \in A \,\middle|\, S_n = s, (S_i)_{i<n}\right] \;=\; K_{q',\mathbf{z}'_i}(s, A) \text{ a.s.} \tag{3.4}$$

such that $(S_i)_{i \in \mathbb{N}_0}$ is a Markov process under $P_{\text{MDP},q',s'_0,\mathbf{z}'}$ with initial distribution $\delta_{s'}$ and transition kernels $K_{q',\mathbf{z}_i}$.

So far the policy parameter sequence does not depend on the observed process. This is precisely what is needed to describe an agent that learns some policy from observation of the process. Therefore we need to define a controlled dynamic over the original MDP that captures the situation of a learning agent. First of all we introduce additional memory variables m_i where $i \in \mathbb{N}_0$ and allow m_n (where $n \geq 1$), to depend on the former memory value m_{n-1}, the sensor value s_{n-1}, the action a_n and the state s_n. Furthermore the values of the parameter vertices, z_n, are now controlled by the agent (i.e. they depend on the current memory value) The former causal model Caus. mod. 19 will be extended to the following one:



Caus. mod. 20 - *Learning algorithm over controlled, parameterized MDP*

The specification of a learning algorithm requires the definition of the memory space, $\mathbf{M}$, a memory update Function:

$$L : \mathbf{S} \times \mathbf{A} \times \mathbf{S} \times \mathbf{M} \to \mathbf{M}$$

and an update rule for the policy parameter from the current memory value:

$$U : \mathbf{M} \to \mathbf{Z}$$

This specification gives rise to the following natural definition of a learning algorithm over an MDP:

Definition 3.1.2 - (Learning algorithm over a Markov decision process and the associated controller extension)

▶ **Definition 3.1.2.1:** *Let* $C = (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a Markov decision process. A learning algorithm over* C *is a triple* $(\mathbf{M}, L, U)$ *where*

- $\mathbf{M}$ *is a measurable set, the set of possible memory values*
- $L : \mathbf{S} \times \mathbf{A} \times \mathbf{S} \times \mathbf{M} \to \mathbf{M}$ *is a measurable function, referred to as memory update rule.*
- $U : \mathbf{M} \to \mathbf{Z}$ *is a measurable function, referred to as policy choice function*

▶ **Definition 3.1.2.2:** *Let* $M = (\mathbf{M}, L, U)$ *be a learning algorithm over some MDP,* $C = (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$*. Then* M *defines a natural controller extension,* $((V', E'), \mathfrak{S}', \mathfrak{T}')$*, of the causal model associated to* C *(compare Definition 1.2.5 and Definition 3.1.1) via:*

- (V', E') *is the graph, Caus. mod. 20,*

- $\mathfrak{S}'_{m_i} := \mathbf{M}$ *for every* $i \in \mathbb{N}_0$*;*
- $\mathfrak{T}'_{m_i} \in \Lambda^{\mathbf{M}}_{\mathfrak{S}_{s_{i-1}} \times \mathfrak{S}_{a_{i-1}} \times \mathfrak{S}_{s_i} \times \mathfrak{S}'_{m_{i-1}}}$ *where* $i \geq 1$ *is defined as* $\mathfrak{T}'_{m_i}\left[(s,a,s',m), A\right] = \delta_{L(s,a,s',m)}(A)$ *and*
- $\mathfrak{T}'_{z_i} \in \Lambda^{\mathbf{Z}}_{\mathfrak{S}_{m_i}}$ *where* $i \geq 0$ *is defined as* $\mathfrak{T}'_{z_i}\left[m, A\right] := \delta_{U(m)}\left[A\right]$

We will also refer to this controller extension as the sensorimotor loop associated to the learning algorithm M*.*

Remark 3.1.2 - (**Technical remark on Definition 3.1.2**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a Markov decision process (with associated causal model* $C' = ((V,E), \mathfrak{S}, \mathfrak{T})$*). And let* $M = (\mathbf{M}, L, U)$ *be a learning algorithm over* C *(with associated sensorimotor loop* $M' = ((V',E'), \mathfrak{S}', \mathfrak{T}')$*). Again we will follow the general construction and notation introduced in Chapter 1. The set of initial vertices of* V' *is*

$$V'_0 = \{x_n \,|n \in \mathbb{N}_0\} \cup \{q, s_0, m_0\}$$

As in Lemma 1.2.4 we assume the configuration of the extended model to be a random variable over the probability space $(\mathfrak{S}, \otimes_{v \in V} \mathcal{F}_v)$*. Let*

$$K'' \in \Lambda^{(\mathfrak{S}, \otimes_{v \in V} \mathcal{F}_v)}_{\left(\mathfrak{S}_{V'_0}, \otimes_{v \in V'_0} \mathcal{F}_v\right)}$$

and $R : \mathbf{M} \times \mathfrak{S}_V \to \mathfrak{S}_{V'}$ *be the process generating kernel and the random variable from Lemma 1.2.4. Let* $q' \in \mathbf{Q}$ *be the world parameter, let* $m' \in \mathbf{M}$ *be the initial memory state and let* $s' \in \mathbf{S}$ *be the initial sensor state then the associated measure on* $\mathcal{F}_V$ *is*

$$P_{q',s',m'}\left[B\right] = \int_{\mathbf{X}^{\mathbb{N}_0}} K''\left[\left(m', q', s', \mathbf{x}\right), B\right] d\left(\otimes_{n \in \mathbb{N}_0} p_x\right)(d\mathbf{x}) \tag{3.5}$$

for every $B \in \mathcal{F}_V$*. The law of the process is*

$$R\left[m', \cdot\right]_* P_{q',s',m'}\left[B\right] = \int_{\mathbf{X}^{\mathbb{N}_0}} R\left[m', \cdot\right]_* K''\left[\left(m', q', s', \mathbf{x}\right), B\right] d\left(\otimes_{n \in \mathbb{N}_0} p_x\right)(d\mathbf{x}) \tag{3.6}$$

where $B \in \mathcal{F}_{V'}$*. The independence properties of Theorem 1.3.1 and Theorem 1.4.1 also hold for the sensorimotor loop associated to* M *under every measure* $P_{q',s',m'}$*. As in the MDP case we will abbreviate* π_v *by an upper case letter that is equal to the corresponding vertex label.*

Note that the sensor process, $(S_i)_{i \in \mathbb{N}_0}$ is in general not Markovian under $P_{q',s',m'}$, the measure of the process with learning algorithm from Remark 3.1.2, anymore. However by Theorem 1.3.1 and by the identity $Q = q'$ a.s. the process of pairs $((S_i, M_i))_{i \in \mathbb{N}_0}$ is a Markov process under $P_{q',s',m'}$, with initial states, $S_0 = s'$, $M_0 = m'$ and transition kernel, $\tilde{K}_{q'} \in \Lambda^{\mathbf{S} \times \mathbf{M}}_{\mathbf{S} \times \mathbf{M}}$ given by:

$$\begin{aligned} & \tilde{K}_{q'}\left[(s,m), A_s \times A_m\right] \\ = & \int_{\mathbf{S} \times \mathbf{X}} T\left[\left(s, \Pi\left(s, x', U(m)\right), q'\right), ds'\right] \mathbb{1}_{A_m}\left[L\left(s, \Pi\left(s, x', U(m)\right), s', m\right)\right] \cdot \\ & \cdot \mathbb{1}_{A_s}\left(s'\right) p_x(dx') \end{aligned} \tag{3.7}$$

with $A_s \in \mathcal{F}_s$ and $A_m \in \mathcal{F}_m$.

Having specified the dynamical models of an MDP and the dynamical model of a learning algorithm over an MDP [1] we will formulate a class of learning problems related to MDPs in the next section. Finally we will prove the convergence of certain gradient-based learning algorithms for this class of problems in Chapter 4.

For the analysis of concrete models in Chapter 4 it is rather inconvenient to write down $\mathbf{M}$, L and U explicitly. We will rather introduce some new variables (G_n for example) and specify how they depend on the old ones (for example: $G_{n+1} = f(S_n, A_{n-1}, G_n)$ where $f : \mathbf{S} \times \mathbf{A} \times \mathbf{G} \to \mathbf{G}$ is some measurable function. It is clear that all new variables are part of the memory and how to translate this specification into our definition of a learning algorithm.

3.2 Optimizing policy functionals under an unknown sensor transition dynamic

3.2.1 Formulation of the problem

We consider the following general optimization problem.

Problem 3.2.1 - (**The general learning problem: Optimizing policy functionals under unknown sensor transition rule**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be an MDP. Consider some function*

$$\phi : \mathbf{Q} \times \mathbf{Z} \to \mathbb{R} \tag{3.8}$$

and the following optimization problem: Find the supremum of ϕ *for given* $q \in \mathbf{Q}$*:*

$$M_{\phi,q} := \sup \{\phi(q,z) \,|z \in \mathbf{Z}\} \tag{3.9}$$

And find the (possibly empty) set of maximizers:

$$L_{\phi,q} := \{z \in \mathbf{Z} \,|\phi(z,q) = M_{\phi,q}\} \tag{3.10}$$

We are looking for learning algorithms that find (or approximate) both $M_{\phi,q}$ and $z \in L_{\phi,q}$ for every possible parameter $q \in \mathbf{Q}$. This parameter is not known to the agent and can at best be estimated from observations to arbitrary precision. In other words optimally we expect some memory variable to converge to $M_{\phi,q'}$ almost surely w.r.t. $P_{q',s',m'}$ for some initial memory state $m' \in \mathbf{M}$, for every $q' \in \mathbf{Q}$ and every $s' \in \mathbf{S}$.

The formulation Problem 3.2.1 is probably the most general learning problem that can be formulated within the MDP framework. We did not require that the map between the "world" transition parameter, $q \in \mathbf{Q}$ and the transition kernel $T \in \Lambda^{\mathbf{S}}_{\mathbf{S}\times\mathbf{A}\times\mathbf{Q}}$ is one-to one. Setting $q = (q_1, q_2)$ and assuming that T does not depend on q_2 even allows the specification of a partially unknown objective function. However since in this case the entire process law does not depend on q_2, it is impossible to learn the optimal value (in some reinforcement learning problems the reward function is not known in the beginning - this is however

---

[1] Note that we do not impose any restriction on the memory space at this point. Therefore the definition is very broad-range. The memory space can be rich enough to store all former memory values, sensor outcomes and actions. Moreover the memory can contain counter variables which can be used to implement time dependent transition functions. On the conceptual level it would be no difference to update the memory using the current sensor value only instead of both the last sensor value and the current one - the former one could simply be stored in memory and be reused. Nevertheless our definition is more handy for the problems that we will discuss in the next chapter.

very different from the scenario outlined here, since the expectation of the reward function can actually be learned from observations we will come back to this problem later on).

Before proceeding with possible applications we will introduce two interesting instances of Problem 3.2.1. Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ be a MDP. As shown in Eq. 3.1 following Definition 3.1.1 the distribution of the sensor process, $(\mathbf{S}_n)_{n \in \mathbb{N}_0}$, is Markovian under the law $P_{\mathrm{MDP},q',s',\mathbf{z}'}$. If the policy sequence is stationary, i.e. $\mathbf{z}_i \equiv z'$ for some $z' \in \mathbf{Z}$ then the sensor process is time-homogeneous with transition kernel $K_{q,z'}$ (compare Eq. 3.3 and Eq. 3.4).

In many applications the policy parameter, z', has to be chosen such that a given functional of the process law, $P_{\mathrm{MDP},q',s',\mathbf{z}'}$, is maximized. Since a time homogeneous Markov process is completely specified by the initial measure and the transition kernel, the following special instance of Problem 3.2.1 will be called sensor process functional:

Problem 3.2.2 - (**Sensor process functionals**)

Let $(\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be an MDP. A function*

$$f : M_1(\mathcal{F}_S) \times \Lambda_{\mathbf{S}}^{\mathbf{S}} \to \mathbb{R} \tag{3.11}$$

will be called sensor process functional. Moreover let $\mu \in M_1(\mathcal{F}_S)$ *and define:*

$$\begin{aligned} \kappa : \quad & \mathbf{Q} \times \mathbf{Z} \to M_1(\mathcal{F}_S) \times \Lambda_{\mathbf{S}}^{\mathbf{S}} && (3.12) \\ & (q, z) \mapsto (\mu, K_{q,z}) && (3.13) \end{aligned}$$

where the kernel $K_{q,z}$ *has been defined in Eq. 3.3. The pair* (μ, f) *naturally defines an instance of Problem 3.2.1 by setting* $\phi := f \circ \kappa$.

often the one-point sensor distribution of a Markov process converges to some invariant distribution (compare Appendix A.1.3). In this case one is often interested in the process in its "equilibrium distribution", such that the following instance of Problem 3.2.1 is very natural:

Problem 3.2.3 - (**Ergodic functionals**)

Let $(\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be an mDP and let*

$$f : M_1(\mathcal{F}_S) \times \Lambda_{\mathbf{S}}^{\mathbf{S}}(\mathcal{F}_S) \to \mathbb{R} \tag{3.14}$$

be a sensor transition functional (compare Problem 3.2.2). Let

$$\begin{aligned} \mathrm{I\tilde{N}V} : \quad & \Lambda_{\mathbf{S}}^{\mathbf{S}} \to M_1(\mathcal{F}_S) \times \Lambda_{\mathbf{S}}^{\mathbf{S}} && (3.15) \\ & K \mapsto (\nu, K) \ \text{ with } \nu K = \nu && (3.16) \end{aligned}$$

map a transition kernel to a pair consisting of an invariant distribution for this kernel and a copy of this kernel. We will consider scenarios only where $\mathrm{I\tilde{N}V}(K)$ *exists and is uniquely defined for every reachable kernel,* K*. The functional* $f \circ \mathrm{I\tilde{N}V}$ *will be called ergodic functional associated to* f*. Ergodic functionals naturally define an instance of Problem 3.2.1 by setting*

$$\phi(q, z) := f \circ \mathrm{I\tilde{N}V}(K_{q,z}), \tag{3.17}$$

where the kernel $K_{q,z}$ *has been defined in Eq. 3.3.*

3.2.2 Finite state and action spaces

A special case of high relevance for practical applications are MDPs with finite state and action spaces. In this case it is possible to write down explicit formulas for first-order stationarity conditions for Problem 3.2.2 and Problem 3.2.3. We assume that the MDP satisfy all the conditions in Assumption 4.1.1 in Chapter 4. The set of policy parameters is the $\mathbf{S}$-fold ϵ-simplex and by Example 2.1.2 the tangent cone at some policy $z \in \mathbf{Z}_\epsilon$ is given by:

$$T\mathbf{Z}_\epsilon\left(z\right) = \left\{ C \in \mathbb{R}^{\mathbf{S}\times\mathbf{A}} \,\middle|\, \sum_{a\in\mathbf{A}} C_{s,a} = 0; c_{s,a} \geq 0 \text{ whenever } z\left(s,\{a\}\right) = \epsilon \right\} \tag{3.18}$$

The sensor transition kernel depends linearly on the policy matrix:

$$K_{q,z}\left(s,\{s'\}\right) := \sum_{a\in\mathbf{A}} z\left(s,a\right) q\left(\left(s,a\right),\{s'\}\right) \tag{3.19}$$

Let f be a sensor process functional that continuously differentiable with respect to the policy parameter. Let ϕ denote the associated policy functional (compare Problem 3.2.2). By the chain rule the directional derivative with respect to the policy of the objective function, ϕ, is:

$$D_z\phi(q,z)\left[C\right] = \sum_{s'\in\mathbf{S}} \frac{\partial f}{\partial K\left(s,\{s'\}\right)}\left(\mu, K_{q,z}\right)\left(\sum_{a\in\mathbf{A}} q\left(\left(s,a\right),\{s'\}\right) C_{s,a}\right) \tag{3.20}$$

for every $C \in T\mathbf{Z}_\epsilon\left(z\right)$. This directional derivative can be used to write down an explicit formula for the Euclidean gradient and the Fisher gradient of ϕ (compare Appendix A.2.2). The result is:

Remark 3.2.1 - (**Gradient formulas for policy functionals**)

Let f be a sensor process functional defining the policy functional ϕ (as in Problem 3.2.2). The Euclidean policy gradient of ϕ is

$$\begin{aligned} &\nabla_{E,z}\phi\left(q,z\right)_{s,a} \\ = &\sum_{s'\in\mathbf{S}} \frac{\partial f}{\partial K\left(s,\{s'\}\right)}\left(\mu, K_{q,z}\right)\cdot\left(q\left(\left(s,a\right),\{s'\}\right) - \frac{1}{|\mathbf{A}|}\sum_{a'\in\mathbf{A}} q\left(\left(s,a'\right),\{s'\}\right)\right) \end{aligned} \tag{3.21}$$

and the Fisher policy gradient is

$$\begin{aligned} &\nabla_{F,z}\phi\left(q,z\right)\left(q,z\right)_{s,a} \\ = \;& z\left(s,\{a\}\right)\sum_{s'\in\mathbf{S}} \frac{\partial f}{\partial K\left(s,\{s'\}\right)}\left(\mu, K_{q,z}\right)\cdot\left(q\left(\left(s,a\right),\{s'\}\right) - K_{q,z}\left(s,\{s'\}\right)\right) \end{aligned} \tag{3.22}$$

The policy derivative in Problem 3.2.3 is based on the derivative of the (generally set-valued) map

$$\begin{aligned} \text{INV}\, \Lambda_{\mathbf{S}}^{\mathbf{S}} &\to 2^{M_1(\mathbf{S})} \\ K &\mapsto \left\{\mu \in M_1\left(\mathbf{S}\right) \middle| \mu K = \mu\right\} \end{aligned} \tag{3.23}$$

The map INV is single-valued on $\mathbf{Q} \times \mathbf{Z}$ by the ergodic theorem for Markov chains (see Appendix Theorem 6.0.2) and analytical (see Schweitzer [168]). We collected the result of Schweitzer and some related perturbation statements for finite state space Markov chains in Theorem 6.0.3 in the Appendix. The derivative of INV is given by:

$$\left(D\,\text{INV}\right)\left(K\right)\left[C\right] = \text{INV}\left(K\right)CY_K \text{ for every } C \in T_{\Delta_{\mathbf{S}}\mathbf{s}}\left(K\right) \tag{3.24}$$

where

$$Y_K = (\mathbb{1} - K + K_*)^{-1} - K_* \tag{3.25}$$

with $K_* = \lim_{n\to\infty} K^n$.

Now let $f : \Delta_{\mathbf{S}} \times {\Delta_{\mathbf{S}}}^{\mathbf{S}} \to \mathbb{R}$ be a sensor process functional (compare Problem 3.2.2) and consider the associated ergodic functional (compare Problem 3.2.3):

$$\phi_{\text{erg}}(q, z) := f\left(\text{INV}\left(K_{q,z}\right), K_{q,z}\right) \tag{3.26}$$

Using the chain rule and some calculus yields the policy derivative of ϕ_{erg}:

$$\begin{aligned} & D_z\phi_{\text{erg.}}(q, z)\,[C] \\ = \; & \sum_{s'\in\mathbf{S}}\sum_{a'\in\mathbf{A}} q\left((s, a'), \{s'\}\right) C_{s,a'} \mathcal{D}_{\text{erg.}} f\left(K_{q,z}\right)_{s,s'} \end{aligned} \tag{3.27}$$

where we introduced some shortening for what we will call ergodic derivative of f:

$$\begin{aligned} & \mathcal{D}_{\text{erg.}} f\left(K\right)_{s,s'} \\ := \; & \left(\frac{\partial f}{\partial K\left(s, \{s'\}\right)}\left(\text{INV}\left(K\right), K\right) + \text{INV}\left(K\right)(\{s\}) \sum_{s_1\in\mathbf{S}} Y_{K;s',s_1} \frac{\partial f}{\partial p(\{s_1\})}\left(\text{INV}\left(K\right), K\right)\right) \end{aligned} \tag{3.28}$$

with Y_K given by Eq. 3.25. This yields the following two gradient formulas for Problem 3.2.3:

Remark 3.2.2 - (**Gradients formulas for policy functionals - part 2**)

Let f be a sensor process functional giving rise to the ergodic functional $\phi_{\text{erg.}}$ (see Eq. 3.26 and Problem 3.2.3). The Euclidean policy gradient of $\phi_{\text{erg.}}$ is

$$\begin{aligned} & \nabla_E\phi_{\text{erg.}}\left(q, z\right)_{s,a} \\ = \; & \sum_{s'\in\mathbf{S}} \left(q\left((s, a), \{s'\}\right) - \frac{1}{|\mathbf{A}|}\sum_{a'\in\mathbf{A}} q\left((s, a'), \{s'\}\right)\right) \mathcal{D}_{\text{erg.}} f\left(K_{q,z}\right)\left(s, \{s'\}\right) \end{aligned} \tag{3.29}$$

and the Fisher policy gradient is

$$\begin{aligned} & \nabla_F\phi_{\text{erg}}\left(q, z\right)_{s,a} \\ = \; & z\left(s, \{a\}\right)\sum_{s'\in\mathbf{S}} \left(q\left((s, a), \{s'\}\right) - K_{q,z}\left(s, \{s'\}\right)\right) \mathcal{D}_{\text{erg.}} f\left(K_{q,z}\right)_{s,s'} \end{aligned} \tag{3.30}$$

where $\mathcal{D}_{\text{erg.}}$ has been defined in Eq. 3.28.

Now we will prove a fundamental relationship between ”discounted sensor process functionals” and ergodic functionals which is one of the main results of the current section.

Theorem 3.2.1 - (**Relation between discounted and ergodic functionals**)

Let $(\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ be an MDP and assume that the state space, $\mathbf{S}$ is finite. Let $\Delta_{\mathbf{S}}$ be the simplex over $\mathbf{S}$ and let $\Delta_{\epsilon;\mathbf{S}}$ be the ϵ-simplex over $\mathbf{S}$ (compare Eq. 2.7). Let

$$h : \Delta_{\mathbf{S}} \times {\Delta_{\epsilon;\mathbf{S}}}^{\mathbf{S}} \to \mathbb{R} \tag{3.31}$$

be a continuous sensor process functional and let $(n_k)_{k\in\mathbb{N}_0}$ be an arbitrary sequence of natural numbers with $\lim_{k\to\infty} n_k = \infty$. Define the discounted value of h along $(n_k)_{k\in\mathbb{N}_0}$:

$$\begin{aligned} g : [0, 1) \times \Delta_{\mathbf{S}} \times \Delta_{\epsilon;\mathbf{S}}^{\mathbf{S}} \; & \to \; \mathbb{R} \\ (\lambda, \mu, K) \; & \mapsto \; \sum_{k\in\mathbb{N}_0} \lambda^k h\left(\mu K^{n_k}, K\right) \end{aligned} \tag{3.32}$$

where we wrote K^{n_k} for the n_k-fold convolution of K with itself. Define the ergodic value of h :

$$\begin{aligned} g_{\text{erg}} : \Delta_{\epsilon;\mathbf{S}}^{\mathbf{S}} &\to \mathbb{R} \\ K &\mapsto h\left(\text{INV}\left(K\right), K\right) \end{aligned} \tag{3.33}$$

where $\text{INV}(K)$ *denotes the unique invariant distribution of* K *again. Then*

$$\lim_{\lambda \nearrow 1} \left(1-\lambda\right) g\left(\lambda, \mu, K\right) = g_{\text{erg}}\left(K\right) \tag{3.34}$$

for any $\mu \in \Delta_{\mathbf{S}}$ *and* $K \in {\Delta_{\epsilon;\mathbf{S}}}^{\mathbf{S}}$. *Moreover the set valued map*

$$\begin{aligned} G : [0,1] &\to 2^{{\Delta_{\epsilon;\mathbf{S}}}^{\mathbf{S}}} \\ \lambda &\mapsto \begin{cases} \text{argmax}_{K \in {\Delta_{\epsilon;\mathbf{S}}}^{\mathbf{S}}} \left\{ g\left(\lambda, \mu, K\right) \right\} & \text{for } \lambda < 1 \\ \text{argmax}_{K \in {\Delta_{\epsilon;\mathbf{S}}}^{\mathbf{S}}} \left\{ g_{\text{erg}}\left(K\right) \right\} & \text{else} \end{cases} \end{aligned} \tag{3.35}$$

is upper semicontinuous (compare Definition 2.2.2). This especially implies that for any given sequence $\left(\lambda_n\right)_{n\in\mathbb{N}} \in [0,1)^{\mathbb{N}}$ *with* $\lim_{n\to\infty} \lambda_n = 1$ *and any selection* $K_{*,n} \in G(\lambda_n)$, *the limit points of the sequence* $\left(K_{*,n}\right)_{n\in\mathbb{N}}$ *are maximizers of* g_{erg}.

Proof. Fix $\delta > 0$ and set $H := \max\left\{ \left|h\left(\mu, K\right)\right| \middle| \mu \in \Delta_{\mathbf{S}}, K \in {\Delta_{\epsilon;\mathbf{S}}}^{\mathbf{S}} \right\}$. If $H = 0$ then $h \equiv 0$ and the statement is trivial. If $H \neq 0$ by the ergodic theorem for finite state space Markov chains (compare Appendix A.1.4): $\lim_{n\to\infty} \mu K^n = \text{INV}(K)$ and the limit is uniformly in $K \in {\Delta_{\epsilon;\mathbf{S}}}^{\mathbf{S}}$ and $\mu \in \Delta_{\mathbf{S}}$. Therefore by compactness of the domain and continuity of h there exists some $N \in \mathbb{N}$ such that

$$\left|h\left(\mu K^{n_k}, K\right) - h\left(\text{INV}(K), K\right)\right| < \frac{\delta}{2} \text{ for all } K \in {\Delta_{\epsilon;\mathbf{S}}}^{\mathbf{S}} \text{ and all } k \geq N \tag{3.36}$$

Now whenever $1 > \lambda > 1 - \delta/(2(N+1)H)$ then:

$$\begin{aligned} & (1-\lambda) \sum_{k\in\mathbb{N}_0} \lambda^k h\left(\mu K^{n_k}, K\right) \\ = \; & (1-\lambda) \sum_{k=0}^{N} \lambda^k h\left(\mu K^{n_k}, K\right) + (1-\lambda) \sum_{k=N+1}^{\infty} \lambda^k h\left(\mu K^{n_k}, K\right) \\ \leq \; & \frac{\delta}{2} + (1-\lambda) \sum_{k=N+1}^{\infty} \lambda^k \left[h\left(\text{INV}(K), K\right) + \left|h\left(\mu K^{n_k}, K\right) - h\left(\text{INV}(K), K\right)\right| \right] \\ \leq \; & \lambda^N h\left(\text{INV}(K), K\right) + \delta \end{aligned} \tag{3.37}$$

and similarly

$$\begin{aligned} & (1-\lambda) \sum_{k\in\mathbb{N}_0} \lambda^k h\left(\mu K^{n_k}, K\right) \\ = \; & (1-\lambda) \sum_{k=0}^{N} \lambda^k h\left(\mu K^{n_k}, K\right) + (1-\lambda) \sum_{k=N+1}^{\infty} \lambda^k h\left(\mu K^{n_k}, K\right) \\ \geq \; & -\frac{\delta}{2} + (1-\lambda) \sum_{k=N+1}^{\infty} \lambda^k \left[h\left(\text{INV}(K), K\right) - \left|h\left(\mu K^{n_k}, K\right) - h\left(\text{INV}(K), K\right)\right| \right] \\ \geq \; & \lambda^N h\left(\text{INV}(K), K\right) - \delta \end{aligned} \tag{3.38}$$

In total

$$\limsup_{\lambda \nearrow 1} \left(1-\lambda\right) g\left(\lambda, \mu, K\right) \leq g_{\text{erg}}\left(K\right) + \delta \tag{3.39}$$

and

$$\liminf_{\lambda\nearrow 1}\left(1-\lambda\right)g\left(\lambda,\mu,K\right)\geq g_{\text{erg}}\left(K\right)-\delta \tag{3.40}$$

Since δ was arbitrary this implies:

$$\lim_{\lambda\nearrow 1} g\left(\lambda,\mu,K\right)=g_{\text{erg}}\left(K\right) \tag{3.41}$$

As a consequence the function:

$$\gamma\left(\lambda,K\right):=\begin{cases}\left(1-\lambda\right)g\left(\lambda,\mu,K\right) & \text{for } \lambda<1\\ g_{\text{erg}}\left(K\right) & \text{else}\end{cases} \tag{3.42}$$

is continuous on $[0,1]\times\Delta_{\epsilon;\mathbf{S}}$. Therefore the set-valued map

$$\lambda\mapsto\operatorname{argmax}_{K\in\Delta_{\epsilon,\mathbf{S}}^{\mathbf{S}}}\gamma(\lambda,K) \tag{3.43}$$

is upper semi-continuous by the maximum theorem (compare Theorem 2.2.1). But the function $\left(1-\lambda\right)g\left(\lambda,\mu,K\right)$ differs from $g\left(\lambda,\mu,K\right)$ by a positive factor only and therefore possesses the same maximizers. This proves the second statement of the theorem. ■

In the next section we will describe some interesting instances of Problem 3.2.1, Problem 3.2.2 and Problem 3.2.3.

3.3 Examples of policy functionals

3.3.1 The expected reward

Let $(\mathbf{Q},\mathbf{Z},\mathbf{S},\mathbf{A},\mathbf{X},p_x,T,\Pi)$ be an MDP. Consider the discounted reward of the sensor process $(S_k)_{k\in\mathbb{N}}$ [2]:

$$R:=\sum_{k\in\mathbb{N}_0}\lambda^k r(S_k) \tag{3.44}$$

where $0<\lambda<1$ is a discount factor and $r:\mathbf{S}\to\mathbb{R}$ is the reward function (we assume r to be bounded and measurable).

As often done in the theory of Markov processes, the transition kernel $K\in\Lambda_{\mathbf{S}}^{\mathbf{S}}$ can be considered as a linear operator on the set of signed measures of finite total variation on $\mathcal{F}_{\mathbf{S}}$ via (it is convenient to write the action of this operator as a left-action):

$$\left(\nu K\right)(A):=\int_{\mathbf{S}}\nu(ds)K(s,A)\text{ for every } A\in\mathcal{F}_{\mathbf{S}} \tag{3.45}$$

and as a linear operator on the set of bounded measurable functions on $\mathbf{S}$ via

$$K:\mathcal{B}_b(\mathbf{S})\to\mathcal{B}_b(\mathbf{S})\,;\,Kf(s):=\int_{\mathbf{S}}K(s,dt)f(t) \tag{3.46}$$

Moreover the vector space of signed, measures of finite total variation and the vector space of bounded measurable functions are Banach spaces under the total variation norm and the supremum norm respectively. On each of these two Banach spaces the corresponding linear operator associated to K is continuous with operator norm 1.

Applying the n-fold product of this operator to a signed measure (or a bounded measurable function) is the same as applying the n-fold convolution of K to this signed measure (measurable function), where the n-fold convolution is defined as usual:

$$K^0(s,A)=\delta_s(A)\,;\,K^{n+1}(s,A):=\int_{\mathbf{S}}K^n(s,dt)K(t,A) \tag{3.47}$$

---

[2] For a precise definition of the sensor process, see Definition 3.1.1 and Remark 3.1.1

Then $E_\mu\left[r\left(S_n\right)\right] = \mu\left(K^n r\right) = \left(\mu K^n\right) r$ and by the summation formula for the geometric series:

$$\begin{aligned} f_{\text{rew};r,\lambda}(\mu, K) := E\left[R\right] = \sum_{n\in\mathbb{N}_0} \mu\left(\left(\lambda K\right)^n r\right) &= \mu\left(\left(\mathbb{1} - \lambda K\right)^{-1} r\right) \\ &= \left(\mu\left(\mathbb{1} - \lambda K\right)^{-1}\right) r \qquad (3.48) \end{aligned}$$

If the state space is finite then the initial measure, μ, is usually expressed as a row vector, the reward function, r, is expressed as a column vector, the kernel, K, becomes a stochastic matrix and the right-hand side of Eq. 3.48 can be interpreted as a matrix identity. In this case the matrix inversion can be calculated efficiently using the well-known Gauß-Jordan algorithm.

A special feature of the discounted reward problem is that the optimizers, K, do not depend on the initial measure μ. Define the value function

$$V(s) := \sup\left\{f_{\text{rew };r,\lambda}\left(\delta_s, K\right) \middle| K \in \Lambda^{\mathbf{S}}{}_{\mathbf{S}}\right\} \tag{3.49}$$

Then V satisfies the Bellman principle of optimality:

$$V(s) = r(s) + \lambda \sup\left\{\int_{\mathbf{S}} K\left(s, ds'\right) V(s') \middle| K \in \Lambda_{\mathbf{S}}^{\mathbf{S}}\right\} \tag{3.50}$$

On the other hand whenever a solution V to the Bellman equation is known, then any $K_* \in \Lambda^{\mathbf{S}}{}_{\mathbf{S}}$ satisfying

$$\int_{\mathbf{S}} K_*(s, ds')V(s') = \sup\left\{\int_{\mathbf{S}} K(s, ds')V(s') \middle| K \in \Lambda^{\mathbf{S}}{}_{\mathbf{S}}\right\} \text{ for all } s \in \mathbf{S} \tag{3.51}$$

is a solution to the original optimization problem. Since the Bellman equation does not depend on the initial measure, the same holds true for the optimizer. The Bellman equation is usually the starting point for a solution of the maximum-reward problem (a good reference for MDPs are Eugene and Feinberg [71], Dynkin and Yushkevich [68] or Bertsekas and Shreve [23] for example).

