Master Thesis Proposal: Understanding collective feedback behaviour with Gaussian Process Regression

Examiner: Jun.-Prof. Dr. Tatjana Petrov
Department: Modelling of complex, self-organising systems

Background

Population Markov chains are widely used to model the stochastic behaviour of biological populations, such as animal collectives [3]. Discrete entities, or *agents*, make individual decisions and interact with each other, which can be expressed through simple rules within the Markov chain [2, 4]. Unfortunately, the parameters of the model are not always known and individual behaviours cannot be further described. In this work, we analyse a case where only information about the whole population is available, namely as steady-state measurements of the chain (distribution among its BSCCs [1]). The goal is to infer the chain's parameters from this data to analyse individual behaviours. We want to learn more about how biological entities make decisions in a social network.

We discuss a social feedback mechanism in a population of honeybees [5]. In an experiment, a colony of bees is exposed to a threat, whereupon each individual bee may sting, and consequently die. Alarm pheromone is released during stinging, which warns the other bees and increases their aggressiveness, causing them to possibly sting as well. However, the aggressiveness is not endlessly increased; instead, there is a still unknown mechanism that prevents the colony from becoming extinct. We aim to understand how the individual bee is influenced by other bees in the colony from analysing statistical data. The observable outcome of the experiment consists only of the number of living bees. After performing N experiments, the outcomes are captured in a histogram and mapped to the respective BSCCs. From this, the individual behaviour of one bee can be inferred depending on the group size and the amount of pheromone in the system. See Figure 1 for an example of a colony of $n = 3$ bees. On the left side, the population Markov chain with 3 unknown parameters and 4 BSCCs is shown. On the right side, the resulting histogram after a hypothesized experiment is shown, describing the frequency of observing 0, 1, 2, and 3 living bees.

Formal parameter synthesis as well as Bayesian inference methods are developed to estimate the unknown parameters of the chain and derive knowledge about individual behaviours [5]. However, analysing these stochastic models quickly becomes difficult for increasing colony sizes, and experimental data is not available for all possible configurations. Therefore, new methods are required that support the existing parameter inference approaches in order to make predictions of bee colonies when only sparse data is available.

Goals

The goal of this thesis is to develop a data-efficient, scalable method for learning the unknown landscape function that maps the size of the colony to the respective data distribution of surviving bees. Consequently, predictions about colonies of different sizes are possible without performing new experiments. These predictions help to analyse the behaviour of individual bees and how they adapt their response to a threat based on the size of the colony.

Assume that we have experimental data for colonies of $n \in [2, 5, 10]$ bees and the respective histograms. We can map the histograms to the underlying Markov chain and infer the chain's parameters using rational functions provided by the transition probabilities. If every bee of the system is put in a single box, their behaviours are independent, and it is easy to conclude what happens for different colony sizes. The mean number of surviving bees will stay approximately the same, but approaches the true mean with increasing colony size. However, if all bees are in the same box, they influence each other and their behaviours are not independent anymore.
First, we want to apply Machine Learning approaches to learn the most probable data distribution, i.e. the histogram, only from the number of bees in the colony. We aim to describe the dependency between individual bees and different group sizes more clearly by learning the landscape function from available data.
Secondly, we assume that we can control the initial amount of pheromone in the system, and again observe the number of living bees after the experiment. The Machine Learning model should be enriched by this additional input parameter to make even more detailed predictions about the bee's response based on the group size, but also on the amount of pheromone that it is exposed to.
Finally, the overall goal of the project is to describe and understand the parameters of the underlying system, represented in the Markov chain. Therefore, we aim for extending the developed methods to not only predict the group behaviour of the colony (as data distribution of living bees), but also the individual behaviour of a single bee in different colony sizes by learning the chain's parameters. This method is especially useful when the rational functions of the Markov chain cannot be obtained or analytically solved.

