Faculty of Science and
Bio-Engineering Sciences
Department of Computer Science

Guiding the mitigation of epidemics with reinforcement learning

Dissertation submitted in fulfilment of the requirements for the degree of Doctor of Science: Computer science

Pieter Libin

Herent, 14 March 2020

Promotors: Prof. Dr. Ann Nowé (Vrije Universiteit Brussel)
Prof. Dr. Philippe Lemey (Katholieke Universteit Leuven)

2020 Uitgeverij VUBPRESS Brussels University Press
VUBPRESS is an imprint of ASP nv (Academic and Scientific Publishers nv)
Keizerslaan 34
B-1000 Brussels
Tel. +32 (0)2 289 26 50
Fax +32 (0)2 289 26 59
E-mail: info@aspeditions.be
www.aspeditions.be

ISBN 978 90 5718 970 8
NUR 950
Legal deposit D/2020/11.161/042

To Arwen

Summary

Epidemics of infectious diseases are an important threat to public health and global economies. The most efficient way to combat epidemics is through prevention. To develop prevention strategies and to implement them as efficiently as possible, a good understanding of the complex dynamics that underlie these epidemics is essential. Epidemiological studies allow us to obtain insights in the history of such processes. However, to properly understand such processes, and to study emergency scenarios, epidemiological models are necessary. Such models enable us to make predictions and to study the effect of prevention strategies in simulation. The development of prevention strategies, which need to fulfil distinct criteria (i.a., prevalence, mortality, morbidity, cost), remains a challenging process. For this reason, it is important to study how optimization techniques can be used to support decision makers. In this thesis, we study a reinforcement learning[1] approach to automatically learn prevention strategies.

We investigate two main lines of research. Firstly, we study the decision making problem where a number of possible prevention strategies has been defined by decision makers, who need to determine which of these strategies is most efficient. This decision is made by evaluating the prevention strategies in a complex and computationally demanding epidemiological model. To perform this evaluation efficiently, we investigate the use of algorithms in the field of reinforcement learning that are grounded in the Bayesian uncertainty framework. This approach enables us to learn faster and to quantify the uncertainty of the decisions. To make this possible, we extend existing algorithms and create a new algorithm. Furthermore, we provide theoretical insights in how these algorithms operate.

[1] Reinforcement learning is a field within artificial intelligence that is used to automatically learn by interacting with an environment.

Secondly, we extend this approach such that we can learn adaptive strategies in an epidemiological model. This means that, rather than comparing preventive strategies, we will attempt to learn which subsequent steps are necessary to act optimally, while considering the state of the epidemic. Since the state space of the epidemiological models that are necessary to investigate versatile prevention strategies is huge, we need to represent this space efficiently, in a way that reinforcement learning becomes feasible. To this end, we follow a deep reinforcement learning approach.

We evaluate both research trajectories in the context of pandemic influenza, a pathogen that has made many victims in the past. Our experiments show that our first research trajectory is very useful to evaluate prevention strategies. Furthermore, we show that these techniques will also be useful to support other complex decision making problems that involve computationally demanding models. In the experiments to validate the second research trajectory, we create a new epidemiological model to investigate school closure policies in case an influenza pandemic emerges. To evaluate our learning technique, we present a new method to establish a ground truth, through which we show that our learning technique approximates the optimal strategy. Finally, we investigate whether there is a collaborative advantage when designing school closure policies. We formulate this research question as a multi-agent problem and attempt to solve this problem using deep multi-agent reinforcement learning techniques.