For a gradient-ascent based approach, note that the optimization problem is an instance of Problem 3.2.2 with the sensor process functional given by $f_{\text{rew};r,\lambda}$ from Eq. 3.48. The only reward specific quantity in the general gradient formulas, Remark 3.2.1, is the partial derivative of $f_{\text{rew};r,\lambda}$ with respect to the sensor kernel. This derivative can be calculated using Lemma 6.0.9 from the Appendix:

Remark 3.3.1 - (**Kernel derivative for expected discounted reward**)

Let $(\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a finite state space MDP (compare Assumption 4.1.1 and section 3.2.2). Then the partial derivative of 3.2.2 (see Eq. 3.48) with respect to the world kernel is:*

$$\begin{aligned} &\frac{\partial f_{\text{rew};r,\lambda}}{\partial K\left(s, \{s'\}\right)}\left(\mu, K\right) \qquad (3.52) \\ &= \lambda\left\{\sum_{s_1\in\mathbf{S}} \mu\left(\{s_1\}\right)\left(\mathbb{1} - \lambda K\right)^{-1}{}_{s_1,s}\right\}\left\{\sum_{s_1\in\mathbf{S}} \left(\mathbb{1} - \lambda K\right)^{-1}{}_{s',s_1} r\left(s_1\right)\right\} \end{aligned}$$

By Theorem 3.2.1 the discounted expected reward approximates the corresponding ergodic reward for discount factors close to 1. The process functional for the ergodic counterpart to the expected discounted reward problem is:

$$f_{\text{erg.rew };r}(\mu, K) := \int_{\mathbf{S}} r d\mu \tag{3.53}$$

By Remark 3.2.2 the only reward-specific ingredient in the policy gradient formula is what we called ergodic derivative of $f_{\text{erg.rew };r}$. The resulting expression is:

Remark 3.3.2 - (**Ergodic derivative for the expected reward**)

Let $(\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a finite state space MDP (compare Assumption 4.1.1 and section 3.2.2). Then the ergodic derivative of 3.53 (see Eq. 3.53) with respect to the world kernel is:*

$$\mathcal{D}_{\text{erg.}} f_{\text{erg.rew };r} (K)_{s,s'} = \mu(\{s\}) \sum_{s_1 \in \mathbf{S}} Y_K (s', s_1)\, r(s_1)$$

The expected reward is the simplest non-trivial example of a policy functional. As we already highlighted reward maximization has already attracted lots of attention in reinforcement learning. Even though our approach to the gradient formula is new, the final result coincides with a gradient formula that is well-known in robotics:

Remark 3.3.3 - (**Comment on the gradient formula for the expected reward**)

A formula for the gradient of the discounted expected reward and of the reward in the stationary distribution has been given in Sutton et al. [180] (and previously in Marbach and Tsitsiklis [124], Cao and H.F. [45] and Jaakkola, Singh, and Jordan [91]). The authors yield the result by an elegant computational trick. It is very illustrative to see that their formula coincides with ours. Let $q \in \mathbf{Q} = \Lambda^{\mathbf{S}}_{\mathbf{S}\times\mathbf{A}}$ *be a fixed world kernel and let* ρ *be the reward in the stationary distribution, i.e.*

$$\rho(z) = f_{\text{erg.rew.var};r}(\text{INV}(K_{q,z}), K_{q,z}) \tag{3.54}$$

Then the policy gradient from Sutton et al. [180] is:

$$\nabla_E \rho(a\,|s) = \text{INV}(K)(s) \left(Q^z(s,a) - \frac{1}{|\mathbf{A}|} \sum_{a' \in \mathbf{A}} Q^z(s,a) \right) \tag{3.55}$$

where $z \in \mathbf{Z}$ *and*

$$Q^z(s,a) := E_{z,a,s} \left[\sum_{n \in \mathbb{N}} (R_n - \rho(z)) \right] \tag{3.56}$$

where R_n *is the reward at time* n *and* $P_{z,a,s}$ *is the law of the fixed policy process* $(Z_n = z)$ *with initial states* $S_0 = s$ *and* $A_0 = a$*. Let* $r \in \mathbb{R}^{\mathbf{S}}$ *be the reward vector (i.e.* $R_n = r(S_n)$*), then*

$$Q^z(s,a) = \sum_{n \in \mathbb{N}} {p^{(s,a)}}^T (K_{q,z})^n r - \rho(z) \tag{3.57}$$

where

$$p^{(s,a)}(s') := q\left((s,a), \{s'\}\right) \tag{3.58}$$

Note that

$$\rho(z) = {p^{(s,a)}}^T ({K_{q,z}}^*)\, r \tag{3.59}$$

since ${K_{q,z}}^* := \lim_{n\to\infty} {K_{q,z}}^n$ *projects any distribution onto the stationary distribution of* $K_{q,z}$*. Furthermore note that for every* $n \geq 1$

$$(K_{q,z})^n = ({K_{q,z}}^* + K_{q,z} - {K_{q,z}}^*)^n = {K_{q,z}}^* + (K_{q,z} - {K_{q,z}}^*)^n \tag{3.60}$$

since $\left(K_{q,z}{}^{*}\right) K_{q,z} = K_{q,z}\left(K_{q,z}{}^{*}\right) = K_{q,z}{}^{*}$ *and* $\left(K_{q,z}{}^{*}\right)^2 = K_{q,z}{}^{*}$. *Inserting this into Eq. 3.57 and using the summation formula for geometric series gives:*

$$Q^z(s,a) = p^{(s,a)^T} Y_z r \tag{3.61}$$

where

$$Y_z = \left(\mathbb{1} - K_{q,z} + K_{q,z}{}^{*}\right)^{-1} - K_{q,z}{}^{*} \tag{3.62}$$

This shows that Eq. 3.55 is really identical to our formula, Eq. 3.29 and Remark 3.3.1. A similar analysis shows that our gradient of the discounted expected reward coincides with the gradient found in Sutton et al. [180]. Note however that our final gradient ascent algorithm needs an estimation of q, which is usually very straight forward (we will provide an estimator for q in the next chapter and prove its convergence), whereas the algorithm in the paper mentioned here need to estimate $z \mapsto Q^z$ *directly, which is much harder.*

This remark shows, that the gradient formula can be derived in a systematic way without any computational trick. This is essential to deal with policy functionals that do not originate from an expected reward function and cannot be calculated that easily. To give a first example of the latter class of functionals, consider a target functional corresponding to the variance of the discounted expected reward: Consider the variance of the discounted expected reward for some reward function, R, from Eq. 3.44:

$$\text{Var} := E_\mu\left[\left(R - E_\mu\left[R\right]\right)^2\right] \tag{3.63}$$

Transforming this into an analytic expression depending on K and μ is slightly more involved than the previous calculation for the expected reward, but follows the same line of reasoning. The final result is:

$$f_{\text{rew.var.};r,\lambda}(\mu, K) := \text{Var} = \tag{3.64}$$

$$\mu\left[\left(\mathbb{1} - \lambda^2 K\right)^{-1} r^2 + 2\lambda(\mathbb{1} - \lambda K)^{-1}\left[r \cdot K\left(\mathbb{1} - \lambda K\right)^{-1} r\right]\right] - \left(\mu\left(\mathbb{1} - \lambda K\right)^{-1} r\right)^2 \tag{3.65}$$

Having an explicit formula for the variance as a function of the transition kernel, it is easy to derive the policy gradient using Remark 3.2.1. To simplify the calculation of similar quanities, we included a section on derivatives of holomorphic matrix functions in the Appendix (see A.2.3 and the special case of polynomials of matrices: Lemma 6.0.9).

Next we will illustrated the gradient fields for the expected discounted reward graphically. Therefore we consider a finite state space MDP with two sensor values and two actions and use the following parameterization of the policies:

Assumption 3.3.1 - (**Parameterization of policies for graphical illustration of gradient formulas**)

Set $\mathbf{S} := \{s_1, s_2\}$, $\mathbf{A} := \{a_1, a_2\}$ *and fix a transition kernel* $q \in \Delta_{\mathbf{S}}{}^{\mathbf{S}\times\mathbf{A}}$. *Every policy* $z \in \Delta_{\mathbf{A}}{}^{\mathbf{S}}$ *can be parameterized by two real numbers* $p, q \in [0,1]$ *by defining* $z_{p,q}$ *via:*

$$z_{p,q}\left(s_1, \{a_1\}\right) := p\ ;\ z_{p,q}\left(s_2, \{a_2\}\right) := q \tag{3.66}$$

This automatically implies

$$z_{p,q}\left(s_1, \{a_2\}\right) = 1 - p \text{ and } z_{p,q}\left(s_2, \{a_1\}\right) = 1 - q \tag{3.67}$$

The gradient at a given point is always tangent to the affine subspace of matrices with rowsum 1 *and can therefore be described by two real parameters. This reduces the problem to a* $2D$ *problem and allows a graphical representation.*

The ergodic expected reward as a function of the policy parameters p and q is plotted in Figure 3.1 [3].

Theorem 3.2.1 can also be illustrated graphically. We use the following parameterization:

Assumption 3.3.2 - (Parameterization of sensor transition kernels for graphical illustration)

Set $\mathbf{S} := \{s_1, s_2\}$*, any kernel* $k \in \Delta_{\mathbf{S}}{}^{\mathbf{S}}$ *can be parameterized by two parameters:*

$$k\left(s_1, \{s_1\}\right) := p\ ;\ k\left(s_2, \{s_2\}\right) := q \tag{3.68}$$

This automatically implies

$$k\left(s_1, \{s_2\}\right) = 1 - p \text{ and } k\left(s_2, \{s_1\}\right) = 1 - q \tag{3.69}$$

We plotted the discounted expected reward for several discount factors and the ergodic reward. The results are shown in Figure 3.2 [4].

3.3.2 Extensions and modifications of the single sensor value expected reward problem

A generalization from rewards depending on one sensor value to rewards that depend on the outcome of two (or more generally n) successive sensor values is straight forward:

Consider the sliding discounted reward,

$$R_{r,1} := \sum_{k \in \mathbb{N}_0} \lambda^k r\left(S_k, S_{k+1}, \ldots, S_{k+n-1}\right), \tag{3.70}$$

and the block-wise discounted reward,

$$R_{r,2} := \sum_{k \in \mathbb{N}_0} \lambda^k r\left(S_{kn}, S_{kn+1}, \ldots, S_{(k+1)n-1}\right), \tag{3.71}$$

where $r : \mathbf{S}^n \to \mathbb{R}$ is $\otimes^n \mathcal{F}_S / \mathcal{B}_{\mathbb{R}}$ measurable and bounded. Define

$$\Pr_r : \Lambda^{\mathbf{S}}{}_{\mathbf{S}} \to \mathcal{B}_b(\mathbf{S}) \tag{3.72}$$

via

$$\Pr_r\left(K\right)\left(s\right) := \int_{\mathbf{S}^{n-1}} K(s, ds_2) K(s_2, ds_3) \cdots K(s_{n-1}, ds_n) r(s, s_2, \ldots, s_n) \tag{3.73}$$

Then the sensor process functional for the problems are:

$$\begin{aligned} & f_{\text{rew},1;r,\lambda}(\mu, K) := E\left[R_{r,1}\right] \\ = \ & \mu\left(\left(\mathbb{1} - K\right)^{-1} \Pr_r\left(K\right)\right) = \left(\mu\left(\mathbb{1} - K\right)^{-1}\right) \Pr_r\left(K\right) \end{aligned} \tag{3.74}$$

for the sliding expected reward and

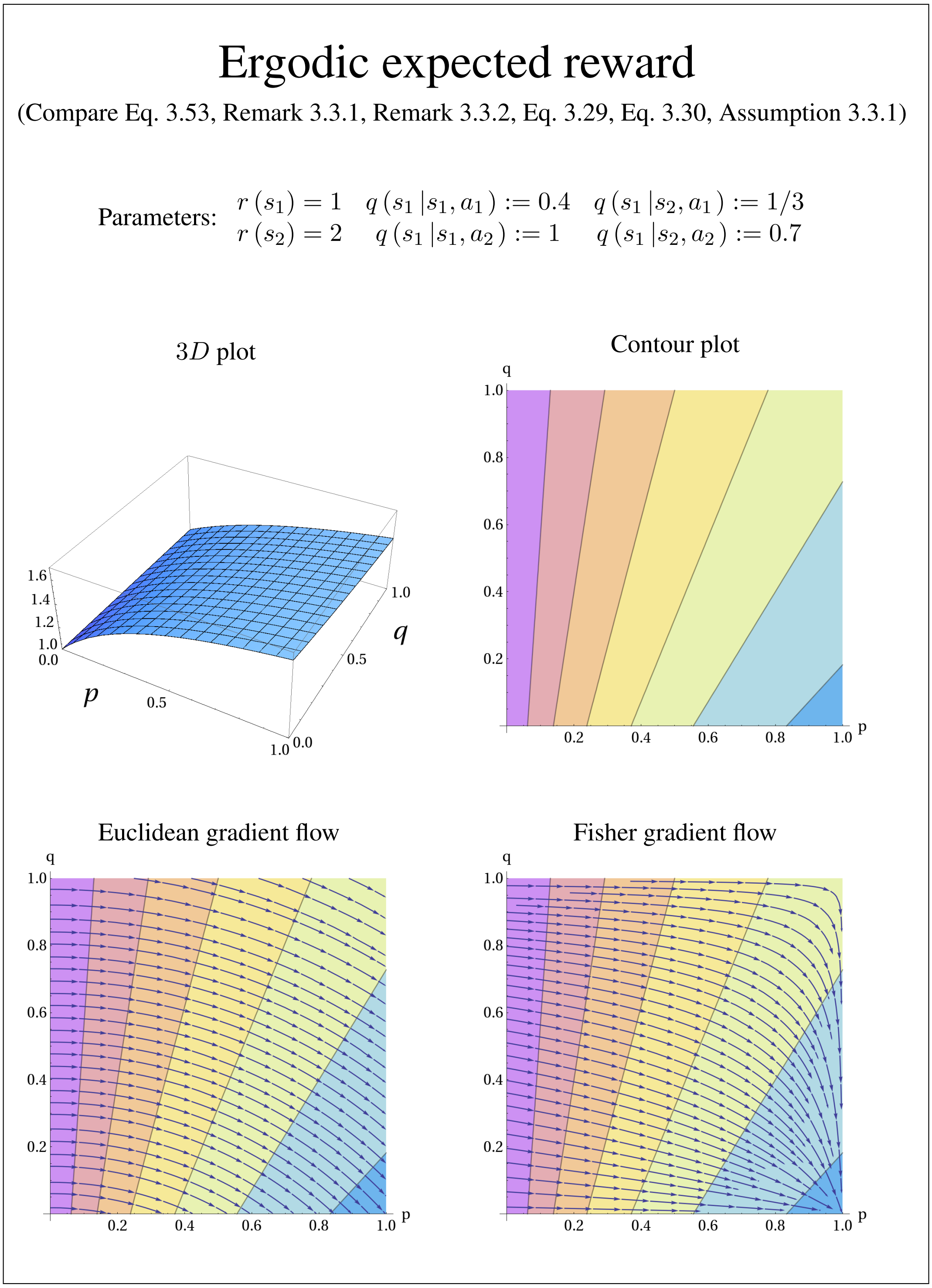
$$\begin{aligned} & f_{\text{rew},2;r,\lambda}(\mu, K) := E\left[R_{r,2}\right] \\ = \ & \mu\left(\left(\mathbb{1} - K^n\right)^{-1} \Pr_r\left(K\right)\right) = \left(\mu\left(\mathbb{1} - K^n\right)^{-1}\right) \Pr_r\left(K\right) \end{aligned} \tag{3.75}$$

for the block-wise discounted reward.

---

[3] The plots have been generated with Wolfram Mathematica 8.0.0.0

[4] The plots have again been generated with Wolfram Mathematica 8.0.0.0



Ergodic expected reward

(Compare Eq. 3.53, Remark 3.3.1, Remark 3.3.2, Eq. 3.29, Eq. 3.30, Assumption 3.3.1)

Parameters: $r(s_1) = 1 \quad q(s_1|s_1,a_1) := 0.4 \quad q(s_1|s_2,a_1) := 1/3$
$r(s_2) = 2 \quad q(s_1|s_1,a_2) := 1 \quad q(s_1|s_2,a_2) := 0.7$

$3D$ plot

Contour plot

Euclidean gradient flow

Fisher gradient flow

Figure 3.1: Ergodic expected reward

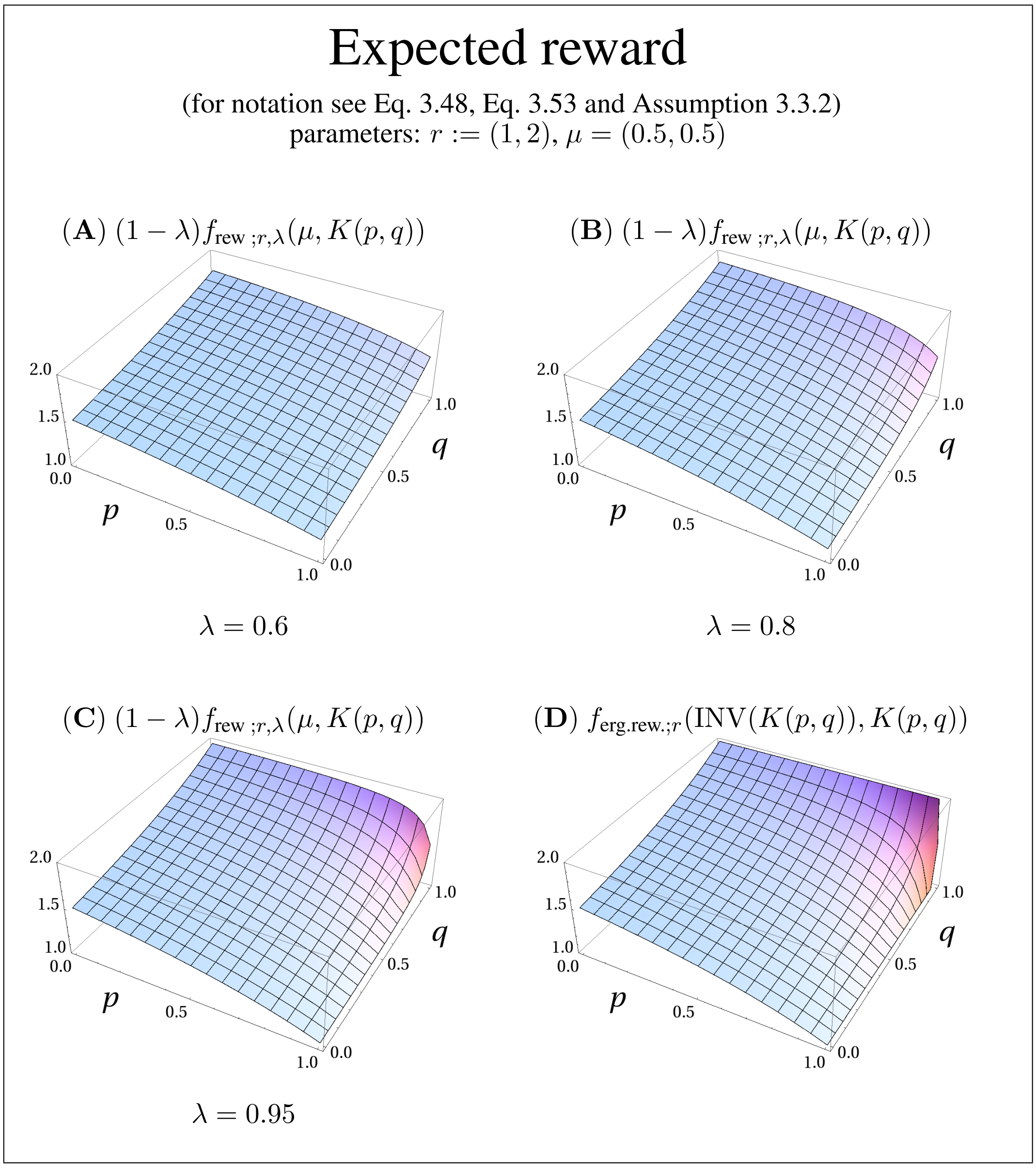


Figure 3.2: Discounted expected reward and expected ergodic reward

Remark 3.3.4 - (**Partial derivative for the multi-point reward**)

► **Remark 3.3.4.1:** *The world kernel derivative of* $f_{\text{rew},1;r,\lambda}$ *reads*

$$\begin{aligned} &\frac{\partial f_{\text{rew},1;r,\lambda}}{\partial K_{s,s'}}\left(\mu,K\right) \\ = \;& \lambda\left\{\sum_{s_1\in\mathbf{S}}\mu\left(\{s_1\}\right)\left(\mathbb{1}-\lambda K\right)^{-1}{}_{s,s_1}\right\}\left\{\sum_{s_1\in\mathbf{S}}\left(\mathbb{1}-\lambda K\right)^{-1}{}_{s',s_1}\operatorname{Pr}_r(K)\left(\{s_1\}\right)\right\}+ \\ &+\sum_{s_1,s_2\in\mathbf{S}}\mu\left(\{s_1\}\right)\left(\mathbb{1}-K\right)^{-1}{}_{s_1,s_2}\sum_{j=1}^{n-1}\tilde{r}_j\left(s\right)\left(s,s'\right) \end{aligned} \tag{3.76}$$

where

$$\begin{aligned} &\tilde{r}_j\left(\tilde{s}\right)\left(s,s'\right) \\ := \;& \sum_{s_1,s_2,\dots,s_n\in\mathbf{S}}\delta_{s_1,\tilde{s}}\delta_{s_j,s}\delta s_{j+1},s'\frac{K\left(s_1,\{s_2\}\right)\cdot\ldots\cdot K\left(s_{n-1},\{s_n\}\right)}{K\left(s_j,\{s_{j+1}\}\right)}r\left(s_1,s_2,\dots,s_n\right) \end{aligned} \tag{3.77}$$

► **Remark 3.3.4.2:** *The ergodic counterpart of Eq 3.74 and Eq. 3.75 is the sensor process functional*

$$f_{\text{erg.rew};r'}(\mu,K) := \int_{\mathbf{S}^n} d\mu\left(ds_1\right)K\left(s_1,ds_s\right)\dots K\left(s_{n-1},ds_n\right)r'\left(s_1,s_2,\dots,s_n\right) \tag{3.78}$$

where now $r':\mathbf{S}^n\to\mathbb{R}$ *is bounded, measurable. The ergodic derivative of* $f_{\text{erg.rew};r'}$ *is*

$$\mathcal{D}_{\text{erg.}}f_{\text{erg.rew };r'}\left(K\right)_{s,s'} = A\left(s,s'\right)+B\left(s,s'\right) \tag{3.79}$$

where

$$A\left(s,s'\right) := \sum_{j=1}^{n-1}\delta_{s_j,s}\delta_{s_{j+1},s'}\frac{\mu\left(\{s_1\}\right)K\left(s_1,\{s_2\}\right)\dots K\left(s_{n-1},\{s_n\}\right)}{K\left(s_j,\{s_{j+1}\}\right)}r'\left(s_1,s_2,\dots,s_n\right)$$

and

$$B\left(s,s'\right) := \operatorname{INV}\left(K\right)\left(\{s\}\right)\sum_{\mathbf{s}\in\mathbf{S}^n}Y_K\left(s',\{s_1\}\right)K\left(s_1,\{s_2\}\right)\cdot\ldots\cdot K\left(s_{n-1},s_n\right)r'\left(s_1,s_2,\dots,s_n\right)$$

More generally it is possible to consider the state-action process for a fixed policy parameter and to take rewards that depend on state-action pairs or state-action-state triplets. However these generalizations are straight forward and we will not write down explicit expressions.

Another generalization is the inclusion of unknown reward functions or probabilistic rewards (by this we mean that the reward at time n depends causally on S_n, A_n and S_{n+1}, i.e. it is independent of the past given these values). This can be modeled easiest by modifying the sensor space into $\mathbf{S} := \mathbf{S}_1\times\mathbb{R}$, where $\mathbf{S}_1$ is a finite set, expressing the set of sensor value and the second factor corresponding to the received reward signal. The transition kernels should then be defined in an appropriate way. Even though this construction is not a finite state action space MDP, the derivation of policy gradients follows the same pattern. For a convergence proof the result in Section 4.2 of Chapter 4 can be employed.

3.3.3 Entropy of the sensor distribution

The expectation of a bounded measurable random variable can be considered as a linear functional of the underlying measure. Conceptually it is very natural to extend these linear functionals to non-linear ones. This is indeed a very fruitful approach. Interesting non-linear functionals include information measures or risk measures to name only two examples.

Consider a finite state space MDP with state space $\mathbf{S}$. The discounted one-point entropy (compare Definition 6.0.5) is generated by the sensor process functional:

$$f_{\text{entr};\lambda}(\mu, K) := \sum_{k\in\mathbb{N}_0} \lambda^k H\left(\mu K^n\right) \tag{3.80}$$

where

$$H : M_1\left(2^{\mathbf{S}}\right) \to \mathbb{R}_{\geq 0}; p \mapsto \sum_{s\in\mathbf{S}} p\left(\{s\}\right) \log_2 p\left(\{s\}\right) \tag{3.81}$$

is the entropy. For non-discrete state spaces, $(\mathbf{S}, \mathcal{F}_S)$, this functional naturally generalizes to:

$$f_{\text{diff.entr};\eta,\lambda}(\mu, K) := \sum_{n\in\mathbb{N}_0} \lambda^n H_\eta\left(\mu K^n\right) \tag{3.82}$$

where η is some σ-finite measure on $\mathcal{F}_{\mathbf{S}}$ and H_η is the generalized entropy (compare Definition 6.0.6). The ergodic functional corresponding to $f_{\text{entr};\lambda}$ is:

$$f_{\text{erg.entr}}(\mu, K) := H\left(\mu\right) \tag{3.83}$$

the resulting ergodic functional is the entropy in the stationary distribution. Again the general formulas from Remark 3.2.1 and Remark 3.2.2 can be used to calculate the policy gradient for these problems:

Remark 3.3.5 - (**World kernel derivative and ergodic derivative for the entropy**)

► **Remark 3.3.5.1:** *Let* $\mathbf{S}$ *be the state space of a finite state and action space MDP. Then the partial derivative of* $f_{\text{entr};\lambda}$ *with respect to the world kernel is*

$$\begin{aligned} &\frac{\partial f_{\text{entr};\lambda}}{\partial K\left(s, \{s'\}\right)}\left(\mu, K\right) \\ = \; & -\sum_{n\in\mathbb{N}_0} \lambda^n \sum_{j=0}^{n-1} \left(\mu K^j\right)\left(\{s\}\right) \sum_{\tilde{s}\in\mathbf{S}} K^{n-1-j}\left(s', \tilde{s}\right) \ln\left(\mu K^n\right)\left(\tilde{s}\right) + R_1\left(s\right) \end{aligned} \tag{3.84}$$

where $R_1(s)$ *is a summand that does not depend on* s'*. As a consequence the summation over this term in Eq. 3.21 and Eq. 3.22 vanishes. Therefore this summand can be ignored.*
► **Remark 3.3.5.2:** *The ergodic derivative of* $f_{\text{erg.entr}}$ *is:*

$$\mathcal{D}_{\text{erg.}} f_{\text{erg.entr}}\left(K\right)_{s,s'} = \mu(\{s\}) \sum_{s_1\in\mathbf{S}} Y_K\left(s', s_1\right) \log\left(\text{INV}\left(K\right)\left(\{s_1\}\right)\right) + R_1(s)$$

where again $R_1(s)$ *is a summand that does not depend on* s'*. As a consequence the summation over this term vanishes in Eq. 3.29 and Eq. 3.30. Therefore this summand can be ignored.*

Unlike the expected reward, the partial derivative of $f_{\text{entr};\lambda}$ with respect to the world kernel is an infinite series, that cannot be simplified in a straight forward way. From a computational prospective the infinite sum is not a problem, since the series decays exponentially such that a finite cutoff at a sufficiently high number of terms yields a convenient approximation. It is also very easy to provide an *a priori* bound for the cutoff error. The ergodic counterpart on the other hand gives rise to a simple explicit expression, that can be calculated efficiently.

3.3.4 Discounted mutual information and predictive information

Consider a finite state space MDP with sensor space $(\mathbf{S}, \mathcal{F}_S)$. An interesting information measure is the sliding discounted mutual information between two neighboring sensor values. The corresponding sensor process functional is

$$f_{\text{M.I.},1;\lambda}(\mu, K) := \sum_{n \in \mathbb{N}_0} \lambda^n I_{\mu K^n \otimes K}(\pi_1, \pi_2) \tag{3.85}$$

a rather similar quantity is the block-wise discounted mutual information of two successive sensor values, with sensor functional

$$f_{\text{M.I.},2;\lambda}(\mu, K) := \sum_{n \in \mathbb{N}_0} \lambda^n I_{\mu K^{2n} \otimes K}(\pi_1, \pi_2) \tag{3.86}$$

where $I_p(\pi_1, \pi_2)$ denotes the mutual information (compare Definition 6.0.6) and $p \otimes K \in M_1(\mathcal{F}_S \otimes \mathcal{F}_S)$ is the unique measure satisfying:

$$(p \otimes K)(A \times B) := \int_A p(ds) K(s, B) \text{ for every } A, B \in \mathcal{F}_S \tag{3.87}$$

Since the process $(S_n)_{n \in \mathbb{N}_0}$ is Markovian (for fixed policy) the mutual information between two successive sensor values, S_n and S_{n+1}, is equal to the mutual information between all predecessors of S_n including S_n and all successors of S_{n+1} including S_{n+1}, i.e.

$$I(S_n, S_{n+1}) = I\left((S_k)_{0 \leq k \leq n}, (S_k)_{k > n}\right). \tag{3.88}$$

This is a consequence of Appendix Lemma 6.0.12. [5]. The quantity on the right-hand side of Eq. 3.88 is also known as predictive information (compare Grassberger [78], Grassberger [77], Bialek, Nemenman, and Tishby [27], Bialek and Tishby [28] and Crutchfield and Feldman [56]). For MDPs, controlled by some learning algorithm Eq. 3.88 is not true anymore - indeed in principle it is possible to save all former sensor values in the memory and to influence the sensor process in a way that allows the reconstruction of the entire past from the future process - in which case the predictive information for the sensor process is equal to the entropy of the entire past process. Therefore the following simple consequence from the graph separation-independence property, Theorem 1.3.1, is remarkable:

Theorem 3.3.1 - (Bounds on the predictive information for controlled MDPs with finite state and memory space)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be an MDP with finite state space,* $\mathbf{S}$ *and let* $M = (\mathbf{M}, L, U)$ *be a learning algorithm over* C *with finite memory space,* $\mathbf{M}$. *Then*

$$I\left[(S_i)_{0 \leq k \leq n}, (S_i)_{k > n}\right] \leq \ln(|\mathbf{S}|) + \ln(|\mathbf{M}|) \tag{3.89}$$

for every $n \in \mathbb{N}$ *under any measure* $P_{q', s', m'}$ *constructed in Remark 3.1.2.*

Proof. By Appendix Lemma 6.0.12 and the Markov property of the process $((S_i, M_i))_{i \in \mathbb{N}_0}$ we have:

$$\begin{aligned} I\left[(S_i)_{0 \leq k \leq n}, (S_i)_{k > n}\right] &\leq I\left[(S_i, M_i)_{0 \leq k \leq n}, (S_i, M_i)_{k > n}\right] \\ &= I\left[(S_n, M_n), (S_{n+1}, M_{n+1})\right] \end{aligned} \tag{3.90}$$

---

[5] In the appendix we defined the mutual information using the KL divergence. Another possibility for finite state spaces is a definition via entropies. However in this case the right-hand side of Eq. 3.88 would be ill-defined except for trivial cases

For any pair of random variables, (X, Y) where X has values in $\mathbf{X}$ and Y has values in $\mathbf{Y}$:

$$I(X,Y) \leq \max\left\{\ln\left(|\mathbf{X}|\right), \ln\left(|\mathbf{Y}|\right)\right\}$$

such that Eq. 3.90 implies the validity of the estimate. ■

Remark 3.3.6 - (**Remark on Theorem 3.3.1**)

With some further effort the estimate can be improved further for a specific learning algorithm. Consider the kernel, $\tilde{K}_{q'} \in \Lambda^{\mathbf{S}\times\mathbf{M}}_{\mathbf{S}\times\mathbf{M}}$*, from Eq. 3.7. For a given measure* $p \in M_1\left(2^{\mathbf{S}\times\mathbf{M}}\right)$ *define the measure* $p \otimes \tilde{K}_{q'} \in M_1\left(2^{(\mathbf{S}\times\mathbf{M})\times(\mathbf{S}\times\mathbf{M})}\right)$ *via:*

$$p \otimes \tilde{K}_{q'}\left(\left\{\left((s,m),(s',m')\right)\right\}\right) := p\left(\{(s,m)\}\right) K_{q'}\left[(s,m), \left\{\left(s',m'\right)\right\}\right] \tag{3.91}$$

Let π_1 *and* π_2 *denote the projections of* $(\mathbf{S}\times\mathbf{M})\times(\mathbf{S}\times\mathbf{M})$ *onto the first and second factor. Then a tighter bound following from Eq. 3.90 is:*

$$I\left[(S_i)_{0\leq k\leq n}, (S_i)_{k>n}\right] \leq \sup\left\{I_{p\otimes K_{q'}}\left(\pi_1,\pi_2\right) \left|p \in M_1\left(2^{\mathbf{S}\times\mathbf{M}}\right)\right.\right\}, \tag{3.92}$$

Which is the channel capacity of a channel with transition kernel $K_{q'}$ *(this is essentially one direction of Shannon's noisy-channel coding theorem, compare Shannon [170]).*

Again in the flavor of Theorem 3.2.1 the discounted sliding mutual information and the block wise mutual information can be considered as approximation of an ergodic functional for discount factors closed to 1. The sensor process functional corresponding to this ergodic functional is:

$$f_{\text{P.I.}}(\mu, K) := I(\mu \otimes K). \tag{3.93}$$

The mutual information in the stationary distribution is also known as predictive information in a narrow sense. The optimization of this quantity has attracted lots of attention in the past (compare Ay et al. [17], Zahedi, Ay, and Der [196] and Ay et al. [16]). The convergence proof of a stochastic gradient algorithm maximizing this quantity was one of the main motivations for this thesis. The calculation of policy gradients of the discounted mutual information and its ergodic counterpart, the predictive information in a narrow sense is very similar to the calculations for the discounted reward and for the entropy. The general gradient formulas Remark 3.2.1 and Remark 3.2.2 can be used again and the partial derivative (and the ergodic derivative) of the corresponding sensor process functionals have to be calculated explicitly.

Remark 3.3.7 - (Partial derivative of the discounted mutual information with respect to world kernel and ergodic derivative of the PI sensor process functional)

Let $(\mathbf{S}, \mathcal{F}_S)$ *be the state space of a finite stat and action space MDP.*

► **Remark 3.3.7.1:** *The partial derivative of* $f_{\text{M.I.},1;\lambda}$ *with respect to the world kernel is*

$$\begin{aligned}
&\frac{\partial f_{\text{M.I.},1;\lambda}}{\partial K(s,\{s'\})}(\mu, K) \qquad (3.94)\\
= &\sum_{n\in\mathbb{N}_0} \lambda^n \Bigg[\mu K^n(\{s\}) \ln\left(\frac{K(s,\{s'\})}{\mu K^{n+1}(\{s'\})}\right) + \\
&+ \sum_{s_1\in\mathbf{S}}\left(\sum_{s_2\in\mathbf{S}} K(s_1,\{s_2\}) \ln\left(\frac{K(s_1,\{s_2\})}{\mu K^{n+1}(\{s_2\})}\right)\right)\left(\sum_{j=0}^{n-1} \mu K^j(\{s\}) K^{n-1-j}(s',\{s_1\})\right)\Bigg]
\end{aligned}$$

where we did not write down s'*-independent summands since they cancel in the summation (compare formulas for entropy).*

► **Remark 3.3.7.2:** *The ergodic derivative of* $f_{\text{P.I.}}$ *from Eq. 3.93 is:*

$$\begin{aligned}
&\mathcal{D}_{\text{erg.}} f_{\text{P.I.}}(K)_{s,s'} = \text{INV}(K)(\{s\}) \cdot \qquad (3.95)\\
&\cdot\left[\log\left(\frac{K(s,\{s'\})}{\text{INV}(K)(\{s'\})}\right) + \sum_{s_1,s_2\in\mathbf{S}} Y_K(s',s_1) K(s_1,\{s_2\}) \log\left(\frac{K(s_1,\{s_2\})}{\text{INV}(K)(\{s_2\})}\right)\right]
\end{aligned}$$

where we omitted s'*-independent summands since they cancel in the summation (compare formulas for entropy).*

Figure 3.3 shows an example of the predictive information as a function of the policy (we use the parameterization suggested in Assumption 3.3.1 again). In Figure 3.4 we illustrate Theorem 3.2.1 for the predictive information. We used the parameterization from Assumption 3.3.2 again.