Methods

The preferred method to solve the previously described problem is in the field of Machine Learning. On the one hand, Machine Learning approaches are data-driven and well suited to make predictions of a system. This is useful when we do not want to conduct more experiments, but make statements for every possible colony size with only the available data. On the other hand, using a black-box, model-agnostic method has the advantage of being applicable to numerous different models, and not only to one specific case (e.g. a continuous-time model where synchronicity is not assumed).
In general, Machine Learning is a very powerful approach and applied in various different research areas. In the field of systems biology, it is more and more used to model biological systems. Compared to traditional methods, it makes faster predictions and can be used when the computations become too complex due to combinatorial propagation of dependencies.

We want to find a Machine Learning technique to learn the landscape function of the model. More precisely, the function maps the colony size $n \in \mathbb{N}$ to the multinomial distribution with $n + 1$ possible outcomes $(0, 1, ..., n) \in \mathbb{N}^{n+1}$ and the associated frequencies (or probabilities) $(p_0, ..., p_n) \in \mathbb{R}^{n+1}$ such that $\sum p_i = 1$.
A powerful method is *Gaussian process regression* which is a non-parametric Bayesian approach and widely used to describe, learn and optimize unknown functions. One advantage of this methods lies in its output that provides not only a parameter estimate, but additionally a quantification of uncertainty [6]. Validating the chosen model will provide a measure of its performance and evidence if there is enough data available for a well-fitting regression model.

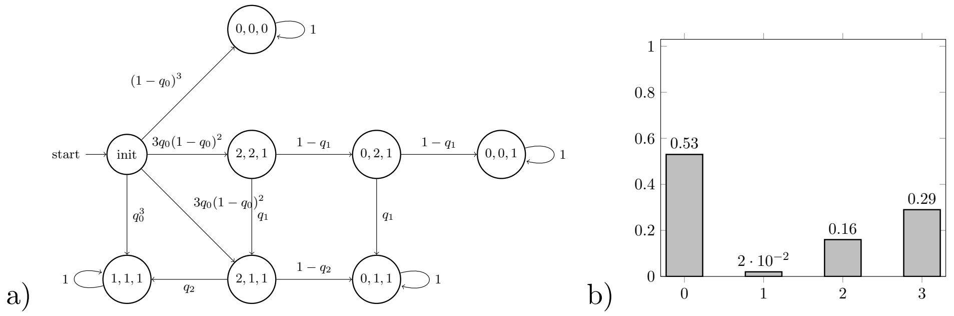

Figure 1: a) Population pMC for colony of $n = 3$ bees. b) Example data histogram of reaching respective BSCC (number of stinging bees) as a result of N experiments.

References

[1] Christel Baier and Joost-Pieter Katoen. *Principles of Model Checking*. Principles of Model Checking. 2008.

[2] Eric Bonabeau. Agent-based modeling: Methods and techniques for simulating human systems. *Proceedings of the National Academy of Sciences of the United States of America*, 99:7280–7287, 2002. doi: 10.1073/pnas.082080899.

[3] Luca Bortolussi and Guido Sanguinetti. *Learning and Designing Stochastic Processes from Logical Constraints*, pages 89–105. Springer Berlin Heidelberg, 2013. doi: 10.1007/978-3-642-40196-1_7.

[4] Donald L. Deangelis and Stephanie G. Diaz. Decision-making in agent-based modeling: A current review and future prospectus. *Frontiers in Ecology and Evolution*, 6, 2019. doi: 10.3389/fevo.2018.00237.

[5] Matej Hajnal, Morgane Nouvian, David Šafránek, and Tatjana Petrov. Data-informed parameter synthesis for population markov chains. *Hybrid Systems Biology (Hsb 2019)*, 11705:147–164, 2019. doi: 10.1007/978-3-030-28042-0_10.

[6] Eric Schulz, Maarten Speekenbrink, and Andreas Krause. A tutorial on gaussian process regression: Modelling, exploring, and exploiting functions. *Journal of Mathematical Psychology*, 85:1–16, 2018.