Samenvatting

Epidemieën van infectieziekten vormen een belangrijke bedreiging voor de volksgezondheid. De meest efficiënte manier om infectieziekten in te dijken is door middel van preventie. Om preventiestrategieën uit te denken en zo efficiënt mogelijk te implementeren is een goed begrip nodig van de complexe dynamiek waarmee dergelijke pathogenen zich in een populatie verspreiden. Epidemiologische studies laten ons toe om inzichten te verwerven in de geschiedenis van zulke processen. Echter, om deze processen beter te begrijpen, is het nodig om epidemiologische modellen te bouwen. Zulke modellen laten ons toe om voorspellingen te maken, elementen te identificeren waarop preventie kan worden toegespitst en het effect van preventiestrategieën te bestuderen. Het ontwikkelen van preventiestrategieën, dewelke aan verschillende criteria dienen te voldoen (i.a., prevalentie, mortaliteit, morbiditeit, kost), blijft echter een zeer uitdagend proces. Daarom is het belangrijk om te onderzoeken hoe optimalisatietechnieken gebruikt kunnen worden om dit proces te ondersteunen. In deze thesis hebben we het gebruik van bekrachtigingsleren[2] onderzocht om optimale preventiestrategieën automatisch te leren.

We hebben twee onderzoeksrichtingen uitgewerkt. Ten eerste hebben we het probleem bestudeerd waarbij er een aantal preventiestrategieën gedefinieerd is, en we willen bepalen welk van deze strategieën het meest efficiënt is door de strategieën te evalueren in een complex en computationeel veeleisend epidemiologisch model. Om deze evaluatie efficiënt aan te pakken bestuderen we algoritmes uit het domein van bekrachtigingsleren die gebaseerd zijn op het Bayesiaanse onzekerheidsraamwerk. Deze aanpak geeft ons de mogelijkheid om sneller te leren en de onzekerheid van beslissingen te kwantificeren. Om dit mogelijk

[2] Bekrachtigingsleren is een deelveld van artificiële intelligentie dat gebruikt wordt om te leren door interactie met een omgeving

te maken hebben we bestaande algoritmes aangepast en nieuwe algoritmes ontwikkeld. Daarnaast hebben we ook een theoretisch inzicht over deze algorithmes gegeven.

Ten tweede breiden we dit uit zodat we adaptieve strategieën kunnen leren binnen een epidemiologisch model. Dus, in plaats van strategieën te vergelijken, trachten we te leren welke opeenvolgende stappen er nodig zijn die rekening houden met de toestand van de epidemie, om optimaal te handelen. Aangezien de toestandsruimte van de epidemiologische modellen die nodig zijn om veelzijdige preventiestrategieën te beschouwen zeer groot is, moeten we deze ruimte zo voorstellen dat de bekrachtigingsleertechniek hierover kan redeneren. Dit doen we door middel van diepe neurale netwerken , i.e., *deep learning*.

We evalueren deze onderzoekstrajecten in de context van pandemische influenza, een epidemie die in het verleden al erg veel slachtoffers heeft gemaakt. Onze experimenten tonen aan dat ons eerste onderzoekstraject zeer nuttig is voor het vergelijken van preventiestrategieën. Verder tonen we aan dat deze technieken ook nuttig zijn om andere complexe beslissingsprocessen te ondersteunen, waar computationeel veeleisende modellen aan te pas komen. In het tweede onderzoekstraject maken we een specifiek model om het sluiten van scholen te onderzoeken indien er een influenza pandemie uitbreekt. In dit model tonen we dat onze leertechnieken de optimale strategie kunnen benaderen. Daarenboven onderzoeken we of er een positief effect is als verschillende geografische districten samenwerken om te beslissen wanneer scholen gesloten dienen te worden. We formuleren deze onderzoeksvraag als een *multi-agent* probleem en trachten dit op te lossen aan de hand van *multi-agent* bekrachtigingsleren in combinatie met diepe neurale netwerken.

Acknowledgments

This book marks the final milestone of a five year research journey that I truly enjoyed. I'm grateful that I got this amazing opportunity to do research, most probably one of the best *jobs* there is, and for all the wonderful people that I met along the way.