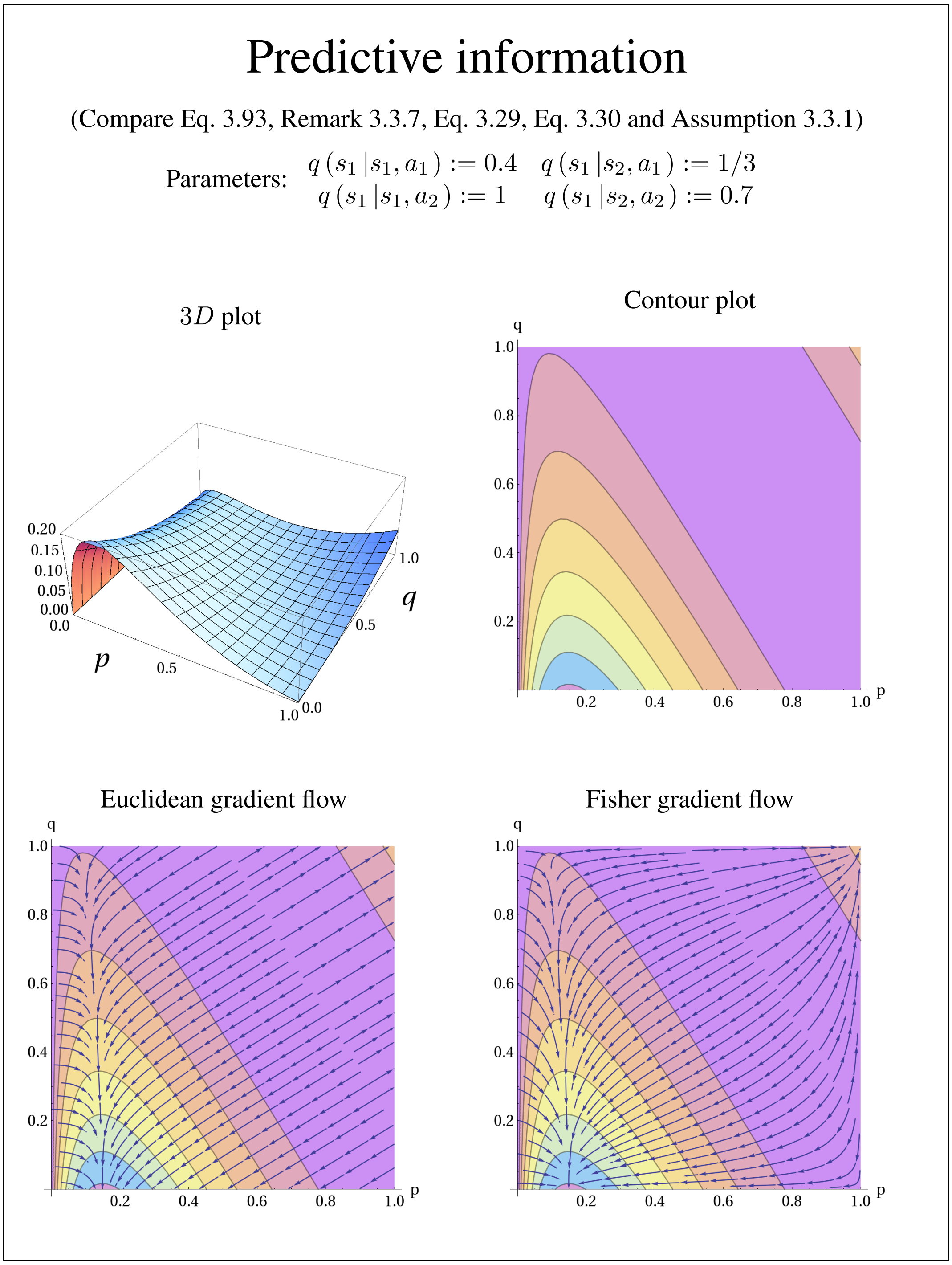


Figure 3.3: Predictive information

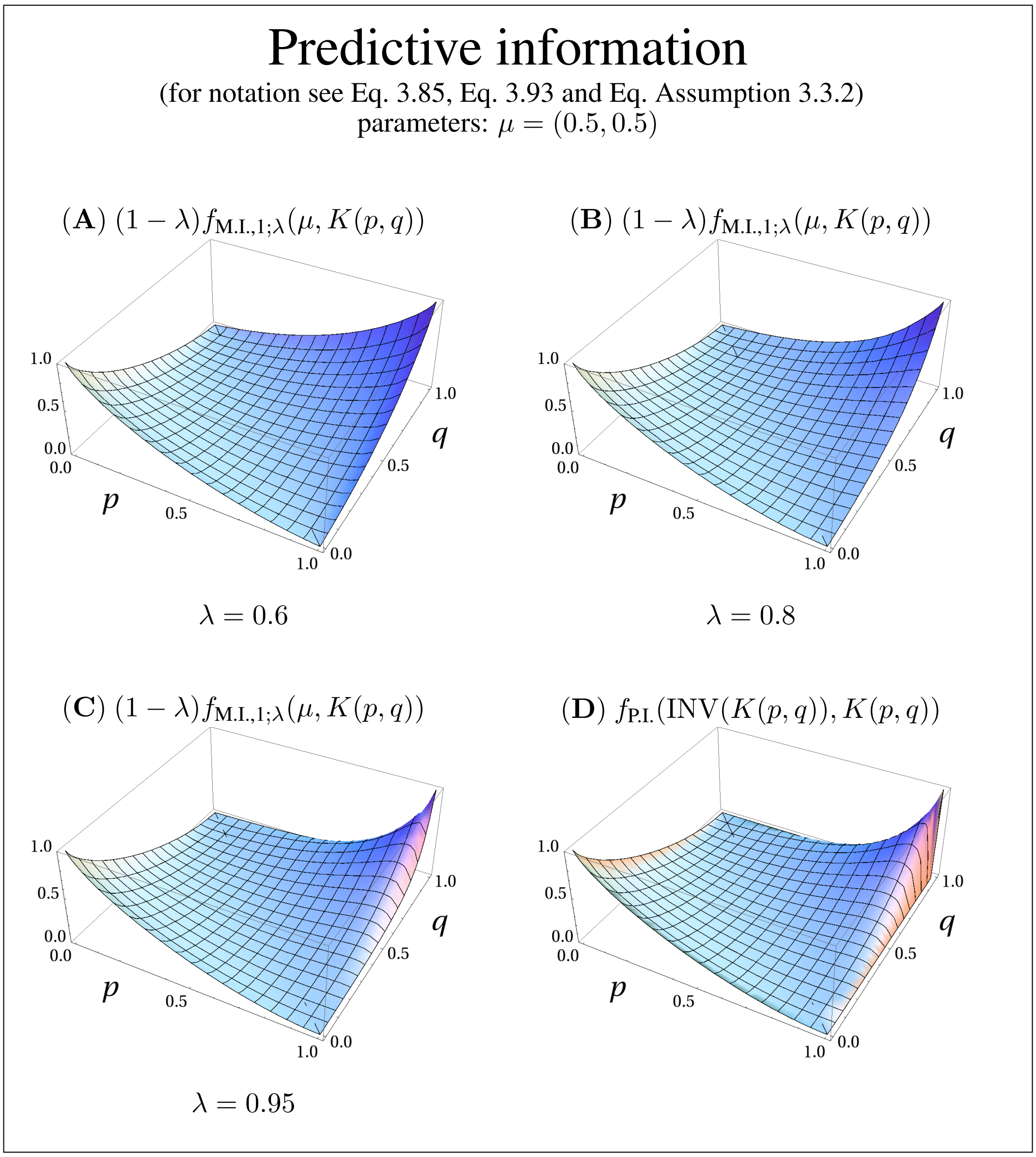


Figure 3.4: Discounted mutual information and predictive information

Chapter 4

Convergence proofs for learning algorithms in the sensorimotor loop

In this chapter we propose a projected stochastic gradient algorithm to solve Problem 3.2.1 and its special instances Problem 3.2.2 and Problem 3.2.3. We start with finite state and action spaces and prove convergence of the algorithm (see Theorem 4.1.1). In this context a special instance of our algorithm is an improvement of the algorithm suggested in Zahedi, Ay, and Der [196] for a maximization of the predictive information. Therefore as a partial result we provide a rigorous convergence proof for this algorithm for the first time.

After this we consider more general MDPs and prove convergence whenever an appropriate Q-estimator exists (see Theorem 4.2.1). Finally we apply this theorem to a linear-Gaussian MDP, for which we also construct a Q-estimator with the desired properties. The resulting algorithm extends results from Ay et al. [16] to a scenario where the system parameters are previously unknown and have to be learnt (see Theorem 4.2.1).

To ease concrete implementations we provide a list of gradients for specific sensor process functionals and ergodic functionals in Appendix A.2.

4.1 Optimization of policy functionals for finite state space MDPs

4.1.1 Model assumptions

In this section we consider a Markov decision process (compare Definition 3.1.1),

$$C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$$

with finite state space and finite action space. The aim is to approach a solution to Problem 3.2.1 for some policy functional $\phi : \mathbf{Q} \times \mathbf{Z} \to \mathbb{R}$. We assume that an explicit expression of ϕ is known but that no analytic expression for the maximum of $M_{\phi,q} := \max \{ I(q,z) \,|z \in \mathbf{Z} \}$ is available.

Ideally $\mathbf{Z}$ and $\mathbf{Q}$ contain all stochastic transition matrices, but we need to impose some mild restrictions on $\mathbf{Q}$ and $\mathbf{Z}$. The algorithm to be stated in the end of this section uses an estimate of the sensor transition kernel T. For consistency of this estimator it is necessary that all state action pairs $(s,a) \in \mathbf{S} \times \mathbf{A}$ are visited infinitely often. Therefore it is reasonable to consider the following parameter set for the world transition:

$$\mathbf{Q} := \left\{ k \in \Lambda^{\mathbf{S}}_{\mathbf{S}\times\mathbf{A}} \,\middle|\, k((s,a),\{s'\}) > 0 \text{ for all } s, s' \in \mathbf{S} \text{ and } a \in \mathbf{A} \right\} \tag{4.1}$$

The relation between the parameter $q \in \mathbf{Q}$ and the kernel T is the following:

$$T\left((k,s,a),A\right) := k\left((s,a),A\right) \text{ for every } a \in \mathbf{A}, s \in \mathbf{S} \tag{4.2}$$

For the set of policies, $\mathbf{Z}$, we fix some parameter $\frac{1}{|\mathbf{A}|} > \epsilon > 0$ and require the policies to be sufficiently mixing, more precisely:

$$\mathbf{Z} = \mathbf{Z}_\epsilon := \left\{ k \in \Lambda_{\mathbf{S}}^{\mathbf{A}} \, |k(s,\{a\}) \geq \epsilon \text{ for all } s \in \mathbf{S} \text{ and } a \in \mathbf{A} \right\} \tag{4.3}$$

We require the policy transition function, Π, and the noise distribution, p_x, to be compatible with the parameter $z \in \mathbf{Z}$, in the following sense:

$$\int_{\mathbf{X}} \Pi\left((s,x,z),A\right) dp_x(x) = z\left(s,A\right) \tag{4.4}$$

One possibility to generate any desired probability distribution on a finite set $\{1,2,\ldots,m\}$ using a random variable, X, with uniform distribution on the unit interval, is to split the unit interval into m disjoint intervals I_i of length $p(\{i\})$ and to decide for i if $X \in I_i$. Therefore we can choose $\mathbf{X} = [0,1]$, $p_x = \nu_{\text{Leb}}$ where ν_{Leb} is the Lebesgue measure. Assume $\mathbf{S} = \{s_1, s_2, \cdots, s_n\}$ and $\mathbf{A} = \{a_1, a_2, \cdots, a_m\}$. Then a canonical choice for Π is the following one:

$$\Pi\left((s,x,z),\{a_i\}\right) = \begin{cases} 1 & \text{if } \sum_{1 \leq k < i} z(s,\{a_k\}) < x \leq \sum_{k \leq i} z(s,\{a_k\}) \\ 0 & \text{else} \end{cases} \tag{4.5}$$

It is clear that this definitions ensure the consistency conditions Eq. 4.4.

Here is a summary of our assumptions on C:

Assumption 4.1.1 - (**Underlying spaces and transition functions**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a Markov decision process.*
► **Assumption 4.1.1.1:** $\mathbf{S} = \{s_1, s_2, \cdots, s_n\}$ *and* $\mathbf{A} = \{a_1, a_2, \cdots, a_m\}$ *are finite sets.*
► **Assumption 4.1.1.2:** $\mathbf{X} = [0,1]$ *and* $p_x = \nu_{\text{Leb}}$.
► **Assumption 4.1.1.3:** $\mathbf{Q}$ *and* $\mathbf{Z}$ *are given by Eq. 4.1 and by Eq. 4.3.*
► **Assumption 4.1.1.4:** *The transition kernels are given by Eq. 4.2 and by Eq. 4.5.*

We will use the same notation as in Chapter 3. The finite state space MDP,

$$C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi),$$

naturally defines a causal model C' (compare Definition 3.1.1). Equivalently every learning algorithm, $M := (\mathbf{M}, L, U)$, over C gives rise to a causal model M' (which we will call sensorimotor loop) and any collection of initial values $q' \in \mathbf{Q}$, $s' \in \mathbf{S}$ and $m' \in \mathbf{M}$ induces a probability measure, $P_{q',s',m'} \in M_1\left(\otimes_{v \in V} \mathcal{F}_v\right)$ where (V,E) is the MDP graph, Caus. mod. 19. We will write $E_{q',s',m'}$ for the expectation with respect to $P_{q',s',m'}$. In the next subsection we address the problem of estimating the sensor kernel parameter $q' \in \mathbf{Q}$.

4.1.2 Estimating the sensor kernel parameter, $q' \in \mathbf{Q}$ for finite state space MDPs

We will frequently need the following random times:

Definition 4.1.1 - (**Recurrence times for state-action pairs**)

For $a \in A$ and $s \in S$ define:

$$T_{a,s,0} := 0 \; ; \; T_{a,s,1} := \inf \{ n \geq 0 \, | S_n = s, A_n = a \}$$

and inductively for $i \geq 1$

$$T_{a,s,i+1} := \inf \{ n > T_{a,s,i} \, | S_n = s, A_n = a \}$$

Define the filtration, $\mathbb{F}_n = \sigma \left(\{Q, S_i, M_i, X_i\}_{0 \leq i \leq n} \right)$ of $\mathcal{F}$. Then the processes $(S_n)_{n \in \mathbb{N}_0}$, $(A_n)_{n \in \mathbb{N}_0}$, $(X_n)_{n \in \mathbb{N}_0}$, $(M_n)_{n \in \mathbb{N}_0}$ and $(Z_n)_{n \in \mathbb{N}}$ are $\mathbb{F}$-adapted. Consequently the random times, $T_{a,s,i}$, are $\mathbb{F}$-stopping times.
Our restrictions on $\mathbf{Q}$ and $\mathbf{Z}$ (see Assumption 4.1.1) imply the following bounds:

Lemma 4.1.1 - (**Bounds on moments of $T_{a,s,i}$**)

Let $C := (\mathbf{Q}, \mathbf{Z}_\epsilon, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ be a MDP satisfying Assumption 4.1.1. Let $M = (\mathbf{M}, L, U)$ be an arbitrary learning algorithm over C (compare Definition 3.1.2). Fix $q' \in \mathbf{Q}$ and set $d(q') := \inf \{ q'((s,a), \{s'\}) \, | s, s' \in \mathbf{S}, a \in \mathbf{A} \}$.
► **Lemma 4.1.1.1:** *We have*

$$E_{q',s_0,m_0} [T_{a,s,n}] \leq \frac{n}{\epsilon d(q')}$$

for every initial states $s_0 \in \mathbf{S}$ and initial memory values $m_0 \in \mathbf{M}$.
► **Lemma 4.1.1.2:** *The random variable $T_{a,s,n}$ possesses a moment generating function in the neighborhood of 0, moreover the following estimate holds:*

$$E_{q',s_0,m_0} [\exp(\beta T_{a,s,n})] \leq \phi(q', \beta, \epsilon)^n \text{ whenever } \beta < -\ln(1 - \epsilon d(q'))$$

where

$$\phi(q', \beta, \epsilon) = \frac{\exp(\beta)}{1 - \exp(\beta + \ln(1 - \epsilon d(q')))}$$

Proof. For this proof we fix $s_0 \in \mathbf{S}$, $m_0 \in \mathbf{M}$, $q' \in \mathbf{Q}$ and write P instead P_{q',s_0,m_0} and E instead of E_{q',s_0,m_0}.
We will prove the first statement, $E[T_{a,s,n}] \leq \frac{n}{\epsilon d(q')}$, by induction over n.
For $n = 0$ the statement is obviously true by definition of $T_{a,s,0}$.
Assume that $E[T_{a,s,n}] \leq \frac{n}{\epsilon d(q')}$. This implies $P[\{T_{a,s,n} = \infty\}] = 0$. Therefore the discrete random sets (compare Definition 1.4.1)

$$\tau_j := \left\{ s_{T_{a,s,n}+j}, a_{T_{a,s,n}+j} \right\} \qquad ; j \geq 0 \tag{4.6}$$

are well-defined for almost all $\omega \in \Omega$. We also define a random set, I_j relative to τ_j (compare Definition 1.4.2) as follows

$$I_j(\{s_k, a_k\}) := \mathrm{An}(\{s_k, a_k\}) \cup \{s_k, a_k\} \tag{4.7}$$

Obviously $I_j \subseteq I_{j+1}$ for every $j \geq 0$ such that $\mathbb{G}_n := \mathcal{F}_{I_n}$ (where $\mathcal{F}_{I_n}$ denotes the inference $\sigma-$algebra of I_n, compare Definition 1.4.3) is a filtration of $\mathcal{F}$. By Assumption 4.1.1 on the parameter spaces, Theorem 1.4.1 and Corrolary 1.4.1 we have:

$$P\left[\left\{ S_{T_{a,s,n}+j+1} = s' \right\} | \mathbb{G}_j \right] = q'((S_{T_{a,s,n}+j}, A_{T_{a,s,n}+j}), \{s'\}) \geq d(q')$$

almost surely. By Assumption 4.1.1 and the strong Markov property again:

$$\begin{aligned} & P\left(\left\{A_{T_{s,a,n}+j+1}=a\right\}\left|\mathbb{G}_j, S_{T_{s,a,n}+j+1}, Z_{T_{s,a,n}+j+1},\right.\right) \\ = \; & Z_{T_{s,a,n}+j+1}\left(S_{T_{s,a,n}+j+1},\{a\}\right) \\ \geq \; & \epsilon \qquad\qquad \text{a.s.} \end{aligned} \tag{4.8}$$

By the tower property of the conditional expectation, these two equations yield:

$$\begin{aligned} & P\left[\left\{A_{T_{s,a,n}+j+1}=a, S_{T_{s,a,n}+j+1}=s\right\}|\mathbb{G}_j\right] \\ = \; & E\left[\mathbb{1}_{\left\{S_{T_{s,a,n}+j+1}=s\right\}} E\left[\mathbb{1}_{\left\{A_{T_{s,a,n}+j+1}=a\right\}}\left|\mathbb{G}_j, S_{T_{s,a,n}+j+1}, Z_{T_{s,a,n}+j+1}\right.\right] |\mathbb{G}_j\right] \\ = \; & E\left[\mathbb{1}_{\left\{S_{T_{s,a,n}+j+1}=s\right\}} Z_{T_{s,a,n}+j+1}(S_{T_{s,a,n}+j+1},\{a\})\,|\mathbb{G}_j\right] \\ \geq \; & \epsilon d(q') \qquad\qquad \text{a.s.} \end{aligned} \tag{4.9}$$

We will prove by induction that

$$\begin{aligned} & P\left(\{T_{a,s,n+1}-T_{a,s,n}\geq k\}\,\middle|\,(T_{a,s,m})_{0\leq m\leq n}\right) \\ \leq \; & P\left(\cap_{1\leq j<k}\left\{\left(S_{T_{a,s,n}+j}, A_{T_{a,s,n}+j}\right)\neq(s,a)\right\}\,\middle|\,(T_{a,s,m})_{0\leq m\leq n}\right) \\ \leq \; & (1-\epsilon d(q'))^{k-1} \qquad\qquad \text{a.s.,} \end{aligned} \tag{4.10}$$

where the less-than sign in the second line is actually an equality whenever $n\geq 1$.
For $k=1$ the statement is trivial. The induction step from $k-1$ to k for $k\geq 2$ follows from Eq. 4.9 and the fact that $S_{T_{a,s,n}+j}$, $A_{T_{a,s,n}+j}$ and $T_{a,s,n}$ are $\mathbb{G}_j$ measurable for every $j\geq 0$:

$$\begin{aligned} & P\left[\cap_{1\leq j<k}\left\{(S_{T_{a,s,n}+j}, A_{T_{a,s,n}+j})\neq(s,a)\right\}\,\middle|\,(T_{a,s,m})_{0\leq m\leq n}\right] \\ = \; & E\left[\mathbb{1}_{\cap_{1\leq j<k-1}\left\{(S_{T_{a,s,n}+j}, A_{T_{a,s,n}+j})\neq(s,a)\right\}}\right. \\ & \left.\left(1-E\left[\mathbb{1}_{\left\{(S_{T_{a,s,n}+k}, A_{T_{a,s,n}+k})=(s,a)\right\}}|\mathbb{G}_{k-1}\right]\right)\,\middle|\,(T_{a,s,m})_{0\leq m\leq n}\right] \\ \leq \; & (1-\epsilon d(q'))^{k-2}(1-\epsilon d(q'))=(1-\epsilon d(q'))^{k-1} \end{aligned} \tag{4.11}$$

This settles the induction step and proves Eq. 4.10.
For any $\mathbb{N}_0\cup\{\infty\}$ −valued random variable T the expectation value can be calculated from the following identity:

$$E[T]=\sum_{k\in\mathbb{N}} P(\{T\geq k\})$$

Hence by Eq. 4.10 and the well-known limit of a geometric series:

$$E[T_{a,s,n+1}-T_{a,s,n}]\leq\frac{1}{\epsilon d(q')} \tag{4.12}$$

Therefore the inductive assumption yields:

$$E[T_{a,s,n}]\leq\frac{n}{\epsilon d(q')} \tag{4.13}$$

For the existence of the characteristic function and the bound in Lemma 4.1.1 consider the following estimate (using the bound Eq. 4.10):

$$\begin{aligned} & E\left[\exp(\beta(T_{a,s,n+1}-T_{a,s,n}))\,|T_{a,s,i}=t_i \text{ for } 1\leq i\leq n\right] \\ = \; & \sum_{k\in\mathbb{N}}\exp(\beta k)P\left[\{T_{a,s,n+1}-T_{a,s,n}=k\}\,|T_{a,s,i}=t_i \text{ for } 1\leq i\leq n\right] \\ \leq \; & \sum_{k\in\mathbb{N}}\exp(\beta k)P\left[\{T_{a,s,n+1}-T_{a,s,n}\geq k\}\,|T_{a,s,i}=t_i \text{ for } 1\leq i\leq n\right] \\ \leq \; & \sum_{k\in\mathbb{N}}\exp(\beta k)\left(1-\epsilon d(q')\right)^{k-1} \\ \leq \; & \exp(\beta)\sum_{k\in\mathbb{N}_0}\exp(\beta+\ln(1-\epsilon d(q')))^k=\phi(q',\beta,\epsilon) \end{aligned}$$

Therefore:

$$E\left[\exp(\beta T_{a,s,n+1})\,\middle|\,\{T_{a,s,i}\}_{1\leq i\leq n}\right] \leq \phi(q',\beta,\epsilon)\exp(\beta T_{a,s,n}) \tag{4.14}$$

So by induction over n:

$$E\left[\exp(\beta T_{a,s,n})\right] \leq \phi(q',\beta,\epsilon)^n \tag{4.15}$$

■

An immediate consequence of Lemma 4.1.1 is that every $T_{a,s,n}$ is almost surely finite and possesses finite moments of arbitrary high order. Therefore every state-action pair is visited infinitely often such that the agent can estimate the sensor kernel parameter q' asymptotically exact.

The next lemma addresses the speed of convergence for a kernel estimator. Whenever a certain state-action pair (s,a) is observed, the distribution of the next sensor value does not depend on the history and is given by the distribution $q'((s,a),\cdot)$. In other words for fixed $s\in\mathbf{S}$ and $a\in\mathbf{A}$ the random variables

$$\tilde{S}_{a,s,n} := S_{T_{a,s,n}+1} \qquad n\geq 1$$

are independent, identically distributed with distribution $q'((s,a),\cdot)$. This insight and the result Lemma 4.1.1 can be used to prove the following result on the speed of convergence of the maximum likelyhood estimator of q':

Lemma 4.1.2 - (**Convergence of the Maximum likelihood estimator**)

Let $C := (\mathbf{Q},\mathbf{Z}_\epsilon,\mathbf{S},\mathbf{A},\mathbf{X},p_x,T,\Pi)$ be a MDP satisfying Assumption 4.1.1. Let $M = (\mathbf{M},L,U)$ be an arbitrary learning algorithm over C (compare Definition 3.1.2). Define $\mathbb{F}_n := \sigma\left(\{S_i,A_i\}_{0\leq i\leq n}\right)$ and define the ($\mathbb{F}-$adapted) maximum likelihood estimator:

$$\hat{q}_n^{\text{(MLE)}}((s,a),\{s'\}) := \begin{cases} \frac{1}{|\mathbf{S}|} & \text{if } \sum_{0\leq j<n}\mathbb{1}_{\{S_j=s,A_j=a\}}=0 \\ \frac{\sum_{0\leq j<n}\mathbb{1}_{\left\{S_j=s,A_j=a,S_{j+1}=s'\right\}}}{\sum_{0\leq j<n}\mathbb{1}_{\left\{S_j=s,A_j=a\right\}}} & \text{else} \end{cases} \tag{4.16}$$

Then for every $\alpha>\frac{1}{2}$:

$$\sum_{n\in\mathbb{N}} n^{-\alpha}\left\|\hat{q}_n-q'\right\|_1 < \infty \text{ a.s. w.r.t. } P_{q',s_0,m_0}$$

for every $q'\in\mathbf{Q}$, $s_0\in\mathbf{S}$ and $m_0\in\mathbf{M}$

Proof. Again we will fix $q'\in\mathbf{Q}$, $s_0\in\mathbf{S}$ and $m_0\in\mathbf{M}$ and we will write P instead of P_{q',s_0,m_0} and E instead of E_{q',s_0,m_0}.

Fix some $\delta>0$ such that

$$\alpha > 0.5+\delta. \tag{4.17}$$

Let $d(q') := \inf\left\{q'((s,a),\{s'\})\,|s,s'\in\mathbf{S},a\in\mathbf{A}\right\}$ again. Fix $0<\beta<-\ln(1-\epsilon d(q'))$ and note that by Markov's inequality (compare Appendix Lemma 6.0.6) and Lemma 4.1.1 for any $C>0$:

$$P\left(\{T_{a,s,n}\geq Cn\}\right) = P\left[\{\exp(\beta T_{a,s,n})\geq\exp(\beta Cn)\}\right] \leq \frac{\phi(q',\beta,\epsilon)^n}{\exp(\beta Cn)}$$

Fix some $C>\frac{\ln(\phi(q',\beta,\epsilon))}{\beta}$. Then the Borel-Cantelli lemma (compare Appendix Lemma 6.0.5) implies that there exists some set $N_0\in\mathcal{F}$ such that $P(N_0)=0$ and for any $\omega\in\Omega\setminus N_0$ the inequality $T_{a,s,n}(\omega)<Cn$ holds for all but finitely many n. Consequently for for all $\omega\in\Omega\setminus N_0$ there exists some constant $C_{0,\omega}\geq 1$ such that

$$T_{a,s,n}(\omega) \leq C_{0,\omega}n \text{ for all } n\in\mathbb{N} \tag{4.18}$$

Moreover the estimator $\hat{q}_n^{(\mathrm{MLE})}$ satisfies the following identity:

$$\tilde{q}_{a,s,k} := \hat{q}^{(\mathrm{MLE})}_{T_{a,s,k}+1}((s,a),\{s'\}) = \frac{\sum_{1\leq j\leq k} \mathbb{1}_{\{\tilde{S}_{a,s,j}=s'\}}}{k}$$

where again

$$\tilde{S}_{a,s,k} := S_{T_{a,s,k}+1}.$$

The random variables $\mathbb{1}_{\{\tilde{S}_{a,s,n}=s'\}}$ are i.i.d. and bounded (for fixed $a \in \mathbf{A}$ and $s \in \mathbf{S}$) with expectation value $q'((s,a),s')$. Hence theorem 4.23 of Kallenberg [99] (alternatively one could use the maybe better-known but technically more involved law of the iterated logarithm) yields:

$$\lim_{k\to\infty} k^{0.5-\delta} \left|\tilde{q}_{a,s,k}((s,a),\{s'\}) - q'((s,a),\{s'\})\right| = 0 \text{ a.s.} \tag{4.19}$$

In other words there exists some set $N_1 \in \mathcal{F}$ of measure zero, such that for every $\omega \in \Omega\backslash N_1$ there exists some constant $C_{1,\omega}$ such that:

$$\left|\tilde{q}_{a,s,k}((s,a),\{s'\}) - q'((s,a),\{s'\})\right| \leq C_{1,\omega} k^{-0.5+\delta} \text{ for every } s \in \mathbf{S}, a \in \mathbf{A}, k \in \mathbb{N} \tag{4.20}$$

Consider the (random) sum

$$I := \sum_{n\in\mathbb{N}} n^{-\alpha} \left\|\hat{q}_n - q'\right\|_1 = \sum_{s,s'\in\mathbf{S},a\in\mathbf{A}} \sum_{n\in\mathbb{N}} n^{-\alpha} \left|\hat{q}_n((s,a),\{s'\}) - q'((s,a),\{s'\})\right|$$

Fix $\omega \in \Omega \setminus (N_0 \cup N_1)$ and constants $C_{0,\omega}$ and $C_{1,\omega}$ such that 4.18 and 4.20 hold.
Define $f(a,s,n)(\omega) := \sup\{k \in \mathbb{N}\,|T_{a,s,k}(\omega) < n\}$.
Then $\hat{q}_n((s,a),\{s'\})(\omega) = \tilde{q}_{f(a,s,n)(\omega)}((s,a),\{s'\})(\omega)$ and by definition of N_1 and $C_{1,\omega}$:

$$I(\omega) \leq C_{1,\omega} \sum_{s,s'\in\mathbf{S},a\in\mathbf{A}} \sum_{n\in\mathbb{N}} n^{-\alpha}(f(a,s,n)(\omega))^{-0.5+\delta}$$

By definition of $C_{0,\omega}$

$$T_{a,s,\left\lfloor \frac{n}{2C_{0,\omega}}\right\rfloor}(\omega) \leq C_{0,\omega} \left\lfloor \frac{n}{2C_{0,\omega}} \right\rfloor \lesssim n$$

the sign "$\lesssim$" indicates that the last inequality is true whenever n is large enough - in this case $n \geq 2C_{0,\omega}$ is sufficient. Hence $f(a,s,n)(\omega) \gtrsim n/(2C_{0,\omega})$. Consequently there exists some constant $C_{3,\omega} > 0$ such that:

$$I(\omega) \leq C_{3,\omega} \sum_{n\in\mathbb{N}} n^{-\alpha-0.5+\delta} < \infty, \tag{4.21}$$

where the last identity follows from the choice of δ. This statement is true for any $\omega \in \Omega \setminus (N_0 \cup N_1)$ and $N_0 \cup N_1$ is a P-null set. Hence the statement is true. ■

4.1.3 A projected stochastic gradient algorithm for finite state space MDPs with non-linear objectives

In this section we formulate an algorithm to approach Problem 3.2.1. In order to apply a stochastic gradient ascent algorithm we will impose further restrictions on the objective function. We canonically identify $\mathbf{Q}$ with $(\Delta_{\mathbf{S}})^{\mathbf{S}\times\mathbf{A}^\circ}$ and $\mathbf{Z}_\epsilon$ with $(\Delta_{\epsilon,\mathbf{A}})^{\mathbf{S}}$ (compare Appendix A.1.4 and section 2.1). We require the following regularity assumptions to be satisfied by the function ϕ in Problem 3.2.1:

Assumption 4.1.2 - (**Regularity of objective function**)

Let $\phi : (\Delta_{\mathbf{S}})^{\mathbf{S}\times\mathbf{A}^\circ} \times (\Delta_{\mathbf{A}})^{\mathbf{S}^\circ} \to \mathbb{R}$. *We further assume:*

► **Assumption 4.1.2.1:** *For every* $q \in (\Delta_{\mathbf{S}})^{\mathbf{S}\times\mathbf{A}^\circ}$ *the function* $\phi_q : z \mapsto \phi(q, z)$ *is continuously differentiable. We further assume that the function* ϕ_q *restricted to* $(\Delta_{\epsilon,\mathbf{A}})^{\mathbf{S}}$ *has a negligible set of critical values for all* $\epsilon > 0$ *(compare Definition 2.4.3).*
► **Assumption 4.1.2.2:** *The function* $q \mapsto D_2\phi(q, z)$ *is continuous uniformly in* $z \in (\Delta_{\epsilon,\mathbf{A}})^{\mathbf{S}}$ *for every* $\epsilon > 0$. *By this we mean that for fixed* $q \in \mathbf{Q}$ *and* $\epsilon > 0$ *and any* $\epsilon' > 0$ *there exist* $\delta > 0$ *such that:*

$$\sup\left\{\left|D_2\phi(q', z) - D_2\phi(q, z)\right| \,\middle|\, z \in (\Delta_{\epsilon,\mathbf{A}})^{\mathbf{S}}\right\} < \epsilon' \text{ whenever } \left\|q' - q\right\|_1 < \delta \qquad (4.22)$$

Consider the following algorithm:

Algorithm 4.1.1 - (**Gradient ascent on finite state space MDP**)

► **Free parameters:**

- *Two ascent decay parameter* $c \in \mathbb{R}_+$ *and* α, *satisfying* $\frac{1}{2} < \alpha \leq 1$.
- *A metric* g *on* $(\Delta_{\mathbf{A}})^{\mathbf{S}^\circ}$ *that is compatible with the orthogonal projection (compare Definition 2.3.1 and the subsequent discussion)*

► **Variables and initializations:**

- *State-action counter:* $n \in {\mathbb{N}_0}^{\mathbf{S}\times\mathbf{A}}$; $n_0(s, a) \leftarrow 0$ *for every* $(s, a) \in \mathbf{S} \times \mathbf{A}$
- *Estimator for sensor transition kernel:* $\hat{q} \in (\Delta_{\mathbf{S}})^{\mathbf{S}\times\mathbf{A}}$; $\hat{q}_0\left((s, a), \{s'\}\right) \leftarrow \frac{1}{|\mathbf{S}|}$ *for every* $s, s' \in \mathbf{S}$, $a \in \mathbf{A}$.
- *Current policy:* $\hat{z} \in (\Delta_{\mathbf{A}})^{\mathbf{S}}$; $\hat{z}_0\left(s, \{a\}\right) \leftarrow \frac{1}{|\mathbf{A}|}$
- *Step counter:* $t \in \mathbb{N}_0$; $t \leftarrow 0$

► **Algorithm:**

```
repeat this:
    $n(S_t, A_t) \leftarrow n(S_t, A_t) + 1$
    foreach $s' \in \mathbf{S}$ do
        $\hat{q}\left[(S_t, A_t), \{s'\}\right] \leftarrow \hat{q}\left[(S_t, A_t), \{s'\}\right] + \frac{1}{n(S_t,A_t)}\left(\mathbb{1}_{\{s'\}}(S_{t+1}) - \hat{q}\left[(S_t, A_t), \{s'\}\right]\right)$
    end
    $\hat{z} \leftarrow \mathrm{Pr}_Z\left[\hat{z} + \frac{c}{(t+1)^\alpha}\nabla_{g,2}\phi(\hat{q}, \hat{z})\right]$
    $t \leftarrow t + 1$
forever
```

Where $\mathrm{Pr}_{\mathbf{Z}}$ *denotes the projection onto* $\mathbf{Z} = (\Delta_{\epsilon,\mathbf{A}})^{\mathbf{S}}$ *(compare Algorithm 4.1.2).*

The projection $\mathrm{Pr}_{\mathbf{Z}}$ can be done efficiently using an adaption of the algorithm suggested in Michelot [128] for standard simplices to a product of $\epsilon-$simplices:

Algorithm 4.1.2 - (**Projection onto** $(\Delta_{\epsilon,\mathbf{A}})^{\mathbf{S}}$)

▶ **Input data:** $x \in \mathbb{R}^{\mathbf{S}\times\mathbf{A}}$
▶ **Variables:** $I, I' \in 2^{\mathbf{A}}$, $\sigma \in \mathbb{R}$
▶ **Algorithm:**

```
foreach $s \in \mathbf{S}$ do
    $I \leftarrow \mathbf{A}$
    repeat
        $I' \leftarrow \emptyset$
        $\sigma := \sum_{a\in I} x_{s,a}$
        foreach $a \in I$ do
            if $x_{s,a} \geq \frac{\sigma-1+\epsilon|\mathbf{A}|}{|I|}$ then
                $x_{s,a} \leftarrow x_{s,a} + \frac{1-\sigma+\epsilon(|I|-|\mathbf{A}|)}{|I|}$
            end
            else
                $x_{s,a} \leftarrow \epsilon$
                $I' \leftarrow I' \cup \{a\}$
            end
        end
        $I \leftarrow I \setminus I'$
    until $I' = \emptyset$
end
return $x$
```

After this preparation we formulate and prove one of the main theorems of this section:

Theorem 4.1.1 - (**Convergence of Algorithm 4.1.1**)

Consider the algorithm Algorithm 4.1.1 and set $K := (\Delta_{\epsilon,\mathbf{A}})^{\mathbf{S}}$*. Then the iterative sequence* $(\hat{z}_n)_{n\in\mathbb{N}}$ *converges to the set of first order optimal points of* ϕ_q*:*

$$S_q := \{z \in K \,|D_2\phi\,(q,z) \in N_K(z)\}\,, \tag{4.23}$$

almost surely with respect to P_{q,s_0,m_0} *for every* $q \in \mathbf{Q}$ *and* $s_0 \in \mathbf{S}$.

Proof. $\hat{q}$ is the maximum likelihood estimator from Lemma 4.1.2. Write

$$\nabla_{g,2}\phi(\hat{q}_{k+1}, \hat{z}_k) := \nabla_{g,2}\phi(q, \hat{z}_k) + \beta_k \tag{4.24}$$

with

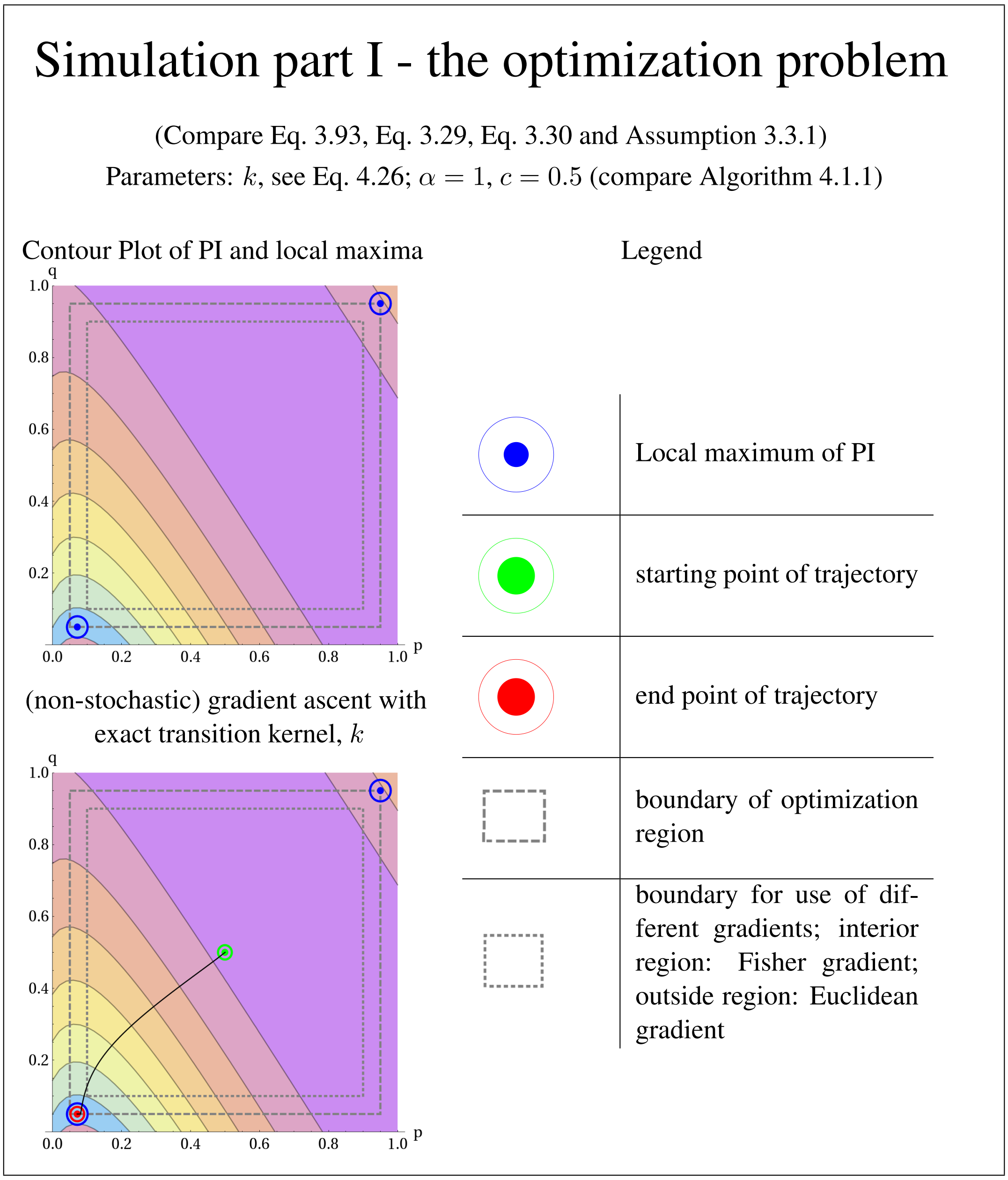
$$\beta_k = \nabla_{g,2}\phi(\hat{q}_{k+1}, \hat{z}_k) - \nabla_{g,2}\phi(q, \hat{z}_k) \tag{4.25}$$

By Lemma 4.1.2 and Assumption 4.1.2 $(\beta_k)_{k\in\mathbb{N}}$ converges to zero almost surely. Therefore the result follows immediately from Theorem 2.4.3. ■

4.1.4 Simulations

In this section we will illustrate the convergence result for the learning algorithms over a finite state spaces MDP, Theorem 4.1.1. We tested the statement for two sensor values and two actions, i.e. $\mathbf{S} = \{s_1, s_2\}$, $\mathbf{A} = \{a_1, a_2\}$ and the following world transition kernel:

$$\begin{array}{ll} k\left[(s_1,a_1),\{s_1\}\right] = 0.4 & k\left[(s_1,a_1),\{s_2\}\right] = 0.6 \\ k\left[(s_1,a_2),\{s_1\}\right] = 0.95 & k\left[(s_1,a_2),\{s_2\}\right] = 0.05 \\ k\left[(s_2,a_1),\{s_1\}\right] = \frac{1}{3} & k\left[(s_2,a_1),\{s_2\}\right] = \frac{2}{3} \\ k\left[(s_2,a_2),\{s_1\}\right] = 0.7 & k\left[(s_2,a_2),\{s_2\}\right] = 0.3 \end{array} \tag{4.26}$$



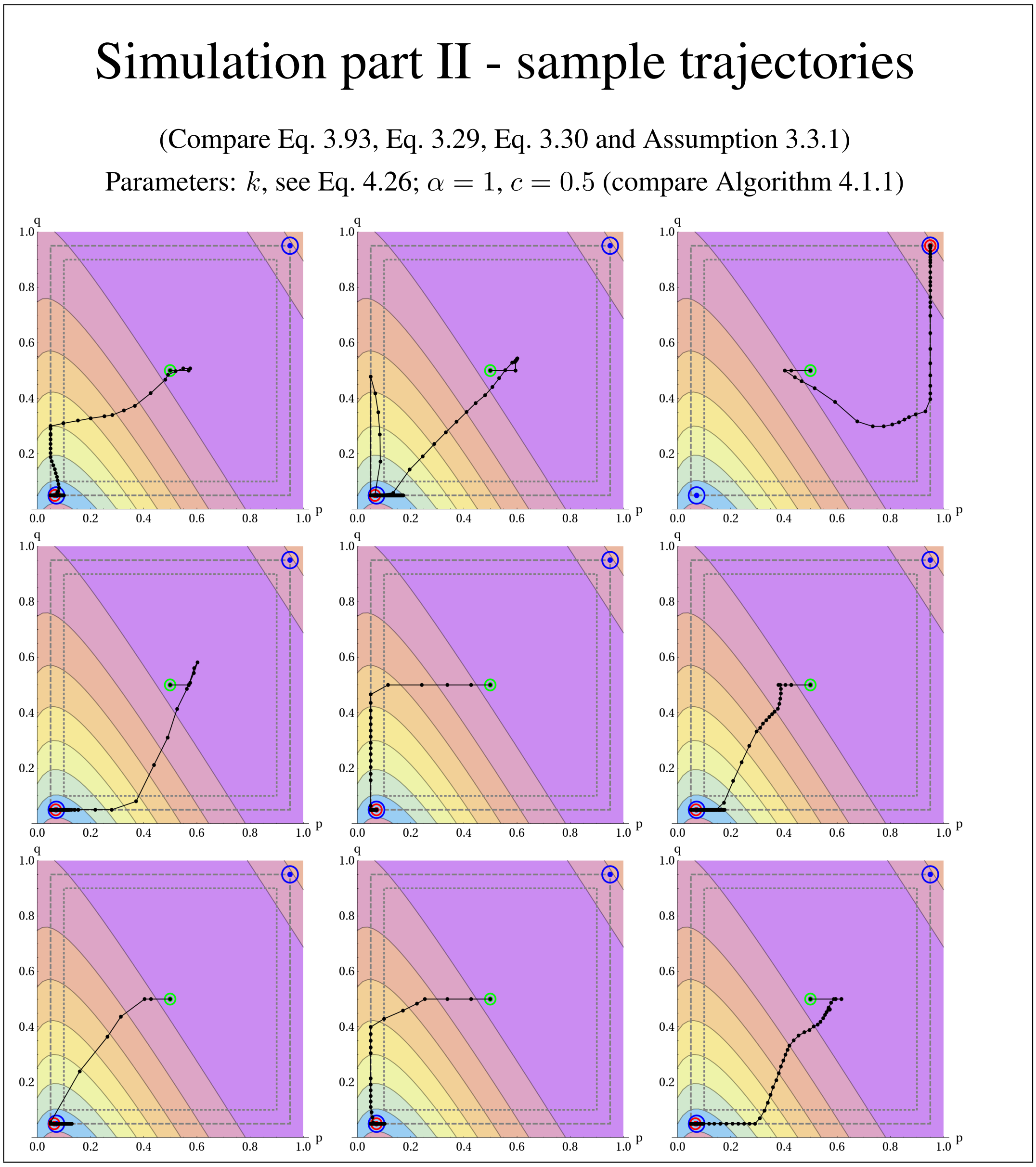
◉	Local maximum of PI
●	starting point of trajectory
●	end point of trajectory
⬚	boundary of optimization region
⬚	boundary for use of different gradients; interior region: Fisher gradient; outside region: Euclidean gradient

Figure 4.1: Simulation of PI maximization - part 1

Any policy kernel $z \in \Lambda_{\mathbf{S}}^{\mathbf{A}}$ can be parameterized by two numbers $p, q \in [0, 1]$:

$$\begin{aligned} z\left(s_1, \{a_1\}\right) &= p & z\left(s_1, \{a_2\}\right) &= 1 - p \\ z\left(s_2, \{a_1\}\right) &= 1 - q & z\left(s_2, \{a_2\}\right) &= q \end{aligned} \tag{4.27}$$

As target functional we choose the predictive information (compare Figure 3.4 and Figure 3.3). In our simulation we set $\epsilon := 0.05$ and perform a gradient ascent with the Fisher gradient whenever $z\left(s, \{a\}\right) \geq 0.1$ for every $a \in \mathbf{A}$ and $s \in \mathbf{S}$ and perform a gradient ascent with Euclidean gradient otherwise. This ensures compatibility with the othorgonal projection onto the ϵ-simplex (compare Definition 2.3.1 and Example 2.3.1). Figure 4.1 shows the level set of the predictive information as a function of the parameters p and q, the two local maximas and a gradient ascent with the true world kernel, k. In Figure 4.2 we show several trajectories. Here the kernel, k, is assumed to be unknown and is learnt by the agent. Last but not least we plot the PI and the squared error of the world kernel estimator for two sample trajectories (see Figure 4.3.

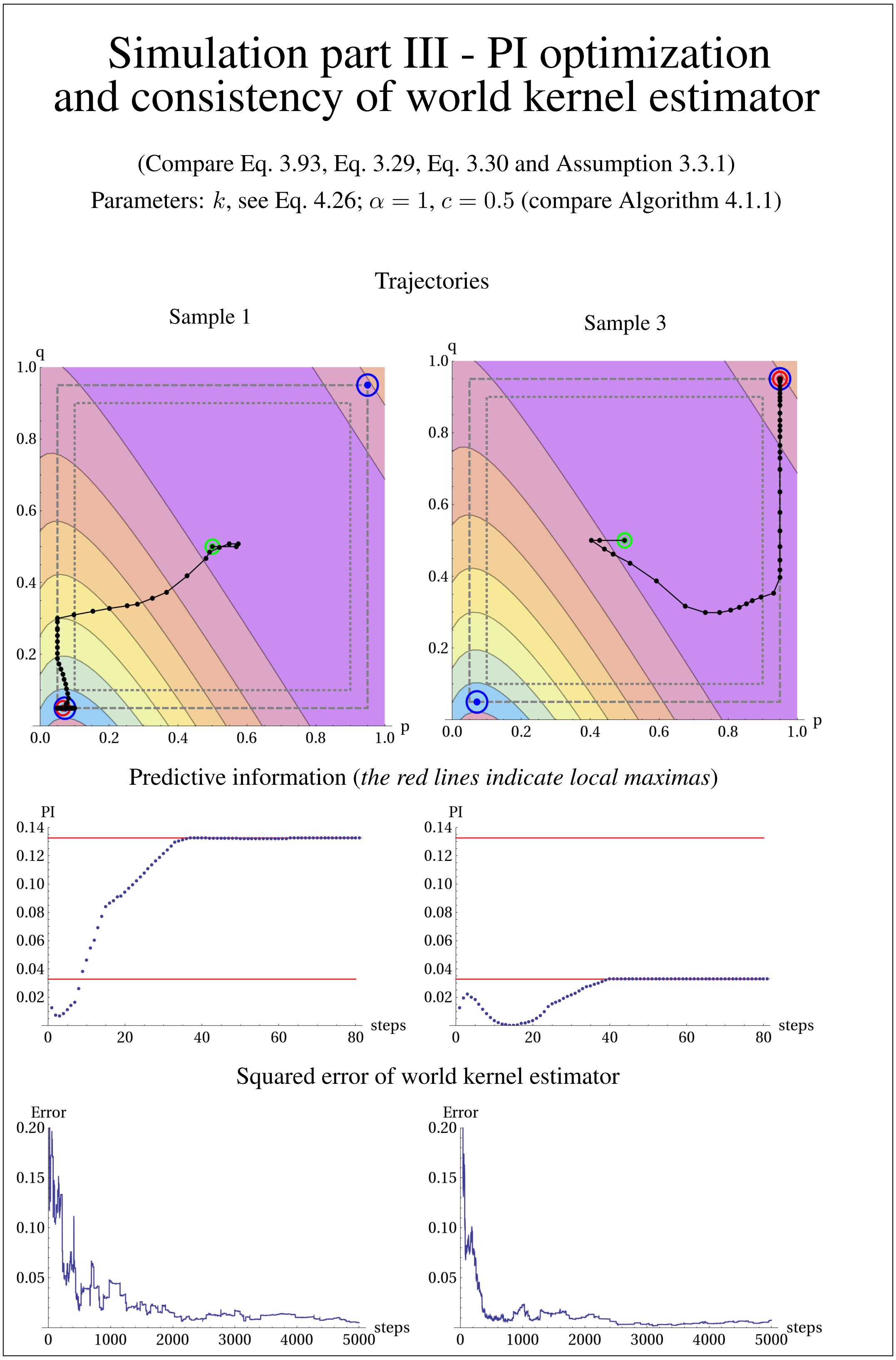


Simulation part II - sample trajectories

(Compare Eq. 3.93, Eq. 3.29, Eq. 3.30 and Assumption 3.3.1)

Parameters: k, see Eq. 4.26; $\alpha = 1$, $c = 0.5$ (compare Algorithm 4.1.1)

Figure 4.2: Simulation of PI maximization - part 2



Simulation part III - PI optimization and consistency of world kernel estimator

(Compare Eq. 3.93, Eq. 3.29, Eq. 3.30 and Assumption 3.3.1)

Parameters: k, see Eq. 4.26; $\alpha = 1$, $c = 0.5$ (compare Algorithm 4.1.1)

Trajectories

Sample 1

Sample 3

Predictive information (*the red lines indicate local maximas*)

Squared error of world kernel estimator

Figure 4.3: Simulation of PI maximization - part 3

4.2 Optimization of policy functionals for general MDPs

The results from the the previous section hold true for general state spaces as long as the parameter set for the policies is a compact subset of $\mathbb{R}^n$. A look on the proof of Theorem 4.1.1 shows that it relies solely on the existence of a consistent estimator of the parameter of the world transition kernel (consistent in the sense that it converges to the right value for any policy sequence that might arise from any possible learning algorithm). We will make this point more precise in this section. Then the proof automatically carries over to general state spaces - provided there exists a consistent estimator. We impose the following restrictions on the underlying Markov decision process:

Assumption 4.2.1 - (**Underlying spaces and transition functions**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a Markov decision process.*
► **Assumption 4.2.1.1:** $\mathbf{X} = [0, 1]$ *and* $p_x = \nu_{\text{Leb}}$.
► **Assumption 4.2.1.2:** $\mathbf{Q}$ *is a topological space and* $\mathbf{Z}$ *is a compact subset of* $\mathbb{R}^n$ *for some* $n \in \mathbb{N}$.

Definition 4.2.1 - (Q**-estimators**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a MDP with associated causal model* $C' = ((V, E), \mathfrak{S}, \mathfrak{T})$ *(compare: Definition 3.1.1). Let* $\mathbb{F}_0 := \{\emptyset, \mathfrak{S}\}$ *and let* $\mathbb{F}_n := \sigma\left(\{S_i, A_j\}_{0 \le i \le n; 0 \le j < n}\right)$ *for* $n \geq 1$. *Then a Q-estimator is a* $\mathbb{F}$ *adapted sequence of random variables,* $\left(\hat{Q}_n\right)_{n \in \mathbb{N}_0}$, *with values in* $\mathbf{Q}$*:*

$$\hat{Q}_n : \mathfrak{S} \to \mathbf{Q} \text{ ; } \hat{Q}_n \text{ is } \mathbb{F}_n/\mathcal{F}_q\text{-measurable} \tag{4.28}$$

By consistency of a Q-estimator we mean that it finally converges to the true value of Q. We consider two different concepts of consistency:

Definition 4.2.2 - (**Consistency of** Q**-estimators**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a MDP with associated causal model* $C' = ((V, E), \mathfrak{S}, \mathfrak{T})$ *(compare: Definition 3.1.1). Let* $\mathbb{F}_0 := \{\emptyset, \mathfrak{S}\}$ *and* $\mathbb{F}_n := \sigma\left(\{S_i, A_j\}_{0 \le i \le n; 0 \le j < n}\right)$ *for* $n \geq 1$.

► **Definition 4.2.2.1:** *A Q-estimator,* $\left(\hat{Q}_n\right)_{n \in \mathbb{N}}$, *will be called consistent, if*

$$\lim_{n \to \infty} \hat{Q}_n = q' \tag{4.29}$$

almost surely with respect to $P_{\text{MDP}, q', s', \mathbf{z}'}$ *for every* $q' \in \mathbf{Q}$, $s' \in \mathbf{S}$, $\mathbf{z}' \in \mathbf{Z}^{\mathbb{N}_0}$ *(compare Remark 3.1.1).*
► **Definition 4.2.2.2:** *A Q-estimator,* $\left(\hat{Q}_n\right)_{n \in \mathbb{N}}$, *will be called strongly consistent, if*

$$\lim_{n \to \infty} \hat{Q}_n = q' \tag{4.30}$$

almost surely with respect to $P_{q', s', Z}$, *for every* $q' \in \mathbf{Q}$, $s' \in \mathbf{S}$ *and every* $\mathbb{F}$*-adapted process,* Z, *with values in* $\mathbf{Z}$ *(see Remark 4.2.1).*

Remark 4.2.1 - (**Comment on Definition 4.2.2**)

Every $\mathbb{F}$-adapted policy process, Z, naturally defines a "learning algorithm" over the MDP in the sense of Definition 3.1.2 and Remark 3.1.2. To see this set:

$$\mathbf{M} := \mathbb{N} \times \mathbf{S}^{\mathbb{N}_0} \times \mathbf{A}^{\mathbb{N}_0} \times \mathbf{Z} \tag{4.31}$$

Since $\mathbb{F}_0$ is trivial, we have that $Z_0 = z'$ almost surely for some $z' \in \mathbf{Z}$. Since Z is $\mathbb{F}$-adapted for every $n \geq 1$ there exists some $(\otimes_{0\leq i\leq n}\mathcal{F}_S) \otimes (\otimes_{0\leq i<n}\mathcal{F}_A)/\mathcal{F}_Z$-measurable functions, $T_n : \mathbf{S}^{n+1} \times \mathbf{A}^n \to \mathbf{Z}$, such that

$$Z_n = T_n\left((S_i)_{0\leq i\leq n}, (A_i)_{0\leq i<n}\right)$$

Fix arbitrary $\mathbf{a}_0 \in \mathbf{A}^{\mathbb{N}_0}$, $\mathbf{s}_0 \in \mathbf{S}^{\mathbb{N}_0}$ and define $m' := (0, \mathbf{s}_0, \mathbf{a}_0, z')$. Set

$$U : \mathbf{M} \to \mathbf{Z} \; ; (n, \mathbf{s}, \mathbf{a}, z) \mapsto z$$

and

$$L : \mathbf{S} \times \mathbf{A} \times \mathbf{S} \times \mathbf{M} \to \mathbf{Z} \; ; \; \big(s, a, s', (n, \mathbf{s}, \mathbf{a}, z)\big) \mapsto (n+1, \mathbf{s}', \mathbf{a}', z')$$

where

$$\mathbf{a}'_k := \begin{cases} a & \text{if } k = n \\ \mathbf{a}_k & \text{else} \end{cases},$$

$$\mathbf{s}'_k := \begin{cases} s & \text{if } k = n \\ s' & \text{if } k = n+1 \\ \mathbf{s}_k & \text{else} \end{cases}$$

and

$$z' := T_{n+1}\left((\mathbf{s}_0, \mathbf{s}_1, \ldots, \mathbf{s}_{n-1}, s, s'), (\mathbf{a}_0, \mathbf{s}_1, \ldots, \mathbf{a}_{n-1}, a)\right).$$

Then the law generated by the learning algorithm $(\mathbf{M}, L, U)$ for initial sensor state $s' \in \mathbf{S}$ and parameter $q' \in \mathbf{Q}$, $P_{q',s',m'}$ (compare Remark 3.1.2) does not depend on the choice of $\mathbf{s}_0$ and $\mathbf{a}_0$ but only on the adapted process, Z. So we will simply write $P_{q',s',Z}$ for the measure $P_{q',s',m'}$.

We assume the existence of a strongly consistent estimator of the world parameter $q \in \mathbf{Q}$

Assumption 4.2.2 - (**Existence of consistent estimator**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ be a Markov decision process satisfying Assumption 4.2.1. We assume the existence of a strongly consistent Q-estimator, $\left(\hat{Q}_n\right)_{n\in\mathbb{N}}$.

Remark 4.2.2 - (**Consistency and strong consistency**)

Obviously every deterministic sequence of policies is adapted to the filtration, $\mathbb{F}$, such that strong consistency includes consistency. To see that strong consistency is really stronger than mere consistency consider the following example:

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ be an MDP with

- $\mathbf{Q} := \mathbb{R}$
- $\mathbf{Z} := \{+1, -1\}$
- $\mathbf{S} := \{s_1, s_2\}$
- $\mathbf{A} := \{+1, -1\}$
- $\mathbf{X}$ *and* $\mathbf{p}_x$ *are irrelevant*

- $$T\left[\left(s_1, a', q'\right), \{s_1\}\right] := \frac{1}{2}\left(1 + a' \cdot \tanh\left(q'\right)\right) = 1 - T\left[\left(s_1, a', q'\right), \{s_2\}\right]$$

  and

  $$T\left[\left(s_2, a', q'\right), \{s_2\}\right] := \frac{1}{2}\left(1 + a' \cdot \tanh\left(q'\right)\right) = 1 - T\left[\left(s_2, a', q'\right), \{s_1\}\right]$$

- $\Pi\left(s, x, z\right) := z$

for a fixed sequence of policies, $\mathbf{z}' \in \mathbf{Z}^{\mathbb{N}_0}$*, the sensor process is Markovian with transition matrices*

$$\hat{K}_n := \begin{pmatrix} \frac{1}{2}\left(1 + \mathbf{z}'_n \cdot \tanh\left(q'\right)\right) & \frac{1}{2}\left(1 - \mathbf{z}'_n \cdot \tanh\left(q'\right)\right) \\ \frac{1}{2}\left(1 - \mathbf{z}'_n \cdot \tanh\left(q'\right)\right) & \frac{1}{2}\left(1 + \mathbf{z}'_n \cdot \tanh\left(q'\right)\right) \end{pmatrix} \tag{4.32}$$

with left-Eigenvectors $(0.5, 0.5)$ *(with respect to Eigenvalue* 1*) and* $(0.5, -0.5)$ *(with respect to Eigenvalue* $\lambda_2 := \mathbf{z}_n \tanh\left(q'\right)$*). Therefore the distribution,* p_n*, of the* n*-th sensor value converges to* $(0.5, 0.5)$ *with the error estimate:*

$$\left\|p_n - (0.5, 0.5)\right\|_1 \leq \left|p_0 - q_0\right| \left|\tanh\left(q'\right)\right|^n$$

By this insight and by the law of large numbers:

$$\begin{aligned} & \lim_{n\to\infty} \frac{\sum_{k=0}^{n} A_k \cdot \left(\mathbb{1}_{\{S_k = s_1, S_{k+1} = s_1\}} - 0.5\right)}{n+1} \\ = \; & \lim_{n\to\infty} \frac{\sum_{k=0}^{n} A_k \cdot \left(\mathbb{1}_{\{S_k = s_1, S_{k+1} = s_1\}} - 0.5\right)}{\max\left\{\sum_{k=0}^{n} \mathbb{1}_{\{S_k = s_1\}}, 1\right\}} \frac{\max\left\{\sum_{k=0}^{n} \mathbb{1}_{\{S_k = s_1\}}, 1\right\}}{n+1} \\ = \; & \frac{1}{4} \tanh\left(q'\right) \end{aligned} \tag{4.33}$$

a.s. with respect to $P_{\mathrm{MDP}, q', s', \mathbf{z}'}$*. Therefore*

$$\hat{Q}_n := \begin{cases} 0 & \text{if } \left|\frac{\sum_{k=0}^{n} A_k \cdot \left(\mathbb{1}_{\{S_k = s_1, S_{k+1} = s_1\}} - 0.5\right)}{n+1}\right| > \frac{1}{4} \\ \operatorname{artanh}\left(4 \frac{\sum_{k=0}^{n} A_k \cdot \left(\mathbb{1}_{\{S_k = s_1, S_{k+1} = s_1\}} - 0.5\right)}{n+1}\right) & \text{else} \end{cases} \tag{4.34}$$

is a consistent estimator. Now consider the adapted policy sequence

$$Z_k := \begin{cases} +1 & \text{if } S_k = s_1 \\ -1 & \text{else} \end{cases}$$

Then S_k *is still Markovian but with transition matrix*

$$\hat{K}'_n := \begin{pmatrix} \frac{1}{2}\left(1 + \tanh\left(q'\right)\right) & \frac{1}{2}\left(1 - \tanh\left(q'\right)\right) \\ \frac{1}{2}\left(1 + \tanh\left(q'\right)\right) & \frac{1}{2}\left(1 - \tanh\left(q'\right)\right) \end{pmatrix} \tag{4.35}$$

such that the stationary distribution becomes $\left(\frac{1}{2}\left(1 + \tanh\left(q'\right)\right), \frac{1}{2}\left(1 - \tanh\left(q'\right)\right)\right)$ *and a repetition of the former calculations yields*

$$\lim_{n\to\infty} \hat{Q}_n = \begin{cases} \operatorname{artanh}\left[\tanh(q')\left(1 + \tanh(q')\right)\right] & \text{if } \left|\tanh(q')\left(1 + \tanh(q')\right)\right| \leq 1 \\ 0 & \text{else} \end{cases} \tag{4.36}$$

almost surely with respect to $P_{q', s', Z}$*, such that* $\left(\hat{Q}_n\right)_{n\in\mathbb{N}_0}$ *is not strongly consistent.*

Since every learning algorithm results in an adapted sequence of policy parameters, we get:

Corrolary 4.2.1 - (**Consistent estimators and learning algorithms**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a causal model and let* $\hat{Q}_n$ *be a strongly consistent Q-estimator (according to Assumption 4.2.2). Let* $(\mathbf{M}, L, U)$ *be a learning algorithm over* C *(compare Definition 3.1.2) then:*

$$\lim_{n\to\infty} \hat{Q}_n = q' \text{ a.s. w.r.t. } P_{q',s',m'} \tag{4.37}$$

for every $q' \in \mathbf{Q}$*,* $s' \in \mathbf{S}$ *and* $m' \in \mathbf{M}$.

Last but not least we require the objective function to satisfy the following regularity properties:

Assumption 4.2.3 - (**Regularity of objective function**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a Markov decision process and assume that Assumption 4.2.1 holds. Let*

$$\phi : \mathbf{Q} \times \mathbb{R}^n \to \mathbb{R} \tag{4.38}$$

be a function. We assume:

▶ **Assumption 4.2.3.1:** *For every* $q \in \mathbf{Q}$ *the function* $\phi_q : z \mapsto \phi(q, z)$ *is continuously differentiable in a neighborhood of* $\mathbf{Z}$*. We further assume that the function* ϕ_q *restricted to* $\mathbf{Z}$ *has a negligible set of critical values (compare Definition 2.4.3).*
▶ **Assumption 4.2.3.2:** *The function* $q \mapsto D_2\phi(q, z)$ *is continuous uniformly in* z *over* $\mathbf{Z}$*. By this we mean that for fixed* $q \in \mathbf{Q}$ *and any* $\epsilon > 0$ *there exist* $\delta > 0$ *such that:*

$$\sup\left\{\left|D_2\phi(q', z) - D_2\phi(q, z)\right| \middle| z \in \mathbf{Z}\right\} < \epsilon \text{ whenever } \left\|q' - q\right\| < \delta \tag{4.39}$$

Consider the following algorithm:

Algorithm 4.2.1 - (**Gradient ascent on general MDP**)

▶ **Free parameters:**

- *Two ascent decay parameters,* $c \in \mathbb{R}_+$ *and* α*, satisfying* $\frac{1}{2} < \alpha \leq 1$.
- *An essentially Lipschitz continuous quasi-projector,* Pr*, of* $\mathbb{R}^n$ *onto* $\mathbf{Q}$ *(compare Definition 2.3.1)*
- *A (set-valued) metric* G *on* $\mathbf{Z}$ *that is compatible with the quasi-projection,* Pr*, (compare Definition 2.3.1) and selection rules* λ_n*, i.e. a* $\mathbb{F}$*-adapted process with values in* $\mathrm{bil}(\mathbb{R}^n)$ *such that:*

$$\lambda_n \in G(Z_n) \tag{4.40}$$

▶ **Variables and initializations:**

- *An Estimator for the parameter* $q' \in \mathbf{Q}$*:* $\hat{Q} \in \mathbf{Q}$.
- *Current policy:* $\hat{z} \in \mathbb{R}^n$*;* $\hat{z} \leftarrow z_0$ *for some* $z_0 \in \mathbf{Z}$
- *Current metric* $g \in \mathrm{bil}(\mathbb{R}^n)$
- *A step counter:* $t \in \mathbb{N}_0$*,* $t \leftarrow 0$

► **Algorithm:**

```
repeat this:
    $\hat{Q} \leftarrow \hat{Q}_t$
    $g \leftarrow \lambda_t$
    $\hat{z} \leftarrow \mathrm{Pr}\left[\hat{z} + \frac{c}{(t+1)^\alpha}\nabla_{g,2}\phi\left(\hat{Q},\hat{z}\right)\right]$
    $t \leftarrow t+1$
forever
```

Now we are able to formulate and prove the second main theorem of this section:

Theorem 4.2.1 - (**Convergence of Algorithm 4.2.1**)

Assume that Assumption 4.2.1, Assumption 4.2.2 and Assumption 4.2.3 are satisfied. Consider the algorithm Algorithm 4.2.1. Then the iterative sequence $(\hat{z}_n)_{n\in\mathbb{N}}$ *converges to the set of first order optimal points of* ϕ_q*:*

$$S_{*,q} := \{z \in \mathbf{Z} \,|D_2\phi\,(q,z) \in N_{\mathbf{Z}}(z)\}\,, \tag{4.41}$$

almost surely with respect to P_{q,s_0,m_0} *(definition: see Definition 3.1.2) for every* $q \in \mathbf{Q}$ *and* $s_0 \in \mathbf{S}$.

Proof. Write

$$\nabla_{g,2}\phi(\hat{q},\hat{z}) := \nabla_{g,2}\phi(q,\hat{z}) + \beta_k \tag{4.42}$$

with

$$\beta_k = \nabla_{g,2}\phi(\hat{q},\hat{z}) - \nabla_{g,2}\phi(q,\hat{z}) \tag{4.43}$$

By Assumption 4.2.2 and Corrolary 4.2.1 the sequence $(\beta_k)_{k\in\mathbb{N}}$ converges to zero almost surely. Therefore the result follows immediately from Theorem 2.4.3. ■

4.3 Example: Linear dynamic with Gaussian noise

In this section we apply Algorithm 4.2.1 to a linear Gaussian dynamic. The underlying problem and the algorithm extends the ideas from Ay et al. [16] to a dynamic that includes learning of the system parameters. Assume that the sensor process takes on values in $\mathbb{R}^M$and fix some norm, $\|\cdot\|$ on $\mathbb{R}_M$. In this section we identify the dual space, $\mathbb{R}^{M,*}$, of $\mathbb{R}^M$ with $\mathbb{R}^M$ (via the Euclidean Riesz isomorphism). Denote the dual norm by $\|\cdot\|_*$ (usually $\|\cdot\|$ and $\|\cdot\|_*$ are either equal to the standard Euclidean norm or $\|\cdot\|$ is the $1-$norm and $\|\cdot\|_*$ is the maximum norm. The former choice is compatible with the standard scalar product, where the later one is often computationally easier to handle). Assume further that the action process takes on values in $\mathbb{R}^N$. We use the same symbol to denote a norm on the action space, since it is always clear from the context, which norm is meant. Usually the dimension of the action space, N, is much lower than the dimension of the sensor space, M.

Consider the following dynamic for the sensor values, S_n, taking on values in $\mathbb{R}^M$:

$$S_{n+1} = RA_n + N_n \; ; S_0 := s_0 \tag{4.44}$$

where N_n are IID random variables having Gaussian distribution with zero mean and (unknown) covariance matrix Σ,

$$R \in \left\{X \in \mathbb{R}^{M\times N} \middle| \|X\|_{\text{Op}} < 1\right\}$$

is a fixed (but also unknown) matrix where we defined

$$\|R\|_{\text{Op}} := \sup\left\{\lambda^T Rx \,\middle|\, \lambda \in \mathbb{R}^M; \|\lambda\|_* = 1 \; ; x \in \mathbb{R}^N; \|x\| = 1\right\} \tag{4.45}$$

We assume further that the actions depend linearly on the current sensor value, i.e.

$$A_n = C_n S_n + \epsilon Y_n \tag{4.46}$$

where

$$C_n \in \left\{X \in \mathbb{R}^{N\times M} \middle| \|X\|_{\text{Op}} \le 1\right\} \tag{4.47}$$

is a sequence of policy matrices and ϵY_n is a regularization term forcing the agent to explore the entire action space. We assume $(Y_n)_{n\in\mathbb{N}_0}$ to be a sequence of IID random variables, independent from $(N_k)_{k\in\mathbb{N}_0}$ with M-dimensional standard Gaussian distribution.

Algorithm 4.2.1 can be used to optimize a functional of the policy matrix, $C \in \mathbb{R}^{N\times M}$ and the unknown system parameters, R and Σ with respect to C. In Eq. 4.60 we show how the stationary distribution of a stationary policy process can be calculated from the policy matrix and the system parameters.
Finally we will present an algorithm that converges to the set of first-order optimal points almost surely for every pair of admissible system parameters. The following list summarizes our assumptions on the underlying Markov decision process:

Assumption 4.3.1 - (**Assumptions on linear Gaussian MDP**)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a Markov decision process. We assume*
► **Assumption 4.3.1.1:** $\mathbf{S} = \mathbb{R}^M$ *and* $\mathbf{A} = \mathbb{R}^N$
► **Assumption 4.3.1.2:** $\mathbf{X} = \mathbb{R}^M$*,* p_x *is the standard Gaussian distribution in* M *dimensions.*
► **Assumption 4.3.1.3:**

$$\mathbf{Q} = \mathbf{R} \times \mathbf{N}$$

where

$$\mathbf{R} = \left\{ X \in \mathbb{R}^{M \times N} \, \|X\|_{\text{Op}} < 1 \right\} \tag{4.48}$$

and

$$\mathbf{N} = \left\{ X \in \mathbb{R}^{M \times M} \, \middle| X = X^T > 0 \right\} \tag{4.49}$$

► Assumption 4.3.1.4:

$$\mathbf{Z} = \left\{ X \in \mathbb{R}^{N \times M} \, \middle| \|X\|_{\text{Op}} \leq 1 \right\}$$

► Assumption 4.3.1.5:

$$\Pi\left(s, x, z\right) := zs + \epsilon x \tag{4.50}$$

and for every $A \in \mathcal{B}_{\mathbb{R}^M}$*:*

$$T\left(\left(s, a, \left(R, \Sigma\right)\right), A\right) = \int_{\mathbb{R}^M} \mathbb{1}_A(x) \rho_{Ra,\Sigma}(x) d\nu_{\text{Leb}}(x) \tag{4.51}$$

where

$$\rho_{\mu,\Sigma}(x) = \frac{1}{\sqrt{2\pi}^N} \exp\left(-\frac{\left(x - \mu\right)^T \Sigma^{-1} \left(x - \mu\right)}{2} \right) \tag{4.52}$$

denotes the Gaussian density.

Remark 4.3.1 - (**Comment on definition Assumption 4.3.1**)

Here it is more convenient to model the dynamic on the measurable space spanned by the values of the sensor noise sequence, $\left(N_k\right)_{k \in \mathbb{N}_0}$*, and the values of the policy noise,* $\left(Y_k\right)_{k \in \mathbb{N}_0}$ *(compare Eq. 4.44 and Eq. 4.46). This is possible by Definition 1.2.5, Lemma 1.2.4 and by the observation that all other process values depend deterministically on the noise variables and former process values. So we assume*

$$\left(\Omega, \mathcal{F}\right) := \left(\mathbb{R}^{M^{\mathbb{N}_0}} \times \mathbb{R}^{N^{\mathbb{N}_0}}, \left(\otimes_{n \in \mathbb{N}_0} \mathcal{B}_{\mathbb{R}^M}\right) \otimes_{n \in \mathbb{N}_0} \mathcal{B}_{\mathbb{R}^N} \right)$$

Therefore the measure of the open-loop dynamic (i.e. the MDP without learning), $P_{(R,\Sigma),s_0,\mathbf{z}}$ *where* $(R, \Sigma) \in Q$*,* $s_0 \in \mathbf{S}$ *and* $\mathbf{z} \in \mathbf{Z}^{\mathbb{N}_0}$ *(compare Remark 4.2.1 and Definition 4.2.2), is interpreted as a measure on* $\mathcal{F}$ *in this section. The same holds true for the measure* $P_{(R,\Sigma),s_0,Z}$ *(compare Remark 4.2.1) where again* $(R, \Sigma) \in Q$*,* $s_0 \in \mathbf{S}$ *and* Z *is a policy process with values in* $\mathbf{Z}$ *that is adapted to the filtration*

$$\mathbb{F}_0 := \{\emptyset, \Omega\} \text{ and } \mathbb{F}_n := \sigma(N_k, Y_k)_{0 \leq k < n} \text{ for } n \geq 1$$

Since the processes $\left(S_n\right)_{n \in \mathbb{N}_0}$ *and* $\left(A_n\right)_{n \in \mathbb{N}_0}$ *are* $\mathbb{F}$*-adapted, any policy sequence adapted to the process* $\left(S_n, A_n\right)_{n \in \mathbb{N}_0}$ *is automatically* $\mathbb{F}$*-adapted.*

We will give a short overview about the open-loop dynamic (i.e. the dynamic that results from the MDP without learning, with a fixed policy sequence $C \in \mathbf{Z}^{\mathbb{N}_0}$).

Remark 4.3.2 - (**Discussion of the open-loop dynamic of a linear Gaussian MDP**)

Let P *be the distribution of the noise variables* $(N_k)_{k \in \mathbb{N}_0}$ *and* $(Y_k)_{k \in \mathbb{N}_0}$*. Then the distribution of the* $n-$*th sensor value,* $S_{n,*}P$*, is Gaussian with mean:*

$$\mu_n = \left(\prod_{k=0}^{n-1} RC_k \right) s_0, \tag{4.53}$$

where the multiplication always starts with the lowest index on the rightmost position and the highest index on the leftmost position. Since we assumed $\|R\|_{\text{Op}} < 1$ *and* $\|C_n\|_{\text{Op}} \leq 1$*, the expectation value of the* $n-$*th sensor value,* μ_n*, converges to zero exponentially fast as* n *approaches infinity with the estimate:*

$$\|\mu_n\| \leq \|s_0\| \cdot {\|R\|_{\text{Op}}}^n \tag{4.54}$$

The covariance matrix of S_1 *is*

$$\Sigma_1 := \Sigma + \epsilon^2 RR^T \tag{4.55}$$

and the covariance matrix of S_n *for* $n \geq 2$ *is*

$$\Sigma_n := \Sigma + \epsilon^2 RR^T + \sum_{k=1}^{n-1} \left(\prod_{j=1}^{k} RC_j \right) \left(\Sigma + \epsilon^2 RR^T \right) \left(\prod_{j=1}^{k} RC_j \right)^T \tag{4.56}$$

Therefore the covariance matrices of the sensor values are uniformly bounded by:

$$\|\Sigma_n\|_{\text{Op},*} \leq \frac{\left\|\Sigma + \epsilon^2 RR^T\right\|_{\text{Op},*}}{1 - {\|R\|_{\text{Op}}}^2} \tag{4.57}$$

where we wrote

$$\|A\|_{\text{Op},*} := \sup \left\{ \lambda^T A v \,\middle|\, \lambda, v \in \mathbb{R}^M; \|\lambda\|_* = \|v\|_* = 1 \right\} \tag{4.58}$$

One consequence is that the family of Gaussian measures is exponentially tight (compare König [104], Dembo and Zeitouni [62], Hollander [88]), i.e. for every $L \in \mathbb{R}_{\geq 0}$ *there exists a compact set* $K \subset \mathbb{R}^M$ *such that*

$$\limsup_{n \to \infty} \frac{1}{n} \ln P \left[S_n \in \mathbb{R}^M \setminus K \right] \leq -L \tag{4.59}$$

If $C_n \equiv C$ *then the distribution* $S_{n,*}P$*, converges* [1] *to the unique invariant distribution,* P_∞*, that is Gaussian with mean zero and covariance matrix:*

$$\Sigma_\infty := \sum_{k=0}^{\infty} (RC)^k \left(\Sigma + \epsilon^2 RR^T \right) \left(C^T R^T \right)^k \tag{4.60}$$

In order to apply Algorithm 4.2.1 to the Gaussian Markov decision process, a strongly consistent estimator for R and σ is needed:

Lemma 4.3.1 - (Strongly consistent estimators for the linear Gaussian MDP)

Let $C := (\mathbf{Q}, \mathbf{Z}, \mathbf{S}, \mathbf{A}, \mathbf{X}, p_x, T, \Pi)$ *be a Markov decision process satisfying Assumption 4.3.1. Then*

▶ **Lemma 4.3.1.1:** *The matrix*

$$M_n := \frac{1}{n} \sum_{k=0}^{n-1} A_k {A_k}^T \tag{4.61}$$

is invertible almost surely for sufficiently large n *(even the stronger statement* $\limsup_{n \to \infty} \left\| {M_n}^{-1} \right\| \leq \frac{1}{\epsilon^2}$ *holds true almost surely.).*

▶ **Lemma 4.3.1.2:** *Moreover*

$$\hat{R}_n := \left(\frac{1}{n} \sum_{k=0}^{n-1} S_{k+1} {A_k}^T \right) {M_n}^{-1} \tag{4.62}$$

---

[1] weakly, strongly, in KL-divergence and in total variation norm

is a strongly consistent estimator for R and

$$\hat{\Sigma}_n := \frac{1}{n}\sum_{k=0}^{n-1} S_{k+1}{S_{k+1}}^T - \hat{R}_n M_n \hat{R}_n^T \tag{4.63}$$

is a strongly consistent estimator for Σ.

Before starting a formal proof, we prove a simple but important bound on the moments of the sensor and action values valid for an arbitrary adapted policy process:

Lemma 4.3.2 - (**Bounds on moments of sensor and action values for closed loop dynamic**)

Let $\mathbb{F}$ be the filtration of Remark 4.3.1 and assume that $C := (C_n)_{n\in\mathbb{N}_0}$ is a $\mathbb{F}$-adapted process. Under the law $P_{(R,\Sigma),s_0,C}$ (compare Remark 4.2.1, Definition 4.2.2 and Remark 4.3.1) for any $p > 0$ there exist some constants $c \in \mathbb{R}_{>0}$ (depending on p, ϵ, the dimension of the action space, M, the dimension of the sensor space, N and the covariance matrix, Σ) such that:

$$E_{(R,\Sigma),s_0,C}\left[\|S_n\|^p\right]^{\frac{1}{p}} \leq \|s_0\| + \frac{c}{1-\|R\|_{Op}} \tag{4.64}$$

and there exists a constant $c_2 > r$ (depending on p, ϵ and N) such that

$$E_{(R,\Sigma),s_0,C}\left[\|A_n\|^p\right]^{\frac{1}{p}} \leq \|s_0\| + \frac{c}{1-\|R\|_{Op}} + c_2 \tag{4.65}$$

Proof of Lemma 4.3.2. Note that for any sequence of positive real numbers $(\alpha_n)_{n\in\mathbb{N}_0}$, for $0 \leq r < 1$ and $c \in \mathbb{R}_{>0}$ satisfying the identity

$$\alpha_{n+1} \leq r\alpha_n + c$$

the following estimate holds true:

$$\alpha_n \leq r^n\alpha_0 + \frac{1-r^n}{1-r}c \leq \alpha_0 + \frac{c}{1-r} \tag{4.66}$$

Inserting Eq. 4.46 into Eq. 4.44 and using $\|C_n\|_{\mathrm{Op}} \leq 1$ together with Minkowski's inequality gives:

$$E\left[\|S_{n+1}\|^p\right]^{\frac{1}{p}} \leq \|R\|_{\mathrm{Op}} E\left[\|S_n\|^p\right]^{\frac{1}{p}} + \epsilon\|R\|_{\mathrm{Op}} E\left[\|Y_n\|^p\right]^{\frac{1}{p}} + E\left[\|N_n\|^p\right]^{\frac{1}{p}} \tag{4.67}$$

Setting

$$\alpha_n := E\left[\|S_n\|^p\right]^{\frac{1}{p}},\ r := \|R\|_{\mathrm{Op}}$$

and

$$c := \epsilon\|R\|_{\mathrm{Op}} E\left[\|Y_n\|^p\right]^{\frac{1}{p}} + E\left[\|N_n\|^p\right]^{\frac{1}{p}}$$

in Eq. 4.66 gives the desired estimate for the p-th moment of S_n. The bound for A_n follows from the bound on the Moment of S_n, $\|C_n\|_{\mathrm{Op}} \leq 1$ and the definition of A_n (see 4.46). A good upper bound for the constants c and c_2 (that also justifies the claim about the dependencies on on ϵ, N, M and p) follows from Appendix Lemma 6.0.7. ■

Proof of Lemma 4.3.1. M_n can be rewritten as (compare: Eq. 4.46):

$$M_n = M_{n,1} + M_{n,2} + {M_{n,2}}^T \tag{4.68}$$

where

$$M_{n,1} := \frac{1}{n}\sum_{k=0}^{n-1}\left(C_k S_k {S_k}^T {C_k}^T + \epsilon^2 Y_k {Y_k}^T\right) \tag{4.69}$$

and

$$M_{n,2} := \frac{\epsilon}{n} \sum_{k=0}^{n-1} C_k S_k {Y_k}^T \tag{4.70}$$

By our independence assumption on the noise variables and since we assumed C to be $\mathbb{F}$-adapted, $\left(C_n S_n {Y_n}^T\right)_{n \in \mathbb{N}_0}$ is a $\mathbb{F}$-martingale. By Lemma 4.3.2 all moments of this sequence are uniformly bounded. Therefore the law of large numbers (compare Lemma 6.0.4) implies:

$$\lim_{n \to \infty} M_{n,2} = 0 \text{ a.s.} \tag{4.71}$$

The law of large numbers implies that

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} \epsilon^2 Y_k {Y_k}^T = \epsilon^2 \mathbb{1} \tag{4.72}$$

and

$$\frac{1}{n} \sum_{k=0}^{n-1} \left(C_k S_k {S_k}^T {C_k}^T\right) \tag{4.73}$$

is a positive matrix for any policy sequence. Therefore $M_n - \epsilon^2 \mathbb{1}$ is a positive matrix and M_n is invertible for suffiently large n with the estimate

$$\limsup_{n \to \infty} \left\|{M_n}^{-1}\right\|_{\text{Op}} \leq \frac{1}{\epsilon^2} \tag{4.74}$$

Multiplying Eq. 4.44 with ${A_n}^T$ from the right, relabeling and summing up gives:

$$\frac{1}{n} \sum_{k=0}^{n-1} S_{k+1} {A_k}^T = R M_n + \frac{1}{n} \sum_{k=0}^{n-1} N_k {A_k}^T \tag{4.75}$$

The last summand converges to zero by the law of large numbers and Lemma 4.3.2 again such that:

$$\lim_{n \to \infty} \hat{R}_n = R \text{ a.s.} \tag{4.76}$$

Inserting Eq. 4.44 into the definition of the estimator for the covariance matrix, Eq. 4.63, gives:

$$\begin{aligned} \hat{\Sigma}_n &= \frac{1}{n} \sum_{k=0}^{n-1} \left(R A_k {A_k}^T R^T - \hat{R}_n A_k {A_k}^T \hat{R}_n^T\right) + \frac{1}{n} \sum_{k=0}^{n-1} N_k {N_k}^T + \\ &\quad + \frac{1}{n} \sum_{k=0}^{n-1} \left(R A_k {N_k}^T + N_k {A_k}^T R^T\right) \end{aligned} \tag{4.77}$$

The convergence of $\hat{R}_n$ to R and the law of large numbers yields

$$\lim_{n \to \infty} \hat{\Sigma}_n = \Sigma \text{ a.s.} \tag{4.78}$$

■

Beside the estimator for the system variables, a quasi-projector onto the unit ball in operator norm and a compatible metric are needed (compare Theorem 2.4.3). As already indicated in section 2.3 an appropriate choice is the retraction onto the unit ball:

$$A \mapsto \min\left\{1, \frac{1}{\|A\|_{\text{Op}}}\right\} A \tag{4.79}$$

where we set $1/0 := \infty$ for convenience. As already stressed out in 2.3 a compatible metric is:

$$g^{(\lambda,v)}{}_A(X,Y) := \left(\lambda^T X v\right)\left(\lambda^T Y v\right) + \mathrm{Tr}\left[\left(X - \frac{\mathrm{Tr}\left[X^T A\right]}{\mathrm{Tr}\left(A^T A\right)} A\right)^T \left(Y - \frac{\mathrm{Tr}\left[Y^T A\right]}{\mathrm{Tr}\left(A^T A\right)} A\right)\right] \tag{4.80}$$

where $\lambda \in \mathbb{R}^N$ with $\|\lambda\|_* = 1$ and $v \in \mathbb{R}^N$ with $\|v\| = 1$ such that $\lambda(Av) = \|A\|_{\mathrm{Op}}$. We will provide a convenient formula for the gradient of some function $\phi : \mathbb{R}^{N\times N} \to \mathbb{R}$ with respect to the metric $g^{(\lambda,v)}{}_A$. Generally the calculation of a gradient requires the inversion of the metric tensor, which is a $N^2 \times N^2$ matrix here. However the special structure of the metric allows a reduction of the problem to two dimensions and can therefore be solved much easier. We will write

$$\langle X, Y\rangle = \mathrm{Tr}\left(X^T Y\right) \tag{4.81}$$

for the Euclidean scalar product and

$$\nabla\phi(x) \text{ where } \nabla\phi(x)_{i,j} := \frac{\partial\phi}{\partial x_{i,j}}(x) \tag{4.82}$$

for the Euclidean gradient. The metric g can be expressed by the Euclidean scalar product as

$$g^{(\lambda,v)}{}_A(X,Y) := \langle X,Y\rangle + \langle v_1, X\rangle\langle v_1, Y\rangle - \langle v_2, X\rangle\langle v_2, Y\rangle \tag{4.83}$$

where

$$v_1 := \lambda v^T \text{ and } v_2 := \frac{A}{\sqrt{\mathrm{Tr}\left(AA^T\right)}} \tag{4.84}$$

Therefore the definition of the gradient, $\nabla_g\phi$, with respect to $g^{(\lambda,v)}{}_A$ can be rewritten in the form:

$$\langle\nabla\phi(x), Y\rangle = \langle\left(\nabla_g\phi(x) + \langle v_1, \nabla_g\phi(x)\rangle v_1 - \langle v_2, \nabla_g\phi(x)\rangle v_2\right), Y\rangle \tag{4.85}$$

for every $Y \in \mathbb{R}^{N\times N}$. Since the Euclidean scalar product is positive definite, Eq. 4.82 implies

$$\nabla\phi(x) = \nabla_g\phi(x) + \langle v_1, \nabla_g\phi(x)\rangle v_1 - \langle v_2, \nabla_g\phi(x)\rangle v_2 \tag{4.86}$$

Let P denote the othorgonal projector of $\mathbb{R}^{N\times N}$ onto the linear span of v_1 and v_2 (we will provide an explicit formula for P immediately after illustrating the main idea). Then

$$\nabla_g\phi(x) = P\left[\nabla_g\phi(x)\right] + (\mathbb{1} - P)\left[\nabla_g\phi(x)\right] \text{ and } (\mathbb{1} - P)\left[\nabla_g\phi(x)\right] = (\mathbb{1} - P)\left[\nabla\phi(x)\right] \tag{4.87}$$

The remaining problem is to find $P\left[\nabla_g\phi(x)\right]$. Expanding $P\left[\nabla\phi(x)\right]$ and $P\left[\nabla\phi_g(x)\right]$ into the basis v_1, v_2 (the case that v_1 and v_2 are linearly dependent has to be treated separately) gives a linear system of dimension two that can be solved in closed form easily. The result is:

Lemma 4.3.3 - **(Gradients with respect to the metric** $g^{(\lambda,v)}{}_A$**)**

Let $\phi : \mathbb{R}^{N\times N} \to \mathbb{R}$ *be continuously differentiable and assume* A, v, λ *to be the corresponding parameters of* $g^{(\lambda,v)}{}_A$*. If*

$$v_1 := \lambda v^T \text{ and } v_2 := \frac{A}{\sqrt{\mathrm{Tr}\left(AA^T\right)}}, \tag{4.88}$$

are linearly independent, set

$$P[x] := \langle w_{1,*}, x\rangle w_1 + \langle w_{2,*}, x\rangle w_2 \tag{4.89}$$

where

$$\langle v_{1,*}, x\rangle = \frac{\mathrm{Tr}\left(A^TA\right)\left(\lambda^Txv\right) - \|A\|_{\mathrm{Op}}\,\mathrm{Tr}\left(A^Tx\right)}{\left(\lambda^T\lambda\right)\left(v^Tv\right)\left(\mathrm{Tr}\left(A^TA\right)\right) - \|A\|^2_{\mathrm{Op}}} \tag{4.90}$$

and

$$\langle v_{2,*}, x\rangle = \sqrt{\mathrm{Tr}\left(A^TA\right)}\frac{\left(\lambda^T\lambda\right)\left(v^Tv\right)\mathrm{Tr}\left(A^Tx\right) - \|A\|_{\mathrm{Op}}\left(\lambda^Txv\right)}{\left(\lambda^T\lambda\right)\left(v^Tv\right)\left(\mathrm{Tr}\left(A^TA\right)\right) - \|A\|^2_{\mathrm{Op}}} \tag{4.91}$$

In this case

$$\begin{aligned}\nabla_g\phi(x) &= \nabla\phi(x) - P\left[\nabla\phi(x)\right] - \frac{\langle v_{2,*}, \nabla\phi(x)\rangle}{\|A\|_{\mathrm{Op}}}v_1 + \frac{\langle v_{1,*}, \nabla\phi(x)\rangle}{\|A\|_{\mathrm{Op}}}v_2 + \\ &\quad + \frac{\left(1 + \left(\lambda^T\lambda\right)\left(v^Tv\right)\right)\langle v_{2,*}, \nabla\phi(x)\rangle}{\|A\|^2_{\mathrm{Op}}}v_2\end{aligned}$$

If v_1 and v_2 are linearly dependent then necessarily $v_1 = \|A\|_{\mathrm{Op}}\,v_2$ and

$$\nabla_g\phi(x) = \nabla\phi(x) + \left(\frac{1}{\|A\|^2_{\mathrm{Op}}} - \frac{1}{\mathrm{Tr}\left(A^TA\right)}\right)A$$

With this formula we finish our example on a linear dynamic with Gaussian noise. For an implementation of the algorithm one has to modify the metric, since ${g^{(\lambda,v)}}_A$ is not defined at the origin. A possible way to modify the metric is shown in Eq. 2.57 of section2.3. Another simple choice is:

$$g_A := \begin{cases}\langle\cdot,\cdot\rangle & \text{if } \|A\|_{\mathrm{Op}} \leq \frac{1}{2} \\ {g^{(\lambda,v)}}_A & \text{else}\end{cases} \tag{4.92}$$

We choose the Gaussian dynamics as an example, because it is rather simple. The open loop dynamic is Gaussian, it is very easy to provide analytic expressions for the estimators, and the invariant distribution for a fixed stationary policy for example. However the algorithm presented in the former section works in more involved cases too.

The projected gradient ascent on the unit ball of matrices equipped with some operator norm is not bounded to the linear Gaussian dynamic of course. The idea to use quasi-projectors different from the best approximation with respect to Euclidean distance and to compensate for the error by using some adapted, non-Euclidean metric is essentially new to our knowledge.

Chapter 5

Outlook

In the first chapter we defined causal models over recursively constructible graphs and investigated their probabilistic properties. We think that these models together with the optimization algorithms introduced in Chapter 2 provide a powerfull tool for the formulation and investigation of (multi-agent) learning algorithms. We applied these methods to agents interacting with an MDP via the sensorimotor loop. There are plenty of other problems that fit into this context, including

- **Partially observable Markov decision processes (POMDP)**
  What changes if the sensor values form a hidden Markov process? It is clear how to describe the dynamics by a causal graph but the formulation of reasonable objective functions becomes more difficult. In this work we focused on objective functions that depend on the fixed-policy sensor process, i.e. on the the transition kernel $k \in \Lambda^{\mathbf{S}}_{\mathbf{S}\times\mathbf{A}}$ and the initial distribution (or a stationary distribution for the sensor process). This is a bad approach for hidden markov model, since the transition probabilities:

  $$P\left[S_{n+1} \in \cdot \,|S_n, A_n\right]$$

  vary with $n \in \mathbb{N}$. For the Markov model it is reasonable to assume that there exists a consistent estimator for the world parameter, Q. Whenever there are hidden notes in the model this assumption can be expected to be false. In the case of hidden notes a theoretically convenient quantity that can be estimated from the data is the conditional expectation of Q given all the accessible vertices. There are two ways to approach learning in models with hidden notes: Either one can optimize a target function given the observable data, a route which naturally leads to stochastic filtering theory or one can develop an optimization algorithm in a Markovian setup, apply it to the hidden Markovian model and investigate its behavior. In the later case the algoithm is more a behavorial heuristic then a propper optimization algorithm. However computer simulations show that the finite-state space model from Chapter 4 perform very well in a controlled POMDP setup (article in preparation, also see: list of publications). Interestingly enough the Fisher gradient gives much better results in this case. This is in accordance with findings from Amari (Amari [3] and Amari and Douglas [4]) and Ay, Montúfar, and Rauh [12].

- **Multi-agent problems** The technics also allow a treatment of systems with several agents who optimize their individual target functions while improving their estimates of the system parameters in the course of time. The prerequisite is a propper understanding of the limiting ODE (compare Theorem 2.4.1).

Beside an investigation of specific learning dynamics it is also interesting to answer fundamental questions about the agent-environment system, especially which limitations for learning originate from the dynamic (a very simple, example is Theorem 3.3.1).

In this work we did not apply the theory to state spaces with infinite vector spaces. The setup introduced here can also be used to describe bang control (i.e. a control at discrete time points) of (stochastic) partial differential equations for example. This is another interesting direction for future research.

Chapter 6

Appendix

A.1 - Probability theory

In the first section of the appendix we will list some important concepts and theorems from measure theoretic probability theory. The purpose of this section is mainly to introduce concepts and notation that are used frequently in the main text. A detailed explanation with essential proofs can be found in any good text book on measure theoretic probability (see for example Kallenberg [99], Bauer [22], König [105]). In A.1.2 we will list some important probabilistic inequalities that we need during the excursion. Last but not least we will provide a very condensed summary of theoretical results about finite state space Markov processes including perturbation theory in the case of finite state spaces. These results are important for the convergence proofs in Chapter 4.

A.1.1 - Basic concepts

Let $(\Omega, \mathcal{F})$ be a measurable space, i.e. a set, Ω together with a σ-algebra, $\mathcal{F} \subseteq 2^{\Omega}$, of events. The set of probability measures on $\mathcal{F}$ will be denoted by $M_1(\mathcal{F})$. A probability space is a triplet $(\Omega, \mathcal{F}, P)$ where $(\Omega, \mathcal{F})$ is a measurable space and $P \in M_1(\mathcal{F})$.

Let $(\mathbf{X}, \mathcal{F}_X)$ be a measurable space. A $\mathbf{X}$-valued random variable is a $\mathcal{F}/\mathcal{F}_X$ measurable map:

$$\phi : \Omega \to \mathbf{X} \tag{6.1}$$

The distribution of X is the measure P_*X, defined by

$$P_*X[A] := P\left[\phi^{-1}[A]\right] \text{ for every } A \in \mathcal{F}_X \tag{6.2}$$

Frequently the set $\mathbf{X}$ carries a natural topology, Σ, (i.e. finite or countable infinite sets are usually equipped with the discrete topology, $\mathbb{R}$ is usually equipped with the standard topology generated by open intervals). Then usually $\mathcal{F}$ is the Borel σ-algebra of Σ. In this case we will frequently denote the σ-algebra by $\mathcal{B}$ instead of $\mathcal{F}$. The standard Borel $\sigma-$algebra on $\mathbb{R}^N$ will be denoted by $\mathcal{B}_{\mathbb{R}^N}$. Throughout this thesis we mainly use the notation used in probability theory. So for a $\mathbf{X}$-valued random variables ϕ and $A \in \mathcal{F}_X$ we use the abbreviation

$$\{\phi \in A\} := \{\omega \in \Omega \,|\phi(\omega) \in A\}$$

for example. If ψ is a measurable map from $(\mathbf{X}, \mathcal{F}_X)$ to $(\mathbf{Y}, \mathcal{F}_Y)$ we will write $\psi(\phi)$ for the concatenation (instead of $\psi \circ \phi$ what is used in most analysis text books). For a positive, (extended) real-valued random variable, X we will write either $E[X]$ or $\int_{\Omega} X(\omega)P(d\omega) = \int_{\mathbb{R}} x(X_*P)(dx)$ for its expectation. Sometimes we will also write $E_P[X]$ if we want to highlight the dependence of the expectation on the measure P. As usual we define:

$$\mathcal{L}_1(dP) = \left\{f : \Omega \to \overline{\mathbb{R}} \,\middle|\, f \text{ is } \mathcal{F}/\mathcal{B}_{\overline{\mathbb{R}}} \text{ measurable ; } E_P[|f|] < \infty\right\} \tag{6.3}$$

and $L_1(dP) = \mathcal{L}_1(dP)/\mathcal{N}_P$ where $\mathcal{N}_P := \{f \in \mathcal{L}_1(dP) \,|E_P[|f|] = 0\}$ is the ideal of random variables that vanish P-almost surely. As usual we will simply write $\phi \in L_1(dP)$

for some $\phi \in \mathcal{L}_1(dP)$ instead of writing $[\phi] \in L_1(dP)$ or $\phi + \mathcal{N}_P \in L_1(dP)$. The positive part of a real-valued random variable, X, will be denoted by X_+ and the negative part by X_-.

Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub σ-algebra and let X be a $\mathcal{F}/\mathcal{B}_{\overline{\mathbb{R}}}$ measurable random variable. Assume that either $X_+ \in L_1(dP)$ or $X_- \in L_1(dP)$. A $\mathcal{F}/\mathcal{B}_{\overline{\mathbb{R}}}$-measurable random variable, Y, will be called a version of the conditional expectation of X given $\mathcal{G}$, if

1. $$Y \text{ is } \mathcal{G}/\mathcal{B}_{\overline{\mathbb{R}}} \text{ measurable ; and} \tag{6.4}$$

2. $$E\left[X\mathbb{1}_A\right] = E\left[Y\mathbb{1}_A\right] \text{ for every } A \in \mathcal{G} \tag{6.5}$$

This relation defines Y uniquely almost everywhere w.r.t. P [1]. We will write $E\left[X\,|\mathcal{G}\right]$ for any random variable, Y, satisfying Eq. 6.4 and Eq. 6.5.

Since many text books define conditional expectations for random variables $X \in L_1(dP)$ only we sketch an existence proof here. Define the two measures $d\mu_+ := X_+dP$ and $d\mu_- := X_-dP$ on $\mathcal{F}$. Both are absolutely continuous with respect to P (i.e. $P(A) = 0$ implies $\mu_+(A) = 0$ and $\mu_-(A) = 0$). Hence $\mu_+\,|_{\mathcal{G}}$ and $\mu_-\,|_{\mathcal{G}}$ are absolutely continuous with respect to the finite measure $P\,|_{\mathcal{G}}$. Hence the Radon-Nikodym derivatives $\frac{d\mu_+|_{\mathcal{G}}}{dP|_{\mathcal{G}}}$ and $\frac{d\mu_-|_{\mathcal{G}}}{dP|_{\mathcal{G}}}$ exist. A version of the conditional expectation is

$$E\left[X\,|\mathcal{G}\right] := \frac{d\mu_+\,|_{\mathcal{G}}}{dP\,|_{\mathcal{G}}} - \frac{d\mu_-\,|_{\mathcal{G}}}{dP\,|_{\mathcal{G}}},$$

then this expression is shown to satisfy Eq. 6.4 and Eq. 6.5. To show uniqueness consider two versions of the conditional expectation of X given $\mathcal{G}$, Y_1 and Y_2. Since $A := \{Y_1 < Y_2\}$ is $\mathcal{G}-$measurable, the defining identity of the conditional expectation yields:

$$E\left[X\mathbb{1}_A\right] = E\left[Y_1\mathbb{1}_A\right] = E\left[Y_2\mathbb{1}_A\right]$$

or $E\left[(Y_2 - Y_1)\mathbb{1}_A\right] = 0$. Since $Y_2 - Y_1 > 0$ on A this implies $P(\{Y_1 < Y_2\}) = 0$. Exchanging the roles of Y_1 and Y_2 yields $P\left[\{Y_1 \neq Y_2\}\right] = 0$.

As common in probability theory we will simply write $E\left[X\,|Y\right]$ for $E\left[X\,|\sigma\{Y\}\right]$, where $\sigma\{Y\}$ denotes the σ-algebra generated by Y. Moreover we will write $P\left[A\,|\mathcal{F}\right]$ instead of $E\left[\mathbb{1}_A\,|\mathcal{F}\right]$. So the expression $P\left[\{X \in A\}\,|\mathcal{G}\right] = \xi(Y)$ a.s. stands for $E\left[\mathbb{1}_{\{X\in A\}}\,|\mathcal{G}\right](\omega) = \xi(Y(\omega))$ for every $\omega \in \Omega \setminus N$ for some $N \in \mathcal{G}$ with $P(N) = 0$.

We will consider time-discrete algorithms so we assume that all processes are indexed by natural numbers. A family of sub $\sigma-$algebras $\mathcal{F}_t \subseteq \mathcal{F}$ will be called filtration if $\mathcal{F}_s \subseteq \mathcal{F}_t$ whenever $s \leq t$. We will usually denote filtrations by math-style double-barred letters, for example $\mathbb{F} := (\mathcal{F}_t)_{t\in\mathbb{N}}$. A filtered probability space is a four-tuple $(\Omega, \mathcal{F}, P, \mathbb{F})$ where $(\Omega, \mathcal{F}, P)$ is a probability space and $\mathbb{F}$ is a filtration of $\mathcal{F}$.

Let $(\Omega, \mathcal{F}, P, \mathbb{F})$ be a filtered probability space and let $X = (X_t)_{t\in\mathbb{N}}$ be a stochastic process over the probability space $(\Omega, \mathcal{F}, P)$, taking on values in some measurable space $(\mathbf{X}, \mathcal{F}_X)$. By this we mean that every X_t is a $\mathcal{F}/\mathcal{F}_X$ measurable random variable. X will be called $\mathbb{F}$-adapted if X_t is $\mathbb{F}_t/\mathcal{F}_X$-measurable for every $t \in \mathbb{N}$.
An adapted process X on the filtered probability space $(\Omega, \mathcal{F}, P, \mathbb{F})$ is called Markov process with respect to $\mathbb{F}$, if for almost every $\omega \in \Omega$:

$$P\left[X_t \in A\,|\mathbb{F}_s\right](\omega) = P\left[X_t \in A\,|X_s\right](\omega) \text{ whenever } s \leq t \text{ and } A \in \mathcal{F}_X. \tag{6.6}$$

---

[1] Note that "almost everywhere" means up to changes on null-sets in $\mathcal{G}$ here. Sometimes it is required explicitly that $\mathcal{G}$ contains all P-null sets of $\mathcal{F}$. In this case there is no difference between "up to null-sets" in $\mathcal{G}$ and "up to null-sets" in $\mathcal{F}$.

Let $(\mathbf{X}, \mathcal{F}_X)$ and $(\mathbf{Y}, \mathcal{F}_Y)$ be two measurable sets. A probability kernel (often also referred to as Markov kernel) is a map:

$$K : \mathbf{X} \times \mathcal{F}_Y \to [0,1]$$

such that $K(\cdot, A)$ is $\mathcal{F}_X/\mathcal{B}_{[0,1]}$ measurable for every $A \in \mathcal{F}_Y$ and such that $K(x, \cdot)$ is a probability measure for every $x \in \mathbf{X}$. The set of probability kernels from some measurable space $(\mathbf{X}, \mathcal{F}_X)$ to another measurable space $(\mathbf{Y}, \mathcal{F}_Y)$ will be denoted by $\Lambda_{\mathbf{X}}^{\mathbf{Y}}$. If we want to emphasize the dependence on the σ-algebras on $\mathbf{X}$ and/or $\mathbf{Y}$ we will also write $\Lambda_{(\mathbf{X},\mathcal{F}_X)}^{(\mathbf{Y},\mathcal{F}_Y)}$, $\Lambda_{(\mathbf{X},\mathcal{F}_X)}^{\mathbf{Y}}$ or $\Lambda_{\mathbf{X}}^{(\mathbf{Y},\mathcal{F}_Y)}$. In the case of finite sets we will not distinguish conceptually between kernels and the associated stochastic matrices.

Let $X = (X_t)_{t \in \mathbb{N}_0}$ be an adapted, real-valued process over the filtered probability space $(\Omega, \mathcal{F}, P, \mathbb{F})$. Then X is called $\mathbb{F}$-martingale if:

$$E[X_t | \mathcal{F}_s] = X_s \text{ a.s. whenever } s \leq t \tag{6.7}$$

An $\mathbb{R}^n$-valued random variable X_t is a martingale if Eq.6.7 holds component-wise.

Let X be a stochastic process over some probability space $(\Omega, \mathcal{F}, P)$ with values in $\mathbb{R}^n$ (more generally the state space could be any other topological space) and let $x \in \mathbb{R}^n$. Then X converges almost surely to x, if

$$\lim_{n \to \infty} X_n(\omega) = x \text{ for } P\text{-almost all } \omega \in \Omega \tag{6.8}$$

A.1.2 - Some important theorems and inequalities from probability theory

We will need a simple version of the law of large numbers for martingale difference sequences (which is a partial result of Stoica [175] theorem 1, see also Rosalsky and Stoica [156], Stoica [176], Wang et al. [191] and Lagodowski and Rychlik [111]):

Lemma 6.0.4 - (**Law of large numbers for martingale difference sequences**)

Let X_n be a real-valued martingale difference sequence with

$$\limsup_{n \to \infty} E[|X_n|^p] < \infty \tag{6.9}$$

for some $p > 1$. Then

$$\lim_{n \to \infty} \frac{\sum_{1 \leq k \leq n} X_k}{n} = 0 \text{ a.s.} \tag{6.10}$$

Another well-known and important theorem about almost sure convergence is the Borel-Cantelli lemma (compare Kallenberg [99], Bauer [22], König [105]):

Lemma 6.0.5 - (**Borel-Cantelli lemma**)

▶ **Lemma 6.0.5.1:** *Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $A_n \in \mathcal{F}$ for every $n \in \mathbb{N}$. Assume that*

$$\sum_{n \in \mathbb{N}} P[A_n] < \infty$$

then the probability to be in infinitely many events, A_n is zero:

$$P[\cap_{n \geq 1} \cup_{k \geq n} A_n] = 0$$

▶ **Lemma 6.0.5.2:** *Let X be a (time-discrete) $\mathbb{R}^n$-valued stochastic process over some probability space, $(\Omega, \mathcal{F}, P)$. Assume that for some $x \in \mathbb{R}^n$:*

$$\sum_{n \in \mathbb{N}} P\left[|X_n - x| \geq \epsilon\right] < \infty \text{ for every } \epsilon > 0 \tag{6.11}$$

Then

$$X_n \to x \text{ a.s.} \tag{6.12}$$

We also need some well-known inequalities:

Lemma 6.0.6 - (**Some probabilistic inequalities**)

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $X : \Omega \to \mathbb{R}$ be a $\mathcal{F}/\mathcal{B}_{\mathbb{R}}$-measurable random variable.

▶ **Lemma 6.0.6.1:** *If X is non-negative and $y > 0$ then Markov's inequality holds*

$$P\left[X \geq y\right] \leq \frac{E\left[X\right]}{y} \tag{6.13}$$

▶ **Lemma 6.0.6.2:** *Let $X \in \mathcal{L}_1(dP)$ and $g : \mathbb{R} \to \mathbb{R}$ be a convex function, then the Jensen's inequality holds true:*

$$g\left(E\left[X\right]\right) \leq E\left[g\left(X\right)\right] \tag{6.14}$$

with equality if and only if $g(X) = a\left(X - E\left[X\right]\right) + g\left(E\left[X\right]\right)$ a.s. for some $a \in \mathbb{R}$ Whenever g is strictly convex equality in Jensen's inequality occurs if and only if X is almost surely constant.

(6.15)

A simple but important consequence of Markov's inequality is an exponential decay of the tail probabilities whenever the characteristic function

$$\phi(t) := E\left[\exp\left(tX\right)\right]$$

of some real-valued random variable X exists in a neighborhood of zero:

$$P\left[X \geq a\right] \leq \frac{E\left[\exp\left(tX\right)\right]}{\exp\left(ta\right)} \tag{6.16}$$

for every $a > 0$. An optimization over t gives:

$$\ln P\left[X \geq a\right] \leq -\mathcal{L}\left[\ln \phi\right](a) \tag{6.17}$$

where

$$\mathcal{L}\left[f\right](y) := \sup\left\{yx - f(x)\,|x \in \mathbb{R}\right\} \tag{6.18}$$

is the Legendre transform. This is actually the essential estimate in the upper bound of Cramer's theorem in large deviation theory (compare König [104], Dembo and Zeitouni [62], Hollander [88], Freĭdlin and Wentzell [73]) and possesses important extensions from $\mathbb{R}$ to locally convex Hausdorff topological vector spaces. We need this estimate for our proof of Lemma 4.1.1 and Lemma 4.1.2.

In section 4.3 of Chapter 4 we need some estimates for the centered moments of multidimensional Gaussian distributions:

Lemma 6.0.7 - (**Bounds on centered moments of a Gaussian distribution**)

▶ **Lemma 6.0.7.1:** *Let X be a M-dimensional Gaussian random variable with mean μ and covariance Σ. Then for any $k > -M$*

$$\begin{aligned} E\left[(\|X-\mu\|_2)^k\right] &\leq \sqrt{2}^k \cdot \frac{\left(\sqrt{\|\Sigma\|_{\text{Op},2}}\right)^{k+M}}{\sqrt{\det(\Sigma)}} \cdot \frac{\Gamma\left(\frac{k+M}{2}\right)}{\Gamma\left(\frac{M}{2}\right)} \\ &\leq \sqrt{2}^k \cdot \left(\sqrt{\|\Sigma\|_{\text{Op},2}}\right)^k \cdot \left(\sqrt{\|\Sigma\|_{\text{Op},2}\,\|\Sigma^{-1}\|_{\text{Op},2}}\right)^M \cdot \frac{\Gamma\left(\frac{k+M}{2}\right)}{\Gamma\left(\frac{M}{2}\right)} \end{aligned} \tag{6.19}$$

where

$$\Gamma(z) := \int_0^\infty x^{z-1}\exp(-x)dx \tag{6.20}$$

is the well-known $\Gamma-$function and

$$\|A\|_{\text{Op},2} = \sup\left\{\sqrt{x^TAA^Tx}\,\middle|x \in \mathbb{R}^M; \|x\|_2 = 1\right\} \tag{6.21}$$

is the operator norm with respect to the Euclidean norm.

Proof. The expression to be estimated is

$$E\left[(\|X-\mu\|_2)^k\right] = \frac{1}{\sqrt{\det(\Sigma)}\sqrt{2\pi}^M}\int_{\mathbb{R}^M}\|x\|^k\exp\left(-\frac{x^T\Sigma^{-1}x}{2}\right)d\nu_{\text{Leb.}}(x) \tag{6.22}$$

The exponential factor in the integrand can be upper bounded by

$$\exp\left(-\frac{x^T\Sigma^{-1}x}{2}\right) \leq \exp\left(-\frac{1}{\|\Sigma\|_{\text{Op},2}}\frac{x^Tx}{2}\right)$$

The calculation of the remaining integral is an exercise in elementary calculus. Observing the spherical symmetry of the integrand, the formula

$$\int_{\mathbb{R}^M}\phi(\|x\|)d\nu_{\text{Leb.}}(x) = A_M\int_0^\infty \phi(r)r^{M-1}dr, \tag{6.23}$$

where

$$A_M := \frac{2\sqrt{\pi}^M}{\Gamma\left(\frac{M}{2}\right)} \tag{6.24}$$

is the volume of the $1-$sphere of dimension $M-1$, reduces the integral to a one dimensional one. Finally the substitution $z := \frac{r^2}{2\|\Sigma\|_{\text{Op},2}}$ brings the integral into the defining integral of the $\Gamma-$function. The second claim in the statement follows from the estimate

$$\det A \geq \left(\frac{1}{\|A^{-1}\|}\right)^M \tag{6.25}$$

valid for every strictly positive matrix, A. The proof shows that both inequalities are equalities whenever Σ is a multiple of the identity matrix. ■

A.1.3 - Essentials on the ergodic theory of Markov chains

A Markov chain can be defined as a special recursively constructible causal model (compare Chapter 1, especially Example 1.4.2). In this section we focus on ergodic properties of time-homogeneous Markov chains. We are mainly interested in finite state spaces but also mention important results for general state spaces. This might be useful to generalize the algorithms discussed in section 4. The definition of ergodicity is used slightly different in the literature about ergodic theory and the literature on Markov chains. We mainly adapt the notation and definitions from the survey article of Diaz-Espinosa [64].

Definition 6.0.1 - (**Markov chains**)

A general state space, time homogeneous Markov chain is a pair $((\mathbf{S}, \mathcal{F}_S), K)$ *where*

- $(\mathbf{S}, \mathcal{F}_S)$ *is a measurable space*
- $K \in \Lambda_{\mathbf{S}}^{\mathbf{S}}$ *is a probability kernel, the transition kernel.*

let $((\mathbf{S}, \mathcal{F}_S), K)$ be a Markov chain. For every initial probability law $\mu \in M_1(\mathcal{F}_S)$ there exists exactly one process law P_μ on $(\mathbf{S}^\infty, \otimes^\infty \mathcal{F}_S)$ that is compatible with the probabilistic structure of the chain, i.e.

- $P_\mu(\{\pi_0 \in A\}) = \mu(A)$ for every $A \in \mathcal{F}_S$
- $P_\mu[\{\pi_{n+1} \in A\} | \pi_n = s, (\pi_k)_{k<n}] = K(s, A)$ a.s.
  for all $n \in \mathbb{N}, A \in \mathcal{F}_S$ and $s \in \mathbf{S}$.

This can be seen as a special case of Theorem 1.2.1. Intuitively $K(s, A)$ describes the probability of landing in $A \in \mathcal{F}_S$ in the next step if the current state is $s \in \mathbf{S}$. Every kernel acts in a natural way on bounded, measurable functions via

$$(K\phi)(s) := \int_{\mathbf{S}} K(s, ds')\, \phi(s') \tag{6.26}$$

the probabilistic interpretation is the following: For any Markov process with transition kernel K:

$$E[\phi(\pi_{n+1}) | \pi_0, \dots \pi_n] = (K\phi)(\pi_n) \text{ a.s.} \tag{6.27}$$

especially $(K\mathbb{1}_A)(s) = K(s, A)$ for every $A \in \mathcal{F}_S$. The adjoint operator [2] acts on measures $\mu \in M_1(\mathcal{F}_S)$ via:

$$(\mu K)(A) = \int_{\mathbf{S}} \mu(ds) K(s, A) \tag{6.28}$$

The probabilistic interpretation is the following: If an initial point is drawn with probability μ then after one time step the probability distribution is given by μK.
A special role is played by invariant measures, invariant functions and functions almost invariant with respect to a given invariant measure:

Definition 6.0.2 - (**Invariant measures, invariant functions**)

► **Definition 6.0.2.1:** *Let* $K \in \Lambda_{(\mathbf{S}, \mathcal{F}_S)}^{(\mathbf{S}, \mathcal{F}_S)}$ *be a probability kernel. A* $\sigma-$*finite measure* $\mu \in M(\mathcal{F}_S)$ *will be called* $K-$ *invariant if*

$$\mu K = \mu \tag{6.29}$$

The $K-$*invariant measures on* $\mathcal{F}_S$ *will be denoted by* $\mathrm{INV}_K(\mathcal{F}_S)$.
► **Definition 6.0.2.2:** *Let* $K \in \Lambda_{(\mathbf{S}, \mathcal{F}_S)}^{(\mathbf{S}, \mathcal{F}_S)}$ *a function* $\phi \in \mathfrak{B}(\mathbf{S}, \mathcal{F}_S)$ *will be called* $K-$*invariant, if*

$$K\phi = \phi \tag{6.30}$$

and almost $K-$*invariant with respect to some invariant measure* μ*, if:*

$$K\phi = \phi \text{ a.s. with respect to } \mu \tag{6.31}$$

► **Definition 6.0.2.3:** *For a given kernel* $K \in \Lambda_{(\mathbf{S}, \mathcal{F}_S)}^{(\mathbf{S}, \mathcal{F}_S)}$ *the invariant* σ*-algebra is:*

$$\mathcal{I}_K := \{A \in \mathcal{F}_S | K\mathbb{1}_A = \mathbb{1}_A\} \tag{6.32}$$

---

[2] The dual space of the bounded measurable functions on $(\mathbf{S}, \mathcal{F}_S)$ equipped with supremum norm (denoted by $\mathfrak{B}(\mathbf{S}, \mathcal{F}_S)$) can be identified with the set of signed, finitely additive measures of finite total variation on $\mathcal{F}_S$ (compare for example Aliprantis and Border [1])

and for a given $K-$invariant measure μ the almost invariant σ-algebra is defined to be:

$$\mathcal{I}_K^{\mu} := \{A \in \mathcal{F}_S \,|K\mathbb{1}_A = \mathbb{1}_A \text{ a.s. with respect to } \mu\} \tag{6.33}$$

▶ **Definition 6.0.2.4:** *Let $K \in \Lambda_{(\mathbf{S},\mathcal{F}_S)}^{(\mathbf{S},\mathcal{F}_S)}$ be a probability kernel and let μ be a K-invariant measure then μ will be called ergodic, if it is $\mathcal{I}_K$-trivial (or equivalently $\mathcal{I}_K^{\mu}$-trivial), i.e.*

$$\mu(A) \in \{0,1\} \text{ for every } A \in \mathcal{I}_K \tag{6.34}$$

the ergodic measures on $\mathcal{F}_S$ will be denoted by $\mathrm{ERG}_K\left(\mathcal{F}_S\right)$

Ergodic measures are necessarily mutually singular and the extremal elements in the set of all invariant measures. Moreover if the state space is Polish, every invariant measure can be decomposed into its ergodic components [3].

Theorem 6.0.1 - (**Ergodic decomposition of invariant measures**)

Let K be a Markov kernel on the Polish space $(\mathbf{S}, \mathcal{F}_S)$ and let μ be a $K-$ invariant probability measure. Then:

$$\mu = \int_{\mathrm{ERG}_K(\mathcal{F}_S)} m dQ(m) \tag{6.35}$$

for some probability measure $Q \in M_1\left(\mathcal{B}_{M_1(\mathcal{F}_S)}\right)$, where $\mathcal{B}_{M_1(\mathcal{F}_S)}$ denotes the Borel $\sigma-$algebra of the weak topology on the set of probability measures on $\mathcal{F}_S$ [4].

A.1.4 - Ergodic theory and singular perturbation of finite state space Markov chains

In this section we will relate the concepts from A.1.3 to the usual terminology for finite state space Markov chains and we will cite some important perturbation results that we need in Chapter 4. So let $\left(\left(\mathbf{S}, 2^{\mathbf{S}}\right), K\right)$ be a Markov chain where $\mathbf{S}$ is a finite set. As before we always equip finite sets with the discrete σ-algebra, $2^{\mathbf{S}}$. For finite state spaces the transition kernel is completely described by the stochastic $\mathbf{S} \times \mathbf{S}$ matrix $P_{a,b} := K(a, \{b\})$, where $a, b \in \mathbf{S}$. Moreover the ergodic structure of the chain can be inferred from the communication graph:

Definition 6.0.3 - (**Communication graph of a finite state Markov chain**)

▶ **Definition 6.0.3.1:** *Let $\left(\left(\mathbf{S}, 2^{\mathbf{S}}\right), K\right)$ be a finite state space Markov chain. The communication graph of the chain, is the directed graph $G_K := (\mathbf{S}, E)$ where*

$$(a, b) \in E \text{ iff } K(a, \{b\}) > 0 \tag{6.36}$$

▶ **Definition 6.0.3.2:** *A subset $A \subseteq \mathbf{S}$ will be called ergodic class, if for every $a, b \in A$: $a \rightsquigarrow b$* and *$b \rightsquigarrow a$ and if there exists no $c \in \mathbf{S} \setminus A$ such that $a \rightsquigarrow c$ for some (then every) $a \in A$. With the terminology from Definition 1.1.2 and the considerations afterwards this is equivalent to saying that an ergodic class is a subset of comparable, $\leq$-maximal elements that is maximal with respect to set-inclusion.*

---

[3] For a proof based on a generalization of Birkhoff's ergodic theorem to arbitrary L_1-$l_\infty-$contractions and a good summary about ergodic properties or Markov operators, see for example: Diaz-Espinosa [64]

[4] By weak topology on $M_1\left(\mathcal{F}_S\right)$ we mean the topology generated by linear functionals $\mu \mapsto E_\mu\left[\phi\right]$ where ϕ is a real-valued bounded, continuous function. Then $\mu \mapsto E_\mu\left[\phi\right]$ is actually measurable for all $\mathcal{B}_S/\mathcal{B}_{\mathbb{R}}$-measurable function, ϕ (compare Diaz-Espinosa [64]). The integral appearing in Theorem 6.0.1 is the Gelfand integral (compare Diaz-Espinosa [64]) on V^* where V is the Banach space of bounded, measurable functions equipped with supremum norm.

The ergodic properties of finite state space Markov chains are summarized by the following theorem:

Theorem 6.0.2 - (**Ergodic properties of finite state space Markov chains**)

Let $\left(\left(\mathbf{S}, 2^{\mathbf{S}}\right), K\right)$ *be a finite state space Markov chain. Then there exists at least one ergodic subset of* $\mathbf{S}$. *Two different ergodic subsets are necessarily disjoint. Let* $A_i \subseteq \mathbf{S}$ *be the collection of all ergodic subsets, where* $1 \leq i \leq k$ *and* $k \leq |\mathbf{S}|$ *is the number of ergodic classes. Then:*

- *For every* $i \in \{1, \ldots, k\}$ *there exists a unique* K*-invariant, positive function* $\phi_i \in \mathbb{R}^{\mathbf{S}}$ *with*
$$\phi_i(a) = \begin{cases} 1 & \text{for } a \in A_i \\ 0 & \text{for } a \in \cup_{1 \leq j \leq k; j \neq i} A_j \end{cases} \tag{6.37}$$
- *These functions form a complete basis of right-Eigenvectors of K for the Eigenvalue* 1, *i.e. the set of all* $K-$*invariant functions is the vector space generated by the set* $\{\phi_i\}_{1 \leq i \leq k}$. *Probabilistically* $\phi_i(a)$ *is the probability to end-up in ergodic class* i *if the starting state is* a.
- *For every* $1 \leq i \leq k$ *there exists a unique* K*-ergodic probability measure,* μ_i*, that is equivalent to the measure*
$$\nu_i(\{a\}) = \begin{cases} 1 & \text{if } a \in A_i \\ 0 & \text{else} \end{cases} \tag{6.38}$$
- *The set of invariant measures is the convex cone generated by* $\{\mu_i\}_{1 \leq i \leq k}$
- *For every probability measure* $\mu \in M_1\left(2^{\mathbf{S}}\right)$*:*
$$\lim_{n \to \infty} \frac{1}{n} \sum_{j=0}^{n-1} \mu K^j = \sum_{1 \leq j \leq k} E_\mu\left[\phi_j\right] \mu_j \tag{6.39}$$
- *If* K *is strictly positive, i.e.* $K(a, \{b\}) > 0$ *for every* $a, b \in \mathbf{S}$ *then* $k = 1$, $A_1 = \mathbf{S}$, $\phi_1 \equiv 1$, *the ergodic measure* μ_1 *has full support and is the only invariant probability measure for* K. *Moreover*
$$\lim_{n \to \infty} \nu K^n = \mu_1 \tag{6.40}$$
for every $\nu \in M_1\left(2^{\mathbf{S}}\right)$. *The speed of convergence is exponential with a rate depending on the second largest Eigenvalue of* K.

Geometrically the set $M_1\left(2^{\mathbf{S}}\right)$ is equivalent to the standard simplex, $\Delta_{0,\mathbf{S}}$, in $\mathbb{R}^{\mathbf{S}}$. The latter is the convex hull of $\{e_s\}_{s \in \mathbf{S}}$. By geometrically equivalence we mean that the isomorphism

$$\mu \mapsto \sum_{s \in \mathbf{S}} \mu\left[\{s\}\right] e_s$$

preserves the topological structure, the differentiable structure and the affine-convex structure. Every element in $\Delta_{0,\mathbf{S}}$ can be written as a convex combination of the extremal points, $\{e_s\}_{s \in \mathbf{S}}$ in a unique way.

A kernel from a finite set $\mathbf{S}_1$ to another finite set $\mathbf{S}_2$ is geometrically equivalent to a $\mathbf{S}_1$-fold product of simplices over $\mathbf{S}_2$, i.e. $\Lambda_{\mathbf{S}_1}^{\mathbf{S}_2}$ is canonically isomorphic to $\left(\Delta_{\mathbf{S}_2}\right)^{\mathbf{S}_1} \subseteq \mathbb{R}^{\mathbf{S}_1 \times \mathbf{S}_2}$.

The interior points of $(\Delta_{\mathbf{S}_2})^{\mathbf{S}_1}$, denoted by $(\Delta_{\mathbf{S}_2})^{\mathbf{S}_1 \circ}$, consists of all strictly positive kernels. By Theorem 6.0.2 the set-valued map

$$\mathrm{INV} : (\Delta_{\mathbf{S}})^{\mathbf{S}} \to 2^{M_1(\mathcal{F}_S)} \,;\, K \mapsto \{\mu \in M_1\left(\mathcal{F}_S\right) | \mu K = \mu\} \tag{6.41}$$

has compact, convex values and is single-valued on $(\Delta_{\mathbf{S}_2})^{\mathbf{S}_1 \circ}$. By Example 2.2.1 the map INV is upper semicontinuous (compare Definition 2.2.2). Moreover it is even analytic in the interior of $(\Delta_{\mathbf{S}_2})^{\mathbf{S}_1}$. This is a consequence of well-known perturbation results for finite state space Markov chains (compare Schweitzer [168], Schweitzer [167],Simon and Ando [171], Hassin and Haviv [84], Eugene and Feinberg [71] and references therein):

Theorem 6.0.3 - (**Perturbation of finite state space Markov chains**)

▶ **Theorem 6.0.3.1:** *Let* $K \in (\Delta_{\mathbf{S}})^{\mathbf{S}} \subset \mathbb{R}^{\mathbf{S}\times\mathbf{S}}$ *then the tangent cone (compare Example 2.1.2) of* $(\Delta_{\mathbf{S}_2})^{\mathbf{S}_1}$ *at point* K *is given by:*

$$T_{(\Delta_{\mathbf{S}})^{\mathbf{S}}}(K) = \left\{ C \in \mathbb{R}^{\mathbf{S}\times\mathbf{S}} \,\middle|\, \sum_{s' \in \mathbf{S}} C_{s,s'} = 0 \text{ for every } s \in \mathbf{S}; C_{s,s'} \geq 0 \text{ whenever } K_{s,s'} = 0 \right\} \tag{6.42}$$

If K *is strictly positive then the tangent cone becomes a vector space (what is clear since the convex-interior of* $(\Delta_{\mathbf{S}})^{\mathbf{S}}$ *is an analytic submanifold of* $\mathbb{R}^{\mathbf{S}\times\mathbf{S}}$*).*

▶ **Theorem 6.0.3.2:** *The map* INV *from Eq. 6.41 restricted to the interior of* $(\Delta_{\mathbf{S}})^{\mathbf{S}}$ *is analytic. For any strictly positive* K $(\Delta_{\mathbf{S}})^{\mathbf{S}}$*, any* $C \in T_K(\Delta_{\mathbf{S}})^{\mathbf{S}}$ *and any sufficiently small* $\epsilon > 0$*, the matrix* $K_\epsilon := K + \epsilon C$ *belongs to the interior of* $(\Delta_{\mathbf{S}})^{\mathbf{S}}$ *and:*

$$\mathrm{INV}(K_\epsilon) = \mathrm{INV}(K)\left(\mathbb{1}_{\mathbf{S}} - \epsilon U\right)^{-1} \tag{6.43}$$

where

$$U = CY_K \tag{6.44}$$

where Y_K *is the deviation matrix of* K*:*

$$Y_K = \left(\mathbb{1} - K + K^*\right)^{-1} - K^* \tag{6.45}$$

and

$$K^* = \lim_{n\to\infty} K^n = \underline{1}\,\mathrm{INV}(K)^T \tag{6.46}$$

where v^T *is the transpose of* v *(we assume all vectors to be column vectors) and* $\underline{1}$ *is the vector whose entries are all equal to* 1.

▶ **Theorem 6.0.3.3:** *Let* $K \in \partial(\Delta_{\mathbf{S}})^{\mathbf{S}}$ *and let* $C \in T_K(\Delta_{\mathbf{S}})^{\mathbf{S}}$ *such that* $C_{s,s'} > 0$ *whenever* $K_{s,s'} = 0$ *then for sufficiently small* $\epsilon > 0$ *the matrix* $K_\epsilon := K + \epsilon C$ *belongs to the interior of* $(\Delta_{\mathbf{S}})^{\mathbf{S}}$*. Then* Y_{K_ϵ} *posses a Laurent series expansion with a non-essential pole of order* s *for some* $s \geq 0$*:*

$$Y_{K_\epsilon} = \sum_{k=-s}^{\infty} Y^{(k)} \epsilon^k \tag{6.47}$$

and

$$\mathrm{INV}(K_\epsilon) = \pi_0 \left(\mathbb{1} - U\right)^{-1} \tag{6.48}$$

where $\pi_0 K = \pi_0$ *and* $U = CY^{(0)}$.

A.2 - Gradients on the probability simplex and some calculus

In this section we will provide some definitions and formulas that are needed in Chapter 3 and Chapter 4.

A.2.1 - Technical lemma on the submultiplicativity of the trace function on positive matrices

First of all we provide and prove a well-known lemma that we need in section 4.3.

Lemma 6.0.8 - (**Submultipicativity of the trace of positive matrices**)

Let $A, B \in \mathbb{R}^{N\times N}$ *be positive matrices, i.e.* $A = A^T$*,* $B = B^T$ *and* $x^T Ax \geq 0$*,* $x^T Bx \geq 0$ *for all* $x \in \mathbb{R}^N$*. Then*

$$\left|\mathrm{Tr}(AB)\right| \leq \left|\mathrm{Tr}(A)\right| \cdot \left|\mathrm{Tr}(B)\right| \tag{6.49}$$

Proof. By the spectral theorem for symmetric matrices there exists some orthogonal projectors P_i of rank one and some positive numbers $\lambda_i \geq 0$ such that

$$A = \sum_{i=1}^{N} \lambda_i P_i \tag{6.50}$$

and some orthogonal projectors $\tilde{P}_i$ together with some real numbers $\mu_i \geq 0$ such that

$$B = \sum_{i=1}^{N} \mu_i \hat{P}_i \tag{6.51}$$

Then

$$|\mathrm{Tr}\,(AB)| = \left| \sum_{i=1}^{N} \sum_{j=1}^{N} \lambda_i \mu_j \,\mathrm{Tr}\left(P_i \hat{P}_j\right) \right| \tag{6.52}$$

since $\left|\mathrm{Tr}(P_i \hat{P}_j)\right| \leq 1$ for any pair of othorgonal projectors of rank one this implies

$$|\mathrm{Tr}\,(AB)| \leq \sum_{i=1}^{N} \sum_{j=1}^{N} |\lambda_i \mu_j| = |\mathrm{Tr}\,(A)| \cdot |\mathrm{Tr}\,(B)| \tag{6.53}$$

■

A.2.2 - Gradients on $(\Delta_{\mathbf{A}})^{\mathbf{S}}$

For a finite state space MDP (compare Assumption 4.1.1 in Chapter 4) consider some policy functional

$$\phi : (\Delta_{\mathbf{S}})^{\mathbf{S}\times\mathbf{A}^\circ} \times (\Delta_{\mathbf{A}})^{\mathbf{S}^\circ} \to \mathbb{R} \tag{6.54}$$

We assume that that $\phi(q,\cdot)$ is continuously differentiable for every $q \in (\Delta_{\mathbf{S}})^{\mathbf{S}\times\mathbf{A}^\circ}$. For a gradient ascent algorithm it is necessary to calculate the derivative of $\phi\,(q,\cdot)$ into direction $v \in T_{(\Delta_{\mathbf{A}})^{\mathbf{S}^\circ}}(z)$:

$$D_2\phi(q,z)\,[v] := \lim_{h\to 0} \frac{\phi\,(q, z+hv) - \phi\,(q,z)}{h} \tag{6.55}$$

We will sometimes also write

$$\frac{\partial \phi}{\partial v}(q,z) := D_2\phi(q,z)\,[v] \tag{6.56}$$

The directional derivative is a base independent concept. If $\{v_i\}_{1\leq i\leq |\mathbf{S}|\cdot(|\mathbf{A}|-1)}$ is a basis of $T_{(\Delta_{\mathbf{A}})^{\mathbf{S}}}\,(z)$ (compare Eq. 6.42), then an expansion with respect to this basis is:

$$D_2\phi(q,z) = \sum_i \frac{\partial \phi}{\partial v_i}(q,z){v_i}^* \tag{6.57}$$

where ${v_i}^*$ is the dual basis of v_i. A (smooth) map that maps a given point, $p \in M$ (where most generally M can be any differentiable manifold) to an $\dim(M)$-tuple of basis vectors in the tangent space at p, T_pM, is called frame in differential geometry and plays a special role in general relativity theory where it encodes the local reference frames of an observer. A special instance of frames are the ones generated by a specific coordinate system for example.

The manifold $M := (\Delta_{\mathbf{A}})^{\mathbf{S}^\circ}$ is a (analytic) submanifold of $\mathbb{R}^{\mathbf{S}\times\mathbf{A}}$. So the tangent space of T_pM can be canonically identified with a subspace of $\mathbb{R}^{\mathbf{S}\times\mathbf{A}}$. Since M is an open submanifold of an affine subspace of $\mathbb{R}^{\mathbf{S}\times\mathbf{A}}$, the tangent space is point independent (compare 6.42):

$$T_pM = \left\{ C \in \mathbb{R}^{\mathbf{S}\times\mathbf{A}} \,\middle|\, \sum_{a\in\mathbf{A}} C_{s,a} = 0 \right\} \tag{6.58}$$

For a general metric,

$$g : M \to \text{bil}\left(\mathbb{R}^{\mathbf{S}\times\mathbf{A}}\right) \ ; g(p) \text{ is symmetric and positive definite} \tag{6.59}$$

the gradient of a function is defined by the following two requirements:

Definition 6.0.4 - (**Definition of gradient on a submanifold of** $\mathbb{R}^{\mathbf{S}\times\mathbf{A}}$)

Let

$$g : M \to \text{bil}\left(\mathbb{R}^{\mathbf{S}\times\mathbf{A}}\right) \ ; g(p) \text{ is symmetric and positive definite} \tag{6.60}$$

be a metric on M and let $\psi \in C_1(M, \mathbb{R})$ be a differentiable function. Then the g-gradient of ψ at a given point p is the unique vector $\nabla_g\psi(p) \in \mathbb{R}^{\mathbf{S}\times\mathbf{A}}$ that satisfies

1. $g(\nabla_g\psi(p), v) = D\psi(p)[v]$ *for every* $v \in T_pM$
2. $\nabla_g\psi(p) \in T_pM$

Consider the Euclidean metric at $z \in M$:

$$g_{E,z}(X, Y) := \langle X, Y\rangle = \sum_{s\in\mathbf{S}, a\in\mathbf{A}} X_{s,a}Y_{s,a} \tag{6.61}$$

Obviously

$$G_E(z)_{s,a} := D\psi(z)[e_{s,a}]$$

satisfies the first requirement in Definition 6.0.4 but fails to satisfy the second one. However $G_E(z) + n$ still satisfies the first requirement whenever n is perpendicular to T_pM and an appropriate choice of n yields a vector that satisfies the second requirement. A short calculation gives:

$$n \perp\!\!\!\perp_{g_E} T_pM \text{ iff } n = \sum_{s\in S} \lambda_s \sum_{a\in A} e_{s,a} \text{ where } \lambda_s \in \mathbb{R} \tag{6.62}$$

and the constraint 6.58 yields

$$\lambda_s = -\frac{\sum_{a\in\mathbf{A}} G_E(p)_{s,a}}{|\mathbf{A}|}.$$

Therefore the Euclidean gradient of ψ is:

$$\nabla_E\psi(z) = \sum_{s\in\mathbf{S}, a\in\mathbf{A}} D\psi(z) \left[e_{s,a} - \frac{1}{|\mathbf{A}|} \sum_{a'\in\mathbf{A}e_{s,a'}} \right] e_{s,a} \tag{6.63}$$

For the Fisher gradient consider the Fisher metric at $z \in M$ (compare A.3):

$$g_{F,z}(X, Y) = \sum_{s\in\mathbf{S}, a\in\mathbf{A}} \frac{X_{s,a}Y_{s,a}}{z_{s,a}} \tag{6.64}$$

The first requirement in Definition 6.0.4 is satisfied by

$$G_F(z)_{s,a} := z_{s,a} D\psi(z)[e_{s,a}]$$

and

$$n \perp\!\!\!\perp_{g_F} T_pM \text{ iff } n = \sum_{s\in S} \lambda_s \sum_{a\in A} z_{s,a} e_{s,a} \text{ where } \lambda_s \in \mathbb{R} \tag{6.65}$$

such that $G_F(z) + n$ satisfies the first and the second defining identity for the gradient whenever

$$\lambda_s = -\sum_{a\in \mathbf{A}} z_{s,a} G_F(z)_{s,a}$$

and therefore the Fisher gradient of ψ is

$$\nabla_F \psi(z) = \sum_{s\in \mathbf{S}, a\in \mathbf{A}} z_{s,a} D\psi(z) \left[e_{s,a} - \sum_{a'\in \mathbf{A}} z_{s,a'} e_{s,a'} \right] e_{s,a} \tag{6.66}$$

Therefore the Euclidean policy gradient of the policy functional is:

$$\nabla_E \phi(q,z)_{s,a} := D_2\phi(q,z) \left[e_{s,a} - \frac{1}{|A|} \sum_{a'\in \mathbf{A}} e_{s,a'} \right] \tag{6.67}$$

Whenever $\phi(q,\cdot)$ can be extended to an open neighborhood of M in $\mathbb{R}^{\mathbf{S}\times\mathbf{A}}$ then Eq. 6.67 can be rewritten as

$$\nabla_E \phi(q,z)_{s,a} := \frac{\partial \phi}{\partial z_{s,a}}(q,z) - \frac{1}{|\mathbf{A}|} \sum_{a'\in \mathbf{A}} \frac{\partial \phi}{\partial z_{s,a}}(q,z)\,. \tag{6.68}$$

This equation is usually more convenient for computations. Eq. 6.67 shows that an extension of the original function to $\mathbb{R}^{\mathbf{S}\times\mathbf{A}}$ is superfluous and that the gradient is completely determined by the vectors tangent to the manifold $(\Delta_{\mathbf{S}})^{\mathbf{S}\times\mathbf{A}^\circ}$. We will switch between both formulations and use the one that is most appropriate for the desired target. The Fisher gradient becomes:

$$\nabla_F \phi(q,z)_{s,a} := z_{s,a} D_2\phi(q,z) \left[e_{s,a} - \sum_{a'\in \mathbf{A}} z_{s,a'} e_{s,a'} \right] \tag{6.69}$$

A.2.3 - Directional derivatives of holomorphic functions of matrices

We start with a lemma on directional derivatives of inverses of matrices and polynomial functions of matrices:

Lemma 6.0.9 - **(Some calculus for functions of matrices)**

Let

$$f(A) := (\mathbb{1} - A)^{-1}$$

then

$$Df(A)[B] = (\mathbb{1} - A)^{-1} B (\mathbb{1} - A)^{-1} \tag{6.70}$$

Let

$$g(A) := A^n,$$

with $n \in \mathbb{N}$ *then*

$$Dg(A)[B] = \sum_{k=0}^{n-1} A^k B A^{n-1-k} \tag{6.71}$$

Now we will address the more general problem of calculating the directional derivative:

$$Df(A)[B] := \lim_{h\to 0} \frac{f(A+hB) - f(A)}{h} \tag{6.72}$$

where $A, B \in M(n)$, $f : \mathbb{C} \to \mathbb{C}$ is holomorphic in a neighborhood of the spectrum, $\sigma(A)$, of A and $f(A)$ is defined via the Banach-space holomorphic functional calculus applied to the $n \times n-$ matrices [5]:

$$f(A) = \frac{1}{2\pi \mathrm{i}} \oint_{\lambda} f(z) \left(z\mathbb{1} - A\right)^{-1} dz$$

where λ denotes an arbitrary Jordan curve with the property that it surrounds the spectrum of A. If A is invertible a short calculation gives:

$$\frac{d}{dh}(A + hB)^{-1}\,|_{h=0} = -B^{-1}AB^{-1} \tag{6.73}$$

Since the zeros of a polynomial depend continuously on the coefficients (compare for example Horn and Johnson [89], Bhatia [26]) $\sigma(A + hB)$ lies also in the interior of λ for sufficiently small h. This reduces the problem to calculating:

$$\frac{d}{dh} f(A + hB)\,|_{h=0} = \frac{d}{dh}\frac{1}{2\pi \mathrm{i}} \oint_{\lambda} f(z) \left(z\mathbb{1} - A - hB\right)^{-1} dz\,|_{h=0}$$

Since for small parameters the integrand is uniformly bounded on the range of λ for all sufficiently small h, Lebesgue's theorem can be applied to exchange differentiation and integration:

$$\begin{aligned} \frac{d}{dh} f(A + hB)\,|_{h=0} &= \frac{1}{2\pi \mathrm{i}} \oint_{\lambda} f(z) \frac{d}{dh}\left(z\mathbb{1} - A - hB\right)^{-1}|_{h=0}\, dz \\ &= \frac{1}{2\pi \mathrm{i}} \oint_{\lambda} f(z) \left(z\mathbb{1} - A\right)^{-1} B \left(z\mathbb{1} - A\right)^{-1} dz \end{aligned} \tag{6.74}$$

In the last step we used 6.73 and the chain rule. Assume that A is diagonalizable, i.e.

$$A = \sum_{k=1}^{r} \lambda_k P_k, \text{ where } 1 \le r \le n \tag{6.75}$$

where the Eigenvalues λ_k are pairwise disjoint and the projection operators onto the corresponding Eigenspaces satisfy

$$P_i P_j = \delta_{i,j} P_i. \tag{6.76}$$

If at least one Eigenvalue is degenerate we have $r < n$. In any case no Eigenvalue lies in the range of λ by assumption and therefore

$$\left(z\mathbb{1} - A\right)^{-1} = \sum_k \left(z - \lambda_k\right)^{-1} P_k \tag{6.77}$$

for all z in the range of λ. Therefore:

$$\begin{aligned} Df(A)[B] &= \sum_{i,j} \frac{1}{2\pi \mathrm{i}} \oint_{\lambda} f(z) \frac{1}{(z - \lambda_i)(z - \lambda_j)} P_i B P_j dz \\ &= \sum_i f'(\lambda_i) P_i B P_i + \sum_{i \ne j} \frac{f(\lambda_i) - f(\lambda_j)}{\lambda_i - \lambda_j} P_i B P_j \end{aligned} \tag{6.78}$$

In the last step we used Cauchy's integral formula.
If A is not diagonalizable it can be written as $A = D + N$ where D is diagonalizable, N is nilpotent and $ND = DN$. The Neumann series for matrix inversion yields:

$$\left(z\mathbb{1} - A\right)^{-1} = \sum_{k=0}^{n-1} N^k \left(z\mathbb{1} - D\right)^{-(k+1)} \tag{6.79}$$

---

[5] This class of functions obviously includes everywhere holomorphic functions like polynomials or exponentials (in which case the holomorphic functional calculus is just a power series of the underlying matrix) as well as more complicated functions like the logarithms or the $n-$th root (Both functions can be defined on an appropriate splitted complex plane $\mathbb{C} \setminus \mathbb{R}_{\ge 0} z_0$, where $z_0 \in \mathbb{C} \setminus 0$ can be chosen arbitrarily - a common value is $z = -1$)

writing

$$D = \sum_{k=1}^{r} \lambda_k P_k, \text{ where } 1 \leq r \leq n \tag{6.80}$$

again gives:

$$Df(A)\,[B] = \sum_{i,j=1}^{r} \sum_{k,l=0}^{n-1} \frac{1}{2\pi \mathrm{i}} \oint_{\lambda} f(z) \frac{1}{(z-\lambda_i)^{k+1}(z-\lambda_j)^{l+1}} N^k P_i B P_j N^l dz \tag{6.81}$$

This can again be splitted into a diagonal part (arising from the summation over pairs (i, j) with $i = j$) and a cross part. Using Cauchy's integral formula again yields

$$Df(A)\,[B]_{\text{diag}} = \sum_{i=1}^{r} \sum_{k,l=0}^{n-1} \frac{f^{(k+l+1)}(\lambda_i)}{(k+l+1)!} N^k P_i B P_i N^l. \tag{6.82}$$

The cross-part is more involved. A calculation using Cauchy's formula again gives:

$$Df(A)\,[B]_{\text{cross}} = \sum_{i \neq j}^{r} \sum_{k,l=0}^{n-1} \left(\alpha\,[l,k,i,j] + \alpha\,[k,l,j,i]\right) N^k P_i B P_j N^l \tag{6.83}$$

where

$$\alpha\,[l,k,i,j] = \sum_{r=0}^{k} \begin{pmatrix} l+r \\ r \end{pmatrix} \frac{(-1)^r}{(\lambda_i - \lambda_j)^{l+r+1}} \frac{f^{(k-r)}(\lambda_i)}{(k-r)!} \tag{6.84}$$

A.3 - Information theory and information geometry

In this section we will give a short overview about some concepts from information theory. The definitions, interpretations, the background and applications can be found in Shannon's original paper (Shannon [170]) and the books Liese and Vajda [118], Amari, Nagaoka, and Harada [5], Csiszár and Korner [58] and Cover and Thomas [54] for example.

We present the relevant quantities for finite state spaces first and we will continue with the general case afterwards.

Definition 6.0.5 - (**Entropy related quantities for finite state spaces**)

▶ **Definition 6.0.5.1:** *Let* $f : [0, \infty) \to \mathbb{R}$ *be a convex function. Define*

$$m_\infty := \lim_{r\to\infty} \frac{f(r)}{r}$$

(hence $m_\infty \in \overline{\mathbb{R}} \setminus \{-\infty\}$*). Let* μ, ν *be measures on* $2^{\mathbf{S}}$ *where* $\mathbf{S}$ *is a finite set. Then*

$$D_f\left(\mu \,\|\nu\right) := \sum_{s\in\mathbf{S};\nu(s)\neq 0} f\left(\frac{\mu(s)}{\nu(s)}\right)\nu(s) + m_\infty \mu\left[\{s\in\mathbf{S}\,|\nu(s)=0\}\right] \tag{6.85}$$

will be called f*-divergence of* μ *and* ν *(in this formula we set* $0\cdot\infty := 0$*). If* $g(x) = x\ln(x)$ *we just write*

$$D\left(\mu \,\|\nu\right) \text{ for } D_g\left(\mu \,\|\nu\right) \tag{6.86}$$

▶ **Definition 6.0.5.2:** *Let* $\mathbf{S}$ *be a finite set the entropy of a measure* $p \in M_1\left(2^{\mathbf{S}}\right)$ *is*

$$H(p) := -D\left(p \,\|\nu\right) \tag{6.87}$$

where ν *is the counting measure* $\nu(A) := |A|$ *for every* $A \subseteq \mathbf{S}$*. Frequently the entropy is defined with the logarithm with base* 2 *instead of base* e *- both expressions differ by a factor of* $\ln(2)$.

▶ **Definition 6.0.5.3:** *Let* $\mathbf{S}$ *be a finite set* $p, q \in M_1\left(2^{\mathbf{S}_1}\right)$ *then the Kullback-Leibler divergence of p and q is:*

$$D_{\mathrm{KL}}\left(p \,\|q\right) := D\left(p \,\|q\right) \tag{6.88}$$

▶ **Definition 6.0.5.4:** *Let* $\mathbf{S}_1$*,* $\mathbf{S}_2$ *be a finite sets, let* $p \in M_1\left(2^{\mathbf{S}_1\times\mathbf{S}_2}\right)$ *and let* π_i *denote the projections of* $\mathbf{S}_1 \times \mathbf{S}_2$ *onto* $\mathbf{S}_i$ *(where* $i \in \{1, 2\}$*) then the mutual of* π_1 *and* π_2 *under p is:*

$$I_p\left(\pi_1 \,\|\pi_2\right) := D_{\mathrm{KL}}\left(p \,\|\pi_{1*}p \otimes \pi_{2*}p\right) \tag{6.89}$$

where $p_1 \otimes p_2$ *denotes the product measure of* p_1 *and* p_2*. If the distribution of the pair* (π_1, π_2) *is clear we will skip the subscript, p.*

The extension to non-discrete state spaces requires a further technical assumption on the underlying measures. A measure μ on some measurable space, $(\Omega, \mathcal{F})$, is σ-finite if there exist $\Omega_n \in \mathcal{F}$ with $\Omega_n \subseteq \Omega_{n+1}$ for every $n \in \mathbb{N}$, with $\cup_{n\in\mathbb{N}}\Omega_n = \mathcal{F}$ and $\mu\left[\Omega_n\right] < \infty$. An important theorem for σ finite measures is the existence of a Radon-Nikodym derivative, i.e. whenever $\nu << \mu$ in the sense that $\mu(A) = 0$ implies $\nu(A) = 0$ for every $A \in \mathcal{F}$, then there exists a $\mathcal{F}/\mathcal{B}_{\mathbb{R}}$ measurable function p such that

$$\nu(A) = \int_A p(\omega)\mu(d\omega) \text{ for every } A \in \mathcal{F} \tag{6.90}$$

this function, p, will be denoted by $\frac{d\nu}{d\mu}$. The following definition of f-divergence originates from Liese and Vajda [118]:

Definition 6.0.6 - (**Entropy related quantities for general σ-finite measures**)

► **Definition 6.0.6.1:** *Let $(\Omega, \mathcal{F})$ be a measurable space, let $f : [0, \infty) \to \mathbb{R}$ be a convex function. Define $m_\infty := \lim_{r\to\infty} \frac{f(r)}{r}$. Let μ, ν be σ-finite measures on $\mathcal{F}$ and let ρ be a measure that dominates both, μ and ν. Set $p := \frac{d\mu}{d\rho}$ and $q := \frac{d\nu}{d\rho}$. Then*[6]

$$D_f(\mu \,\|\nu) := \int_{\{q>0\}} f\left(\frac{p(\omega)}{q(\omega)}\right) \nu(d\omega) + m_\infty \mu\left[\{q = 0\}\right] \tag{6.91}$$

will be called f-divergence of μ and ν whenever this quantity exists (in this formula we set $0 \cdot \infty := 0$ again). If $g(x) = x \ln(x)$ we just write

$$D(\mu \,\|\nu) \text{ for } D_g(\mu \,\|\nu) \tag{6.92}$$

► **Definition 6.0.6.2:** *Let $(\Omega, \mathcal{F})$ be a measurable space, let η be a σ-finite measure on Ω and let $P \in M_1(\mathcal{F})$ then the generalized η-entropy of P is*

$$H_\eta(P) := -D(P \,\|\eta) \tag{6.93}$$

In the special case that $(\Omega, \mathcal{F}) = (\mathbb{R}^n, \mathcal{B}_{\mathbb{R}^n})$ and η being the Lebesgue measure, this quantity is known as differential entropy.

► **Definition 6.0.6.3:** *Let $(\Omega, \mathcal{F})$ be a measurable space and let $p, q \in M_1(\mathcal{F})$ then the Kullback-Leibler divergence of p and q is:*

$$D_{\text{KL}}(p \,\|q) := D(p \,\|q) \tag{6.94}$$

► **Definition 6.0.6.4:** *Let $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$, let $p \in M_1(\mathcal{F}_1 \otimes \mathcal{F}_2)$ and let π_i denote the projections of $\Omega_1 \times \Omega_2$ onto Ω_i (where $i \in \{1, 2\}$). Then the mutual of π_1 and π_2 under p is:*

$$I_p(\pi_1 \,\|\pi_2) := D_{\text{KL}}(p \,\|\pi_{1*}p \otimes \pi_{2*}p) \tag{6.95}$$

where again $p_1 \otimes p_2$ denotes the product measure of p_1 and p_2. If the distribution of (π_1, π_2) is clear we will skip the subscript, p.

A useful bound on the value of f-divergences is given by the following lemma (a slight generalization of remark 1.2 in first chapter of Liese and Vajda [118])

Lemma 6.0.10 - (**Lower bounds on f-divergences**)

Let $f : [0, \infty) \to \mathbb{R}$ be convex and let $\partial f(x_0)$ denote the set of subderivatives of f at x_0:

$$\lambda \in \partial f(x_0) \text{ iff } f(x) \geq f(x_0) + \lambda \cdot (x - x_0) \text{ for all } x \in \mathbb{R}_{\geq 0}$$

Then if

- *μ and ν are finite or*
- *ν is finite and there exists some point $x_0 \in \mathbb{R}_{\geq 0}$ and $\lambda \in \partial f(x_0)$ such that $\lambda \geq 0$ or*
- *μ is finite and there exists some point $x_0 \in \mathbb{R}_{>0}$ and $\lambda \in \partial f(x_0)$ such that $f(x_0) - \lambda x_0 \geq 0$ or $f(0) \geq 0$ and $\partial f(0) \cap \mathbb{R}$ is not empty.*

---

[6] As the definition indicates neither the existence of the f-divergence nor its particular value depends on the dominating measure, ρ, a convenient choice is $\rho := \nu + \mu$ for example. A proof of this statement can be found in Liese and Vajda [118]

then $D_f(\mu\,\|\nu)$ exists and the estimate

$$D_f(\mu\,\|\nu) \geq \sup\left\{\nu(\Omega)\left(f(x_0) - \lambda\cdot x_0\right) + \lambda\cdot\mu(\Omega)\,|x_0 \in \mathbb{R}_{\geq 0}, \lambda \in \partial f(x_0)\right\} \quad (6.96)$$

holds true[7].

Proof. Set

$$\tilde{f}(u,v) := \begin{cases} vf\left(\frac{u}{v}\right) & \text{for } v > 0 \\ u\cdot m_\infty & \text{else} \end{cases} \quad (6.97)$$

where we set $m_\infty \cdot 0 := 0$ if $m_\infty = \infty$ (compare Liese and Vajda [118], page 212). Then

$$D_f(\mu\,\|\nu) = \int \tilde{f}\left(\frac{d\mu}{d\rho}, \frac{d\nu}{d\rho}\right) d\rho \quad (6.98)$$

for $v > 0$

$$\tilde{f}(p,q) \geq q\left(f(x_0) - \lambda x_0\right) + \lambda p \quad (6.99)$$

for every subderivative, $\lambda \in \partial f(x_0)$ by definition. For $v = 0$ and $t > 0$ we have by definition of the subderivative:

$$\frac{f(x_0 + t) - f(x_0)}{t} \geq \lambda \text{ for every } \lambda \in \partial f(x_0) \quad (6.100)$$

such that the limit $t \to \infty$ gives $m_\infty \geq \lambda$ and therefore Eq 6.99 is also valid for $v = 0$. Inserting this estimate for $\tilde{f}$ into Eq 6.98 gives the desired result. ■

In a similar way we can prove an upper bound for the f-divergence (compare Liese and Vajda [118], page 10).

Lemma 6.0.11 - (**Upper bounds on f-divergences**)

Let $f : [0,\infty) \to \mathbb{R}$ be convex, let $(\Omega, \mathcal{F})$ be a measurable space. Let μ and ν be two σ-finite measures on $\mathcal{F}$ dominated by the σ-finite measure, ρ. Set

$$p := \frac{d\mu}{d\rho} \text{ and } q := \frac{d\nu}{d\rho}$$

Then

$$D_f(\mu\,\|\nu) \leq \nu[\Omega]\,f(0) + \mu[\Omega]\,m_\infty \quad (6.101)$$

If $m < \frac{p}{q} \leq M$ for some $m \in \mathbb{R}_{\geq 0}$ and $M \in \overline{\mathbb{R}}_{\geq 0}$ a.e. with respect to ρ. Then:

$$\begin{aligned} D_f(\mu\,\|\nu) \quad &\leq \quad \nu[\Omega]\left(f(m) - m\frac{f(M) - f(m)}{M - m}\right) + \\ &+\mu[\{q > 0\}]\,\frac{f(M) - f(m)}{M - m} + m_\infty\mu[\{q = 0\}] \end{aligned} \quad (6.102)$$

Proof. By convexity of f for every $m < x \leq M$:

$$f(x) - f(m) \leq \frac{f(M) - f(m)}{M - m}(x - m) \quad (6.103)$$

this gives an appropriate estimate for $\tilde{f}$ (compare proof of lower bound, Lemma 6.0.10) on $\{v > 0\}$. Then an integration gives the desired result again. ■

Next we will quote an important property of f-divergences. The proof of the first part can be found in the first chapter of Liese and Vajda [118] again.

---

[7] This implies positivity of the KL-divergence of two probability measures and positivity of the mutual information. It also gives an upper bound for the generalized entropy whenever the reference measure, η, is finite.

Lemma 6.0.12 - (**Monotonicity of the f-divergence for probability measures**)

► **Lemma 6.0.12.1:** *Let $(\mathbf{X}, \mathcal{F}_X)$ and $(\mathbf{Y}, \mathcal{F}_Y)$ be measurable spaces, let $P, Q \in M_1(\mathcal{F}_X)$ and let $K \in \Lambda_{\mathbf{X}}^{\mathbf{Y}}$. For $p \in M_1(\mathcal{F}_X)$ define $K \otimes p \in M_1(\mathcal{F}_X \otimes \mathcal{F}_Y)$ via*

$$(K \otimes p)(A) = \int p(dx) K(x, dy) \mathbb{1}_A((x, y)) \tag{6.104}$$

and $K * p \in M_1(\mathcal{F}_Y)$ via

$$(K * p)(A) := \int_{\mathbf{X}} p(dx) K(x, A) = K \otimes p(\mathbf{X} \times A) \tag{6.105}$$

Let $f : [0, \infty) \to \mathbb{R}$ be convex. Then

$$D_f(K * P \| K * Q) \le D_f(P \| Q) \tag{6.106}$$

with equality if K is sufficient for P and Q, by this we mean that for every $A \in \mathcal{F}_X$ there exists a $\mathcal{F}_Y/\mathcal{B}_{[0,1]}$ measurable function, ϕ_A, with the property that

$$(K \otimes P)[\pi_1 \in A | \pi_2] = \phi_A(\pi_2) \text{ a.s.} \tag{6.107}$$

and

$$(K \otimes Q)[\pi_1 \in A | \pi_2] = \phi_A(\pi_2) \text{ a.s.} \tag{6.108}$$

where $\pi_1 : \mathbf{X} \times \mathbf{Y} \to \mathbf{X}$ is the canonical projection onto the first factor and $\pi_2 : \mathbf{X} \times \mathbf{Y} \to \mathbf{Y}$ is the projection onto the second factor.

If f is strictly convex then equality in Eq 6.106 implies that K is sufficient for P and Q.

► **Lemma 6.0.12.2:** *For every $i \in \{1, 2, 3\}$ let $(\mathbf{X}_i, \mathcal{F}_i)$ be a measurable space. Let $P \in M_1(\otimes_1^3 \mathcal{F}_i)$ and let $\pi_i : \mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3 \to \mathbf{X}_i$ denote the canonical projection. then*

$$D_f(P \| (\pi_1, \pi_2)_* P \otimes \pi_{3*} P) \ge D_f((\pi_2, \pi_3)_* P \| \pi_{2*} P \otimes \pi_{3*} P) \tag{6.109}$$

with equality if

$$\pi_1 \perp\!\!\!\perp \pi_3 | \pi_2 \text{ under } P \tag{6.110}$$

Whenever f is strictly convex then equality in Eq. 6.109 implies conditional independence, Eq. 6.110 [8].

Proof. As already stated, the first proof can be found in the first chapter of Liese and Vajda [118]. The second statement is a spcical case of the first one. Indeed define $K \in \Lambda_{\mathbf{X}_1 \times \mathbf{X}_2 \times \mathbf{X}_3}^{\mathbf{X}_2 \times \mathbf{X}_3}$ via:

$$K[(x, y, z), A] := \delta_{y,z}[A] \tag{6.111}$$

and set $Q := (\pi_1, \pi_2)_* P \otimes \pi_{3*} P$. Then

$$K * P = (\pi_2, \pi_3)_* P \text{ and } K * Q = \pi_{2*} P \otimes \pi_{3*} P \tag{6.112}$$

such that Eq. 6.109 follows. Moreover:

$$\begin{aligned} & (K \otimes P)[\{\pi_1 \in A\} \cap \{\pi_2 \in B\} \cap \{\pi_3 \in C\} | \pi_2, \pi_3] \\ = \; & \mathbb{1}_B(\pi_2) \mathbb{1}_C(\pi_3) P[\{\pi_1 \in C\} | \pi_2, \pi_3] \end{aligned}$$

and

$$\begin{aligned} & (K \otimes Q)[\{\pi_1 \in A\} \cap \{\pi_2 \in B\} \cap \{\pi_3 \in C\} | \pi_2, \pi_3] \\ = \; & \mathbb{1}_B(\pi_2) \mathbb{1}_C(\pi_3) Q[\{\pi_1 \in C\} | \pi_2, \pi_3] \\ = \; & \mathbb{1}_B(\pi_2) \mathbb{1}_C(\pi_3) P[\{\pi_1 \in C\} | \pi_2] \end{aligned}$$

---

[8] Note that $f(x) = x \ln x$ is strictly convex, such that the lemma implies $I((\pi_1, \pi_2), \pi_3) \ge I(\pi_2, \pi_3)$ with equality if and only if π_1 is independent of π_3 given π_2

such that K is sufficient for P and Q if and only iff

$$\pi_1 \perp\!\!\!\perp \pi_3 \left| \pi_2 \right. \tag{6.113}$$

This proves the second statement. ■

Next we present a useful approximation theorem for f-divergences, as can be found in Liese and Vajda [118] and Darbellay and Vajda [59]

Lemma 6.0.13 - (**Approximation of f-divergences for probability measures**)

► **Lemma 6.0.13.1:** *Let $(I, \leq)$ be a directed set (i.e. a partially ordered set in which every pair of elements possesses an upper bound). Let $(\Omega, \mathcal{F})$ be a measurable set and let $(\mathcal{F}_i)_{i \in I}$ be a monotonous family of σ-algebras, i.e. $\mathcal{F}_i \subset \mathcal{F}_j$ for $i \leq j$. Assume that $\mathcal{F} = \sigma\left(\cup_{i \in I} \mathcal{F}_i\right)$. Then*

$$D_f\left(P\left|_{\mathcal{F}_i}\right\|Q\left|_{\mathcal{F}_i}\right.\right) \leq D_f\left(P\left|_{\mathcal{F}_j}\right\|Q\left|_{\mathcal{F}_j}\right.\right) \text{ for } i \leq j \tag{6.114}$$

and

$$\sup\left\{D_f\left(P\left|_{\mathcal{F}_i}\right\|Q\left|_{\mathcal{F}_i}\right.\right)\left|i \in I\right.\right\} = D_f\left(P\left\|Q\right.\right) \tag{6.115}$$

► **Lemma 6.0.13.2:** *Let $(\Omega, \mathcal{F})$ be a measurable space, let $\Omega_{i,j} \in \mathcal{F}$ with $i \in \mathbb{N}$ and $j \in I_i$ where $|I_i|$ is a finite set be such that:*

- *For every $i \in \mathbb{N}$, the sets $\Omega_{i,j}$ form a partition of Ω, i.e. $\Omega_{i,j_1} \cap \Omega_{i,j_2} = \emptyset$ whenever $j_1 \neq j_2$ and*
$$\Omega = \cup_{j \in I_j} \Omega_{i,j}$$
- *The sequence is nested, in the following sense: For every $j \in I_{i+1}$ there exist $J \subseteq I_i$ such that*
$$\Omega_{i+1,j} = \cup_{k \in J} \Omega_{I,k}$$
- $\mathcal{F} = \sigma\left(\left(\Omega_{i,j}\right)_{i \in \mathbb{N}, j \in I_i}\right)$

Then

$$D_f\left(P\left\|Q\right.\right) = \lim_{i \to \infty} \sum_{j \in I_j} \tilde{f}\left(P\left[\Omega_{i,j}\right], Q\left[\Omega_{i,j}\right]\right) \tag{6.116}$$

monotonously with $\tilde{f}$ given by Eq. 6.97.

Last but not least we present a formula for the Fisher gradient, a metric on appropriate manifolds of probability measures. For a good reference on information geometry, see Amari [2], Amari, Nagaoka, and Harada [5], Pistone and Sempi [147], Murray and Rice [134] and Cencov [48]. In this work we only consider the Fisher metric on strictly positive probability measures on a finite set, $\mathbf{S}$. As already claimed this space can be identified with $\Delta_{\mathbf{S}}{}^{\circ}$ again and the tangent space is given by Example 2.1.2. Then the Fisher metric is:

$$g_{F,p}\left(c_1, c_2\right) = \sum_{s \in \mathbf{S}} \frac{c_{1,s} c_{2,s}}{p(s)} \text{ where } c_1, c_2 \in T_p \Delta_{\mathbf{S}}{}^{\circ} \tag{6.117}$$

Glossary

admissible initial measure

see Definition 1.2.2

causal model

see Definition 1.2.1

causal statistical model

see Definition 1.2.2

consistent Q-estimator

see Definition 4.2.2

discrete random sets

see Definition 1.4.1

discrete random sets relative to another one

see Definition 1.4.2

ergodic derivative

see Eq 3.28 and the preceding discussion

ergodic functional

see Problem 3.2.3

essentially Lipschitz continuous quasi-projector

see Definition 2.3.1

inference σ-algebra

see Definition 1.4.3

Learning algorithm over a Markov decision process

see Definition 3.1.2

Markov decision process

see Definition 3.1.1

policy functional

see Problem 3.2.1

Acronyms

a.e.	almost everywhere
a.s.	almost surely
i.e.	that is (latin ”id est”)
iff	if and only if
MDP	Markov decision process
ODE	ordinary differential equation
POMDP	partially observable Markov decision process
w.r.t.	with respect to

List of Symbols

$\mathbb{1}_A$	The characteristic function of the set A, i.e. $$\mathbb{1}_A(\omega) := \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{else} \end{cases}$$
$\lvert\mathbf{S}\rvert$,$\lvert r\rvert$	context dependent - cardinality of $\mathbf{S}$ whenever $\mathbf{S}$ is a set; absolute value of r if $r \in \mathbb{R}$
δ_x	Dirac measure at $x \in \Omega$ where $(\Omega, \mathcal{F})$ is some measurable space: $$\delta_x(A) := \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{else} \end{cases}$$ for every $A \in \mathcal{F}$
$\mathrm{Id}_{\mathbf{X}}$	Identity function: $$\mathrm{Id}_{\mathbf{X}} : \mathbf{X} \to \mathbf{X}\,;\, x \mapsto x$$
$\prod_{i \in I} \mathbf{A}_i$	Cartesian product of the sets $\mathbf{A}_i$ over the index set I (see Eq. 1.2)
$\mathbf{A}^I$	Equivalent to $$\prod_{i \in I} \mathbf{A}$$ compare entry for $\prod_{i \in I} \mathbf{A}_i$, equivalent to the set of functions $$f : I \to \mathbf{A}$$
$(a_n)_{n \in I}$	Indexed family of elements - is the element $f \in \prod_{i \in I} \mathbf{A}_i$ with $f(i) = a_i$ (compare glossary entry for $\prod_{i \in I} \mathbf{A}_i$). If $I = \mathbb{N}$ then $(a_n)_{n \in \mathbb{N}}$ is an ordinary sequence.
$\mathrm{Par}(v)$	Parents of a vertex v in a given directed graph, (V, E). (see Definition 1.1.1)
$\mathrm{Child}(v)$	Children of a vertex v in a given directed graph, (V, E). (see Definition 1.1.1)
$\mathrm{An}(A)$	Ancestral closure of the set $A \subseteq V$ where (V, E) is some directed graph (see Definition 1.1.1)
$u \rightsquigarrow v$	There exists a path from $u \in V$ to $v \in V$ where (V, E) is some directed graph (see Definition 1.1.1)

V_n	V_0 is the set of input vertices of some directed graph (V, E), V_n is the set of vertices with an ”ancestral tree” of depth n (see Definition 1.1.1, Definition 1.1.1, Theorem 1.1.1 and the successive comment)
$V_{<n}, V_{\leq n}$	For definition of V_i see V_n. Then $V_{<n} := \cup_{0\leq k<n} V_k$ and $V_{\leq n} := \cup_{0\leq k\leq n} V_k$
$\mathbb{R}, \mathbb{R}_{\leq 0}, \mathbb{R}_{>0}$	real numbers, positive real numbers, strictly positive real numbers
$\overline{\mathbb{R}}$	extended real numbers $\overline{\mathbb{R}} := \mathbb{R} \cup \{\infty, -\infty\}$
$\mathbb{R}^n$, $\mathbb{R}^{n*}$	$\mathbb{R}^n$ is the n-fold Euclidean space, $\mathbb{R}^{n*}$ is its dual space
$\mathbb{R}^{\mathbf{S}}$, $\mathbb{R}^{\mathbf{S}*}$	$\mathbb{R}^{\mathbf{S}}$ is the set of maps $f : \mathbf{S} \to \mathbb{R}$; isomorphic to $\mathbb{R}^{\lvert\mathbf{S}\rvert}$, which can be considered as $\mathbb{R}^{\{1,2,\ldots,n\}}$. $\mathbb{R}^{\mathbf{S}*}$ is the dual space of $\mathbb{R}^{\mathbf{S}}$
$\mathbb{N}$, $\mathbb{N}_0$	$\mathbb{N}$ is the set of natural numbers (starting with one in this thesis); $\mathbb{N}_0 = \mathbb{N} \cup \{0\}$
2^V	Power set of V: $2^V = \{W \subseteq V\}$
$\otimes_{v\in V} \mathcal{F}_v$	Product σ-algebra (see Eq. 1.4)
$M_1(\mathcal{F})$	Probability measures on the σ-algebra $\mathcal{F}$, i.e. maps $P : \mathcal{F} \to \mathbb{R}_{\geq 0}$ that are σ-additive and normalized.
f_*P	Pushforward of the probability measure $P \in M_1(\mathcal{F}_{\mathbf{X}})$ under the $\mathcal{F}_{\mathbf{X}}$/$\mathcal{F}_{\mathbf{Y}}$ measurable map $f : \mathbf{X} \to \mathbf{Y}$: $f_*P[A] := P\left[f^{-1}(A)\right]$ The measure f_*P is also known as distribution of the random variable f on the probability space $(\mathbf{X}, \mathcal{F}_X, P)$.
$\mathcal{F}$,$\mathcal{G}$	Usually used to denote σ-algebras
$\mathbb{F}$	Usually used to denote filtrations of σ-algebras, i.e. $\mathbb{F}_n$ is a σ-algebra for every $n \in \mathbb{N}$ (or $n \in \mathbb{N}_0$) such that $\mathbb{F}_n \subseteq \mathbb{F}_{n+1}$ for every n.
$\Lambda^{(\mathbf{Y},\mathcal{F}_Y)}_{(\mathbf{X},\mathcal{F}_X)}$, $\Lambda^{\mathbf{Y}}_{\mathbf{X}}$	Probability kernel from $(\mathbf{X}, \mathcal{F}_X)$ to $(\mathbf{Y}, \mathcal{F}_Y)$, i.e. maps $K : \mathbf{X} \times \mathcal{F}_Y \to [0,1]$ such that $K(\cdot, A)$ is measurable for every $A \in \mathcal{F}_Y$ and $K(x, \cdot)$ is a probability measure for every $x \in \mathbf{X}$
$\cup$, $\cap$	Set union / set intersection
$A \setminus B$	Set theoretical difference of A and B
ν_{Leb}	Lebesgue measure
$\perp\!\!\!\perp$	(conditional) independence, see Definition 1.3.2
G^m	Moral graph of G, see Definition 1.3.1

$\mathrm{CON}_{B\cup A\cup S,S}(A)$	Causal connected component of A, see Definition 1.3.3
$\leftrightsquigarrow_{m,S}$	See Definition 1.3.3
$\mathrm{ResAnc}(A\,\vert S\vert\,B)$	Residual ancestors, see Definition 1.3.3
$\mathcal{G}_{2^V}{}^{(N)}$	Discrete σ-algebra on the power set of V, compare Eq. 1.68
$\mathrm{Range}(f)$	Range of the function f
I_τ	Present random set associated to the discrete random set, τ, compare Definition 1.4.4
$\mathcal{F}_{\text{pr},\tau}$	Present σ-algebra associated to the discrete random set, τ, compare Definition 1.4.4
$T_K(x)$	Tangent cone of the set K at x, see. Definition 2.1.1
$N_K(x)$	Normal cone of the set K at x, see. Definition 2.1.1
$d_K(x)$	distance of x from the set K
$B_R(x)$	open ball with center x and radius R
$\overline{A}$	Closure of the set A
$\Delta_{\epsilon;\mathbf{S}}$	ϵ-simplex over the (finite) set $\mathbf{S}$, see. Example 2.1.2
e_s, $e_{s,*}$	$(e_s)_{s\in\mathbf{S}}$ is the canonical basis in $\mathbb{R}^{\mathbf{S}}$: $$e_s(s') := \delta_{s,s'}$$ and $e_{s,*}$ is the corresponding dual basis: $$e_{s,*} : f \mapsto f(s)$$
$\delta_{s,s'}$	Kronecker delta: $$\delta_{s,s'} = \begin{cases} 1 & \text{if } s = s' \\ 0 & \text{else} \end{cases}$$
$\lVert\cdot\rVert_p$	p-norm on $\mathbb{R}^{\mathbf{S}}$: $$\lVert v\rVert_p := \left(\sum_{s\in\mathbf{S}} \lvert v_s\rvert^p\right)^{\frac{1}{p}}$$ for $p = \infty$: $\lVert v\rVert_\infty := \max\left\{\lvert v_s\rvert \,\vert s \in \mathbf{S}\right\}$
$\lVert\cdot\rVert_{\text{Op},p}$	Operator norm for for $\mathbf{S}\times\mathbf{S}$ matrices with respect to p-norm on $\mathbb{R}^{\mathbf{S}}$: $$\lVert A\rVert_{\text{Op},p} := \sup\left\{\lVert Av\rVert_p \,\middle\vert\, v \in \mathbb{R}^{\mathbf{S}}; \lVert v\rVert_p = 1\right\}$$
$\underline{1}$	vector in $\mathbb{R}^{\mathbf{S}}$ with $\lvert\mathbf{S}\rvert < \infty$ satisfying $$\underline{1}_s := 1 \text{ for every } s \in \mathbf{S}$$

$\lVert\cdot\rVert_*$	Dual norm of $\lVert\cdot\rVert$: : $\lVert\lambda\rVert_* := \sup\left\{\left\vert\lambda(v)\right\vert_p \,\middle\vert\, v \in \mathbb{R}^{\mathbf{S}}; \lVert v\rVert = 1\right\}$ for $\lambda \in \mathbb{R}^{\mathbf{S}*}$
$\mathrm{convcone}(A)$	convex cone generated by the set $A \subseteq V$ where V is a vector space
$D\phi(x_*)\,[v]$	directional derivative of ϕ at x_* into direction v
$\mathrm{Cl}\,(\mathbb{R}^n)$	The set of closed subsets of $\mathbb{R}^n$
$\mathrm{Graph}\,(F)$	Graph of the set-valued map $F : \mathbb{R}^n \to \mathbb{R}^m$: $\mathrm{Graph}\,(F) = \{(x, y) \in \mathbb{R}^n \times \mathbb{R}^m \mid y \in F(x)\} \quad (6.118)$
$\mathcal{L}_1(d\mu)$	Set of (extended) real-valued measurable functions from some measurable space $(\Omega, \mathcal{F})$ with bounded integral with respect to μ, i.e. $\mathcal{L}_1(d\mu) = \left\{\phi : \Omega \to \mathbb{R} \,\middle\vert\, \int_\Omega \lvert\phi(\omega)\rvert\, \mu(d\omega) < \infty\right\}$
$L_1\,(d\mu)$	L_1-space with respect to measure $\mu \in M_1\,(\mathcal{F})$ for some σ-algebra $\mathcal{F}$, i.e. equivalent classes, $[\phi]$, of $\mathcal{F}/\mathcal{B}_{\mathbb{R}}$ measurable functions where $\int \lvert\phi\rvert\, d\mu < \infty$ and $\phi_2 \in [\phi]$ if and only if $\int \lvert\phi - \phi_1\rvert\, d\mu = 0$
$\mathfrak{B}\,(\mathbf{S}, \mathcal{F}_S)$	Real-valued, bounded, $\mathcal{F}_S/\mathcal{B}_{\mathbb{R}}$ measurable functions
$\nabla_g\phi(x)$	gradient with respect to the metric g of the differentiable function $\phi : M \to \mathbb{R}$ at point $x \in M$, i.e. the unique vector satisfying $g\,(\nabla_g\phi(x), v) = D\phi(x)\,[v]$ for every $v \in T_xM$
$C_{\hat{P}}(x)$	Preimage cone of the quasi-projector, $\hat{P}$ at x (compare Definition 2.3.1)
$\mathrm{INV}_K\,(\mathcal{F}_S)$	K-invariant measures, compare Definition 6.0.2
INV	Set-valued map that maps a kernel to its invariant distributions: $K \mapsto \{\mu \in M_1\,(\mathcal{F}) \mid \mu K = \mu\}$
A°	Set of (topologically) interior points of A
$\mathrm{bil}\,(\mathbb{R}^n)$	Bilinearforms on $\mathbb{R}^n$, i.e. the set of maps $b : \mathbb{R}^n \times \mathbb{R}^n \to \mathbb{R}$ satisfying $b(x+\alpha y, u+\beta v) = b(x, u)+\alpha b(y, u)+\beta b(x, v)+\alpha\beta b(y, v)$ for every $x, y, u, v \in \mathbb{R}^n$ and $\alpha, \beta \in \mathbb{R}$.
$\mathcal{D}_{\text{erg.}}f$	Ergodic derivative of f, see Eq 3.28 in the Appendix

$\nabla_E \phi$	Euclidean gradient of ϕ, see Appendix A.2.2
$\nabla_F \phi$	Fisher gradient of ϕ, see Appendix A.2.2
K_*, Y_K	see Appendix A.1.4, Theorem 6.0.3
$D_f\left(\mu \,\|\nu\right)$	f-divergence of μ and ν see Appendix A.3, Definition 6.0.5 and Eq. Definition 6.0.6
$D\left(\mu \,\|\nu\right)$	see Definition 6.0.5
$H(p)$	Entropy, see Appendix A.3, Definition 6.0.5
$D_{\text{KL}}\left(\mu \,\|\nu\right)$	KL-divergence, see Appendix A.3, Definition 6.0.5 and Definition 6.0.6
H_η	Generalized differential entropy, see Appendix A.3, Definition 6.0.6
$\otimes_{i \in I} p_i$	product measure of the measures $p_i \in M_1\left(\mathcal{F}_i\right)$ where $\mathcal{F}_i$ are σ-algebras on some sets Ω_i, i.e. $\otimes_{i\in I} p_i \in M_1\left(\otimes_{i \in I} \mathcal{F}_i\right)$ is the unique measure with the property that for any finite subset $J \subseteq I$ and any collection of sets $A_j \in \mathcal{F}_j$ with $j \in J$: $\left(\otimes_{i\in I} p_i\right)\left[\cap_{j \in J}\left\{\pi_j \in A_j\right\}\right] = \prod_{j \in J} p_j\left[A_j\right]$
$P_{\text{MDP},q',s',\mathbf{z}'}$	Law on the causal model associated to an MDP with parameter $q' \in \mathbf{Q}$, initial sensor state $s' \in \mathbf{S}$ and policy sequence $\mathbf{z}' \in \mathbf{Z}^{\mathbb{N}_0}$, compare Eq. 3.1
$P_{q',s',m'}$	Law of an a MDP with controlled policy parameters with parameter $q' \in \mathbf{Q}$, initial sensor state $s' \in \mathbf{S}$ and initial memory value $m' \in \mathbf{M}$, compare Eq. 3.5
$\mathcal{B}_b(\mathbf{S})$	real-valued, bounded measurable functions on the measurable space $(\mathbf{S}, \mathcal{F}_S)$
$f_{\text{rew};r,\lambda}$	policy functional corresponding to expected, discounted (one-point) reward, r, see Eq. 3.48 and Problem 3.2.2
$f_{\text{rew.var.};r,\lambda}$	policy functional corresponding to variance of expected, discounted reward, r, see Eq. 3.64 and Problem 3.2.2
$f_{\text{rew},1;r,\lambda}$	policy functional corresponding to the sliding expected, discounted (multiple point) reward of r, see Eq. 3.74 and Problem 3.2.2
$f_{\text{rew},2;r,\lambda}$	policy functional corresponding to the block-wise expected, discounted (multiple point) reward of r, see Eq. 3.75 and Problem 3.2.2
$f_{\text{entr};\lambda}$	policy functional corresponding to discounted entropy, see Eq. 3.80 and Problem 3.2.2
$f_{\text{diff.entr};\eta,\lambda}$	policy functional corresponding to discounted generalized entropy, see Eq. 3.82 and Problem 3.2.2
$f_{\text{M.I.},1;\lambda}$	policy functional corresponding to sliding discounted mutual information, see Eq. 3.85 and Problem 3.2.2
$f_{\text{M.I.},2;\lambda}$	policy functional corresponding to block-wise discounted mutual information, see Eq. 3.86 and Problem 3.2.2
$f_{\text{erg.rew };r}$	ergodic functional corresponding to expected reward, see Eq. 3.53 and Problem 3.2.3

$f_{\text{erg.entr}}$	ergodic functional corresponding to entropy, see Eq. 3.83 and Problem 3.2.3
$f_{\text{P.I.}}$	ergodic functional corresponding to mutual information/predictive information, see Eq. 3.93 and Problem 3.2.3
$P_{q',s',Z}$	Law on the causal model of an MDP where the policy sequence is an adapted process, for further details see Remark 4.2.1

Bibliography

[1] C.D. Aliprantis and K.C. Border. *Infinite Dimensional Analysis: A Hitchhiker's Guide*. Springer, 2007.

[2] S.I. Amari. *Differential-Geometrical Methods in Statistics*. Springer-Verlag, 1985.

[3] S.I. Amari. Natural gradient works efficiently in learning. *Neural Comput.* **10**, 2 (1998), pp. 251–276.

[4] S.I. Amari and S.C. Douglas. Why natural gradient? *Proceedings of the 1998 IEEE International Conference on Acoustics, Speech and Signal Processing*. 1998, pp. 1213–1216.

[5] S.I. Amari, H. Nagaoka, and D. Harada. *Methods of Information Geometry*. American Mathematical Society, 2007.

[6] G. Anger, J.P. Aubin, and A. Cellina. Differential inclusions, set-valued maps and viability theory. *Journal of Applied Mathematics and Mechanics* **67**, 2 (1987), p. 100.

[7] J.P. Aubin. *Viability Theory*. Birkhäuser, 1991.

[8] J.P. Aubin and H. Frankowska. *Set-Valued Analysis*. Birkhäuser, 1990.

[9] N. Ay. A refinement of the common cause principle. *Discrete Appl. Math.* **157**, 10 (2009), pp. 2439–2457.

[10] N. Ay. An Information-Geometric Approach to a Theory of Pragmatic Structuring. *The Annals of Probability*. Preprint **30**, 1 (2000), pp. 416–436.

[11] N. Ay and I. Erb. On a notion of linear replicator equations. *Journal of Dynamics and Differential Equations* **17** (2005), pp. 427–451.

[12] N. Ay, G.F. Montúfar, and J. Rauh. Selection criteria for neuromanifolds of stochastic dynamics. *Advances in Cognitive Neurodynamics (III): Proceedings of the Third International Conference on Cognitive Neurodynamics - 2011*. Ed. by Yoko Yamaguchi. Springer-Verlag GmbH, 2013.

[13] N. Ay and D. Polani. Information flows in causal networks. *Advances in Complex Systems* **11**, 1 (2008), pp. 17–41.

[14] N. Ay and K. Zahedi. Causal effects for prediction and deliberative decision making of embodied systems. *Advances in Cognitive Neurodynamics (III)*. Ed. by Yoko Yamaguchi. Springer Netherlands, 2013, pp. 499–506.

[15] N. Ay et al. A geometric approach to complexity. *Chaos* **21**, 3 (2011), p. 037103.

[16] N. Ay et al. Information-driven self-organization: the dynamical system approach to autonomous robot behavior. *Theory in Biosciences* **131**, 3 (2012), pp. 161–179.

[17] N. Ay et al. Predictive information and explorative behavior of autonomous robots. *European Physical Journal B* **63** (2008), pp. 329–339.

[18] J.C. Baez, T. Fritz, and T. Leinster. A characterization of entropy in terms of information loss. *ArXiv eprint* (2011).

[19] V. Bagdonavicius, O. Cheminade, and M. Nikulin. Statistical planning and inference in accelerated life testing using the CHSS model. *Journal of Statistical Planning and Inference* **126**, 2 (2004), pp. 535–551.

[20] J. Bang-Jensen and G.Z. Gutin. *Digraphs: Theory, Algorithms and Applications*. Springer Publishing Company, Incorporated, 2008.

[21] O.E. Barndorff-Nielsen and A. Shiryaev. *Change of Time and Change of Measure*. World Scientific, 2010.

[22] H. Bauer. *Probability Theory*. Bod Third Party Titles, 1996.

[23] D.P. Bertsekas and E.S. Shreve. *Stochastic Optimal Control: The Discrete-Time Case*. Athena Scientific, 2007.

[24] D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.

[25] B Bharath and V Borkar. Stochastic approximation algorithms: overview and recent trends. *Sadhana* **24** (1999), pp. 425–452.

[26] R. Bhatia. *Matrix Analysis*. Springer Verlag, 1997.

[27] W. Bialek, I. Nemenman, and N. Tishby. Predictability, complexity, and learning. *Neural Computation* **13**, 11 (2001), pp. 2409–2463.

[28] W. Bialek and N. Tishby. Predictive information. *ArXiv eprint:cond-mat/9902341* (1999).

[29] E. Bierstone and P.D. Milman. Semianalytic and subanalytic sets. *Publications Mathématiques de l'IHÉS* **67** (1988), pp. 5–42.

[30] P. Billingsley. *Convergence of Probability Measures*. Wiley, 2009.

[31] C.M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.

[32] J. Bolte, A. Daniilidis, and A. Lewis. The Łojasiewicz inequality for nonsmooth subanalytic functions with applications to subgradient dynamical systems. *SIAM J. on Optimization* **17**, 4 (2006), pp. 1205–1223.

[33] J. Bolte et al. Clarke critical values of subanalytic Lipschitz continuous functions. *Ann. Polon. Math.* **87** (2005), pp. 13–25.

[34] J. Bolte et al. Clarke subgradients of stratifiable functions. *SIAM J. on Optimization* **18**, 2 (2007), pp. 556–572.

[35] A. Bondy and U.S.R. Murty. *Graph Theory*. Springer, 2008.

[36] K.C. Border. *Notes on the Kolmogorov Extension Problem*. California Institute of Technology. 1998.

[37] V.S. Borkar. Reinforcement learning in Markovian evolutionary games. *Advances in Complex Systems* **5**, 1 (2002), pp. 55–72.

[38] V.S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Cambridge University Press, 2008.

[39] L. Breiman. *Probability*. Society for Industrial and Applied Mathematics, 1968.

[40] H. Brezis. *Functional Analysis, Sobolev Spaces and Partial Differential Equations*. Springer, 2010.

[41] A.M. Bruckner and M. Rosenfeld. On topologizing measure spaces via differentiation bases. *Annali della Scuola Normale Superiore di Pisa - Classe di Scienze* **23** (1969), pp. 243–258.

[42] J. Bucklew. The source coding theorem via Sanov's theorem (Corresp.) *Information Theory, IEEE Transactions on* **33**, 6 (1987), pp. 907–909.

[43] L. Buşoniu et al. *Reinforcement Learning and Dynamic Programming Using Function Approximators*. Boca Raton, Florida: CRC Press, 2010.

[44] L.L. Campbell. An extended cencov characterization of the information metric. *Proceedings of the American Mathematical Society* **98**, 1 (1986), pp. 135–141.

[45] X.R. Cao and Chen H.F. Perturbation realization, potentials, and sensitivity analysis of Markov processes. *IEEE Transactions on Automatic Control* (1997).

[46] L. Carter et al. Unbiased estimation of the MSE matrix of Stein-Rule estimators, confidence ellipsoids, and hypothesis testing. *Econometric Theory* **6**, 1 (1990), pp. 63–74.

[47] A. Caserta. Decomposition of topologies which characterize the upper and lower semicontinuous limits of functions. *Abstract and Applied Analysis* (2011).

[48] N.N. Cencov. *Statistical Decision Rules and Optimal Inference*. American Mathematical Society, 1980.

[49] X. Chen. *Limit Theorems for Functionals of Ergodic Markov Chains with General State Space*. American Mathematical Society, 1999.

[50] F.H. Clarke. *Functional Analysis, Calculus of Variations and Optimal Control*. Springer, 2013.

[51] F.H. Clarke, R.J. Stern, and G. Sabidussi. *Nonlinear Analysis, Differential Equations and Control*. Springer, 1999.

[52] G. Cohen, R.L. Jones, and M. Lin. *On strong laws of large numbers with rates*. URL: `http://www.ee.bgu.ac.il/~guycohen/cjl.pdf`.

[53] S.H. Collins, M. Wisse, and A. Ruina. A three-dimensional passive-dynamic walking robot with two legs and knees. *I. J. Robotic Res.* **20**, 7 (2001), pp. 607–615.

[54] T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley-Interscience, 1991.

[55] H. Crauel and M. Gundlach. *Stochastic Dynamics*. Springer, 1999.

[56] J.P. Crutchfield and D.P. Feldman. Regularities unseen, randomness observed: levels of entropy convergence. *Chaos* **13**, 1 (2003), pp. 25–54.

[57] I. Csiszár. Axiomatic characterizations of information measures. *Entropy* **10**, 3 (2008), pp. 261–273.

[58] I. Csiszár and J. Korner. *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Orlando, FL, USA: Academic Press, Inc., 1982.

[59] G.A. Darbellay and I. Vajda. Estimation of the information by an adaptive partitioning of the observation space. *IEEE Transactions on Information Theory* **45**, 4 (1999), pp. 1315–1321.

[60] M.P. Deisenroth, G. Neumann, and J. Peters. A survey on policy search for robotics. *Foundations and Trends in Robotics* **21** (2013), pp. 388–403.

[61] T. DelSole and M.K. Tippett. Predictability: recent insights from information theory. *Reviews of Geophysics* **45** (2007).

[62] A. Dembo and O. Zeitouni. *Large Deviations Techniques and Applications*. Springer, 1998.

[63] R. Der and G. Martius. *The Playful Machine - Theoretical Foundation and Practical Realization of Self-Organizing Robots*. Springer, 2012.

[64] O.R. Diaz-Espinosa. URL: `http://www.math.duke.edu/~odiaz/mcergo.pdf`.

[65] R. Diestel. *Graph Theory (Graduate Texts in Mathematics)*. Springer, 2005.

[66] S. Dragiev, M. Toussaint, and M. Gienger. Uncertainty aware grasping and tactile exploration. *ICRA*. 2013, pp. 113–119.

[67] Drazin. Pseudo-inverses in associative rings and semigroups. *The Amercan Mathematical Monthly* **65** (1958), pp. 506–514.

[68] E.B. Dynkin and A.A. Yushkevich. *Controlled Markov processes*. Springer, 1979.

[69] B. Efron and C. Morris. Data analysis using Stein's estimator and its generalizations. *Journal of the American Statistical Association* **70** (1975), pp. 311–319.

[70] R.S. Ellis. *The theory of large deviations and applications to statistical mechanics*. Department of Mathematics and Statistics University of Massachusetts. 2009. URL: `http://www.math.umass.edu/~rsellis/pdf-files/Les-Houches-lectures.pdf`.

[71] A. Eugene and A.S. Feinberg. *Handbook of Markov Decision Processes - Methods and Applications*. Ed. by E.A. Feinberg and A. Shwartz. Kluwer International Series, 2002.

[72] A.F. Filippov. *Differential Equations with Discontinuous Righthand Sides: Control Systems (Mathematics and its Applications)*. Springer, 1988.

[73] M.I. Freĭdlin and A.D. Wentzell. *Random Perturbations of Dynamical Systems*. Springer, 1998.

[74] P. Georgiev, A. Cichocki, and S.I. Amari. *On some extensions of the natural gradient algorithm*. 2001.

[75] P.W. Glynn. *Harris recurrent Markov chains*.

[76] C.W.J. Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37**, 3 (1969), pp. 424–438.

[77] P. Grassberger. Randomness, information, and complexity. *ArXiv eprint* (2012).

[78] P. Grassberger. Toward a quantitative theory of self-generated complexity. *International Journal of Theoretical Physics* **25**, 9 (1986), pp. 907–938.

[79] V. Gutev and T. Nogura. A topology generated by selections. *Topology and Its Applications* **153** (2005), pp. 900–911.

[80] J.M. Hammersley and P.E. Clifford. Markov random fields on finite graphs and lattices. Unpublished manuscript (1971).

[81] S.T. Han. *Information-Spectrum Methods in Information Theory*. Springer, 2003.

[82] P. Harremoëos and K. Holst. Convergence of Markov chains in information divergence. *Journal of Theoretical Probability* **22** (2009), pp. 186–202.

[83] R. Hassin and M. Haviv. Mean passage times and nearly uncoupled Markov chains. *SIAM J. Discret. Math.* **5**, 3 (Aug. 1992), pp. 386–397.

[84] R. Hassin and M. Haviv. Mean passage times and nearly uncoupled Markov chains. *SIAM J. Discrete Math.* **5**, 3 (1992), pp. 386–397.

[85] T. Hastie, R. Tibshirani, and J.H. Friedman. *The Elements of Statistical Learning*. Springer, 30, 2003.

[86] E. Hille and R.S. Phillips. *Functional Analysis and Semi-Groups*. American Mathematical Society, 1957.

[87] M. Hoffmann and R. Pfeifer. The implications of embodiment for behavior and cognition: animal and robotic case studies. *CoRR* (2012).

[88] F. Hollander. *Large Deviations*. American Mathematical Society, 2000.

[89] R.A. Horn and C.R. Johnson. *Matrix Analysis*. Cambridge University Press, 1990.

[90] Y.M. Huang and M. Valtorta. Pearl's calculus of intervention is complete. *Proceedings of the Twenty-Second Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*. Arlington, Virginia: AUAI Press, 2006, pp. 217–224.

[91] T. Jaakkola, S.P. Singh, and M.I. Jordan. Reinforcement learning algorithm for partially observable Markov decision problems. *Advances in Neural Information Processing Systems 7*. MIT Press, 1995, pp. 345–352.

[92] J. Jacod and P.E. Protter. *Probability Essentials*. Springer, 2003.

[93] J. Jahn. *Introduction to the Theory of Nonlinear Optimization*. Springer, 2007.

[94] D. Janzing et al. Information-geometric approach to inferring causal directions. *Artif. Intell.* **182-183** (2012), pp. 1–31.

[95] J.L. Jensen. A note on asymptotic expansions for Markov chains using operator theory. *Adv. Appl. Math.* **8**, 4 (1987), pp. 377–392.

[96] O. Johnson and O.T. Johnson. *Information Theory and the Central Limit Theorem*. Imperial College Press, 2004.

[97] J. Jost. *Dynamical Systems: Examples of Complex Behaviour*. Springer, 2005.

[98] L.P. Kaelbling, M.L. Littman, and A.W. Moore. *Reinforcement learning: a survey*. 1996. URL: `http://www.cs.cmu.edu/afs/cs/project/jair/pub/volume4/kaelbling96a.pdf`.

[99] O. Kallenberg. *Foundations of Modern Probability*. Springer, 2002.

[100] A.S. Kechris. *Classical Descriptive Set Theory*. Springer, 1995.

[101] J.G. Kemeny and J.L. Snell. *Finite Markov Chains*. Springer, 1976.

[102] J. Kober, E. Oztop, and Peters J. Reinforcement learning to adjust robot movements to new situations. *IJCAI*. 2011, pp. 2650–2655.

[103] J. Kober and J. Peters. Policy search for motor primitives in robotics. *Machine Learning* **84** (2011), pp. 171–203.

[104] W. König. *Grosse Abweichungen, Techniken und Anwendungen*. German. 2006. URL: `http://www.wias-berlin.de/people/koenig/www/GA.pdf`.

[105] W. König. *Wahrscheinlichkeitstheorie I+II*. German. 2006. URL: `http://www.wias-berlin.de/people/koenig/www/Skripte.html`.

[106] H.J. Kushner. *Approximation and Weak Convergence Methods for Random Processes with Applications to Stochastic Systems Theory*. MIT Press, 1984.

[107] H.J. Kushner. *Heavy Traffic Analysis of Controlled Queueing and Communication Networks*. Applications of Mathematics. Springer, 2001.

[108] H.J. Kushner. Stochastic approximation: a survey. *Computational Statistics* **2**, 1 (2010), pp. 87–96.

[109] H.J. Kushner and D.S. Clark. *Stochastic Approximation Methods for Constrained and Unconstrained Systems*. Springer-Verlag, 1978.

[110] H.J. Kushner and G. Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Springer, 2003.

[111] Z.A. Lagodowski and Z. Rychlik. Rate of convergence in the strong law of large numbers for martingales. *Probability Theory and Related Fields* **71**, 3 (1986), pp. 467–476.

[112] S. Lang. *Fundamentals of Differential Geometry*. Springer Verlag, 1999.

[113] T. Lang, M. Toussaint, and K. Kersting. Exploration in relational domains for model-based reinforcement learning. *Journal of Machine Learning Research* **13** (2012), pp. 3691–3734.

[114] S.L. Lauritzen. *Graphical Models*. Oxford University Press, 1996.

[115] D. Leao Jr., M. Fragoso, and P. Ruffino. Regular conditional probability, disintegration of probability and radon spaces. en. *Proyecciones* **23** (2004), pp. 15–29.

[116] A. Ijspeert, J. Nakanishi, and S. Schaal. Learning attractor landscapes for learning motor primitives. *Advances in Neural Information Processing Systems 15*. Cambridge, MA: MIT Press, 2003, pp. 1547–1554.

[117] G. Lebanon. Axiomatic geometry of conditional models. *IEEE Transactions on Information Theory* **51**, 4 (2005), pp. 1283–1294.

[118] F. Liese and I. Vajda. *Convex statistical distances*. Teubner, 1987.

[119] R. Liptser and A.N. Shiryayev. *Theory of Martingales*. Springer, 1989.

[120] D.Y. Little and F.T. Sommer. Learning in embodied action-perception loops through exploration. *ArXiv eprint* (2011).

[121] W. Löhr and N. Ay. Non-sufficient memories that are sufficient for prediction. *Complex Sciences*. Ed. by Jie Zhou. Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Springer, 2009, pp. 265–276.

[122] W.S. Lovejoy. A survey of algorithmic methods for partially observed Markov decision processes. *Ann. Oper. Res.* **28**, 1-4 (1991), pp. 47–66.

[123] R. Lucchetti and A. Pasquale. A new approach to a hyperspace theory. *Journal of Convex Analysis* **1** (1994), pp. 173–193.

[124] P. Marbach and J.N. Tsitsiklis. *Simulation-based optimization of Markov reward processes*. Tech. rep. IEEE Transactions on Automatic Control, 1998.

[125] B. Marx and W. Vogt. *Dynamische Systeme: Theorie Und Numerik*. Spektrum Akademischer Verlag GmbH, 2011.

[126] C.D. Meyer. The role of the generalized inverse in the theory of finite Markov chains. *SIAM Review* **17**, 3 (1975), pp. 443–464.

[127] S. Meyn, R.L. Tweedie, and P.W. Glynn. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2009.

[128] C. Michelot. A finite algorithm for finding the projection of a point onto the canonical simplex of α^n; *Journal of Optimization Theory and Applications* **50** (1986), pp. 195–200.

[129] H. Miyamoto et al. A Kendama learning robot based on bi-directional theory. *Neural Networks* **9**, 8 (1996), pp. 1281–1302.

[130] G. Montúfar. Mixture decompositions of exponential families using a decomposition of their sample spaces. *Kybernetika* **49**, 1 (2013), pp. 23–39.

[131] K. Muelling, J. Kober, and J. Peters. A biomimetic approach to robot table tennis. *Adaptive Behavior Journal* **19**, 5 (2011).

[132] K. Muelling et al. Learning to select and generalize striking movements in robot table tennis. *International Journal of Robotics Research* **32**, 3 (2013), pp. 263–279.

[133] K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press, 2012.

[134] M.K. Murray and J.W. Rice. *Differential Geometry and Statistics*. Chapman & Hall, 1993.

[135] J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. English. *Journal of the Royal Statistical Society. Series A (General)* **135**, 3 (1972), pp. 370–384.

[136] G. Neumann. Variational inference for policy search in changing situations. *Proceedings of the International Conference on Machine Learning*. 2011.

[137] J. Peters et al. A unifying framework for robot control with redundant DOFs. *Autonomous Robots* **24**, 1 (2008), pp. 1–12.

[138] J. Palis and W. de Melo. *Geometric Theory of Dynamical Systems: An Introduction*. University of Beijing, 1998.

[139] A. Paraschos et al. Probabilistic movement primitives. *Advances in Neural Information Processing Systems, Cambridge, MA: MIT Press.* 2013.

[140] J. Pearl. The do-calculus revisited. *CoRR* (2012).

[141] L. Perko. *Differential Equations and Dynamical Systems*. U.S. Government Printing Office, 2001.

[142] J. Peters and S. Schaal. Policy gradient methods for robotics. *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems*. Beijing, China, 2006.

[143] J. Peters and S. Schaal. Reinforcement learning of motor skills with policy gradients. *Neural Networks* **21** (2008), pp. 682–697.

[144] J. Peters, S. Vijayakumar, and S. Schaal. Natural actor-critic. *Proceedings of the European Machine Learning Conference*. Porto, Portugal, 2005.

[145] Jan Peters, Katharina Mülling, and Yasemin Altün. Relative Entropy Policy Search. *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI 2010)*. Ed. by Maria Fox and David Poole. AAAI Press, 2010, pp. 1607–1612.

[146] D. Petz. Monotone metrics on matrix spaces. *Linear Algebra and its Applications* **244** (1996), pp. 81–96.

[147] G. Pistone and C. Sempi. An infinite-dimensional geometric structure on the space of all the probability measures equivalent to a given one. English. *The Annals of Statistics* **23**, 5 (1995), pp. 1543–1561.

[148] S.T. Račev et al. *The Methods of Distances in the Theory of Probability and Statistics*. Springer, 2013.

[149] K. Rawlik, M. Toussaint, and S. Vijayakumar. An approximate inference approach to temporal optimization in optimal control. *NIPS*. Ed. by John D. Lafferty et al. Curran Associates, Inc., 2010, pp. 2011–2019.

[150] K. Rawlik, M. Toussaint, and S. Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. *Robotics: Science and Systems*. 2012.

[151] F. Redig and F. Völlering. Concentration of additive functionals for Markov processes and applications to interacting particle systems. *ArXiv eprint* (2010).

[152] J. Peters and S. Schaal. Reinforcement learning by reward-weighted regression for operational space control. *Proceedings of the International Conference on Machine Learning*. 2007.

[153] G.O. Roberts and J.S. Rosenthal. General state space Markov chains and MCMC algorithm. *Probability surveys* **1**, 1 (2004), pp. 20–71.

[154] L.C.G. Rogers and D. Williams. *Diffusions, Markov Processes, and Martingales: Volume 1, Foundations*. Cambridge University Press, 2000.

[155] L.C.G. Rogers and D. Williams. *Diffusions, Markov processes, and Martingales: Volume 2, Ito Calculus*. Cambridge University Press, 2000.

[156] A. Rosalsky and G. Stoica. On the strong law of large numbers for identically distributed random variables irrespective of their joint distributions. *Statistics & Probability Letters* **80**, 17-18 (2010), pp. 1265–1270.

[157] S.M. Ross. *Introduction to Probability Models*. Academic Press, Inc., 2006.

[158] E.A. Rückert et al. Learned graphical models for probabilistic planning provide a new class of movement primitives. *Frontiers in Computational Neuroscience* (2012).

[159] T. Rückstieß, M. Felder, and J. Schmidhuber. State-dependent exploration for policy gradient methods. *ECML/PKDD (2)*. Ed. by W. Daelemans, B. Goethals, and K. Morik. Springer, 2008, pp. 234–249.

[160] W. Rudin. *Functional Analysis*. McGraw-Hill Book Co., 1973.

[161] W. Rudin. *Real and Complex Analysis*. McGraw-Hill Book Co., 1987.

[162] T. Sagawa. *Thermodynamics of Information Processing in Small Systems*. Springer Japan, 2013.

[163] A.A. Samer and M.D. Plumbley. A measure of statistical complexity based on predictive information. *CoRR* (2010).

[164] A. Sard. Hausdorff measure of critical images on Banach manifolds. *Am. J. Math.* **87** (1965), pp. 158–174.

[165] A. Sard. The measure of the critical values of differentiable maps. *Bulletin of American Mathematical Society* **48**, 2 (1942), pp. 883–890.

[166] T. Schreiber. Measuring information transfer. *Physical Review Letters* **85** (10, 2000), pp. 461–464.

[167] P.J. Schweitzer. Aggregation methods for large Markov chains. *Computer Performance and Reliability*. Ed. by G. Iazeolla, P.J. Courtois, and A. Hordijk. North-Holland, 1983, pp. 275–286.

[168] P.J. Schweitzer. Perturbation theory and finite Markov chains. *Journal of Applied Probability* **5**, 2 (1968), pp. 401–413.

[169] T. Senoo et al. Skillful manipulation based on high-speed sensory-motor fusion. *ICRA*. 2009, pp. 1611–1612.

[170] C.E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal* **27** (1948), pp. 379–423.

[171] H.A. Simon and A. Ando. Aggregation of variables in dynamic systems. *Econometrica* **29** (1961), pp. 111–138.

[172] S.P. Singh, T. Jaakkola, and M.I. Jordan. Learning without state-estimation in partially observable Markovian decision processes. *In Proceedings of the Eleventh International Conference on Machine Learning*. Morgan Kaufmann, 1994, pp. 284–292.

[173] M. Spivak. *A Comprehensive Introduction to Differential Geometry*. Publish or Perish, Incorporated, 1975.

[174] M. Spivak. *Calculus On Manifolds: A Modern Approach To Classical Theorems Of Advanced Calculus*. Westview Press, 1971.

[175] G. Stoica. A note on the rate of convergence in the strong law of large numbers for martingales. *Journal of Mathematical Analysis and Applications* **381**, 2 (2011), pp. 910–913.

[176] G. Stoica. Baum-Katz-Nagaev type results for martingales. *Journal of Mathematical Analysis and Applications* **336**, 2 (2007), pp. 1489–1492.

[177] E. Stuhlsatz. *Möbius inversion formula*. URL: `http://www.whitman.edu/mathematics/SeniorProjectArchive/2008/stuhlsatz.pdf`.

[178] R.S. Sutton. Learning to predict by the methods of temporal differences. *Machine Learning*. Kluwer Academic Publishers, 1988, pp. 9–44.

[179] R.S. Sutton and A.G. Barto. *Reinforcement Learning: An Introduction (Adaptive Computation and Machine Learning)*. A Bradford Book, 1998.

[180] R.S. Sutton et al. Policy gradient methods for reinforcement learning with function approximation. *Advances in Information Processing Systems* **12** (2000), pp. 1057–1063.

[181] V.B. Tadic. Convergence and convergence rate of stochastic gradient search in the case of multiple and non-isolated extrema. *CDC*. IEEE, 2010, pp. 5321–5326.

[182] J. Nakanishi et al. Operational space control: a theoretical and emprical comparison. *International Journal of Robotics Research* **27**, 6 (2008), pp. 737–757.

[183] M. Toussaint and C. Goerick. A Bayesian View on Motor Control and Planning. *From Motor Learning to Interaction Learning in Robots*. Ed. by O. Sigaud and J. Peters. Springer, 2010, pp. 227–252.

[184] H. Triebel. *Höhere Analysis*. Deutsch Harri GmbH, 1980.

[185] J.L. Troutman. *Variational Calculus and Optimal Control: Optimization With Elementary Convexity*. Springer-Verlag, 1996.

[186] J.N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine Learning* **16** (1994), pp. 185–202.

[187] B. Viorel. *Mathematical Methods in Optimization of Differential Systems*. Kluwer Academic Publishers, 1994.

[188] N. Vlassis and M. Toussaint. Model-free reinforcement learning as mixture learning. *Proceedings of the 26th Annual International Conference on Machine Learning*. Montreal, Quebec, Canada: ACM, 2009, pp. 1081–1088.

[189] N. Vlassis et al. Learning model-free robot control by a Monte Carlo EM algorithm. *Auton. Robots* **27**, 2 (2009), pp. 123–130.

[190] Q. Wang et al. Divergence estimation of continuous distributions based on data-dependent partitions. *IEEE Transactions on Information Theory* **51**, 9 (2005), pp. 3064–3074.

[191] X. Wang et al. Convergence rates in the strong law of large numbers for Martingale difference sequences. *Abstract and Applied Analysis* **2012** (2012), p. 13.

[192] D. Werner. *Functional Analysis*. Springer, 2007.

[193] H. Whitney. A function not constant on a connected set of critical points. *Duke Math. J.* **1** (1935), pp. 514–517.

[194] D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transactions on Evolutionary Computation* **1**, 1 (1997), pp. 67–82.

[195] J.S. Yedidia, W.T. Freeman, and Y. Weiss. *Exploring artificial intelligence in the new millennium*. Ed. by G. Lakemeyer and B. Nebel. Morgan Kaufmann Publishers Inc., 2003. Chap. Understanding belief propagation and its generalizations, pp. 239–269.

[196] K. Zahedi, N. Ay, and R. Der. Higher coordination with less control-a result of information maximization in the sensorimotor loop. *Adaptive Behavior* **18**, 3-4 (2010), pp. 338–355.

[197] D. Zarubin et al. Topological synergies for grasp transfer. *Hand Synergies - how to tame the complexity of grapsing, Workshop, IEEE International Conference on Robotics and Automation (ICRA)*. Karlsruhe, Germany, 2013.

Erklärungen

- Ich erkenne die Promotionsordnung der Fakultät für Mathematik und Informatik der Universität Leipzig vom 22.7.2009 in der Fassung der Ersten Änderungssatzung vom 4. Juli 2011 an.

- Die eingereichte Arbeit wurde nicht in gleicher oder ähnlicher Form einer anderen Prüfungsbehörde zum Zwecke einer Promotion oder eines anderen Prüfungsverfahrens vorgelegt.

- Es haben keine früheren erfolglosen Promotionsversuche stattgefunden.

....
(Holger Bernigau)

Daten zum Autor

Name:	Holger Bernigau
Geburtsdatum:	12.08.1983 in Potsdam
10/2004 - 07/2010	Studium der Physik Universität Potsdam und Toronto (Diplom Physik)
09/2010 - 10/2013	Doktorand am *Max-Planck-Institut für Mathematik in den Naturwissenschaften* (gefördert durch die International Max Planck Research School) in der Gruppe *Information Theory of Cognitive Systems* von Herrn Prof. Dr. Nihat Ay

Selbstständigkeitserklärung

Hiermit erkläre ich, die vorliegende Dissertation selbständig und ohne unzulässige fremde Hilfe angefertigt zu haben. Ich habe keine anderen als die angeführten Quellen und Hilfsmittel benutzt und sämtliche Textstellen, die wörtlich oder sinngemäß aus veröffentlichten oder unveröffentlichten Schriften entnommen wurden, und alle Angaben, die auf mündlichen Auskünften beruhen, als solche kenntlich gemacht. Ebenfalls sind alle von anderen Personen bereitgestellten Materialien oder erbrachten Dienstleistungen als solche gekennzeichnet.

Leipzig, den 25.07.2014

. . . .
(Holger Bernigau)