I would like to thank my promoters prof. dr. Ann Nowé and prof. dr. Philippe Lemey. Thanks Ann, for inviting me to your lab, giving me this great opportunity, for all the scientific freedom to find my own research direction and to shape me as the critical researcher I've become. Thanks Philippe, for passing on your research rigor, teaching me the skills to present my ideas in a crisp way and to remind me to focus on a limited set of projects. I look forward to continue to collaborate with both of you in the near future.

I thank my jury, for reading this manuscript in great detail, and asking me challenging questions that allowed me to gain new insights. Thanks prof. dr. Bernard De Baets (Universiteit Gent), prof. dr. Enda Howley (National University of Ireland Galway), prof. dr. Dominique Maes (VUB), prof. dr. Tom Lenaerts (VUB/ULB) and the president of the jury prof. dr. Wolfgang De Meuter (VUB). Thanks for your notes that enabled me to polish this book.

I want to thank dr. Kristof Theys and dr. Diederik Roijers for their close collaboration and guidance. Through your experience, I was able to advance both my research outcomes and personal research skills, for which I am truly grateful.

While it took quite some effort, I was lucky enough to receive funding from the Flemish research foundation (FWO). This personal research grant supported me to develop a new methodology and gave me the freedom to connect with researchers around the globe, for which I am grateful.

Thanks to all the colleagues from the AI lab (VUB) and the CEV lab (KU Leuven), for the amazing workspace and the pleasant and fruitful collaborations. In particular I want

to speak out to a few of you. Arno, thanks for your continuous support and the fun while investigating our deep RL project. Felipe, thanks for sharing an office with me for almost five years, it has been great fun, and I hope you will be as successful in your company as you were in your research. And finally, I want to thank Timo. Thanks for the cool research ideas and projects that we jointly investigated, and your endless patience to explain me things. While the conferences were great fun, I think I enjoyed most our long sessions at the blackboard, that I will miss a lot.

I also want to thank our secretaries, Brigitte and Lydie, for their continuous help and nice talks.

Thanks to all the researchers from around the globe that I had the pleasure to collaborate with. A special thanks to: prof. dr. Anne-Mieke Vandamme for introducing me to the academic world in her lab, prof. dr. Ana Abecasis for the nice collaboration and research visits in sunny Lisbon, prof. dr. Niel Hens for the support with meta-population modelling and for the research opportunity that is to come, and the late dr. Ricardo Camacho for his help with many ideas when my main focus was on still on HIV.

Thanks to my friends. And especially Vincent, Kristof, Francis and Kristof, for listening to me ramble about my research, and making it look like you were truly interested.

I want to thank my parents, brother and meter for their support, throughout all my life. And a special thanks for your support for the pursuit of my master degree and the PhD study at the unholy age of 30+.

Finally I want to thank my lovely wife Karen, for all the good times, but also for the patience, the support, your love, and our amazing daughter.

Preface

This thesis presents the research that was conducted to investigate how the mitigation of epidemics can be guided with reinforcement learning[3]. This work is the result of a multi-disciplinary research effort that advances the research fields of epidemiological modelling, machine learning and epidemiology. At the foundation of this thesis lies a substantial amount of exploration that led to additional research results, that I will introduce in this chapter. Here, I focus on papers on which I'm one of the main contributors (first author, or shared first author, shared first authorship is annotated with a $*$). All other research results are enumerated in my academic CV (Chapter 14).

Related to the results presented in this thesis, I explored the field of arboviruses and HIV. Firstly, the work I conducted on arboviruses, allowed me to set up some collaborations that resulted in two first-author journal publications that concern the nomenclature and genotyping of arboviruses (see Section 1 and Section 2). Secondly, the work I conducted with respect to HIV resulted in a first-author survey journal paper about HIV transmission (see Section 3) and a first-author methodological journal paper on how to understand HIV epidemics by investigating phylogenetic trees (see Section 4). In collaboration with Master students, I also worked on the modelling of HIV epidemics: the modelling of the undiagnosed HIV population (see Section 6) and the inference of set-point viral load transmission models (see Section 5).

[3]The research presented in this thesis was funded by the VUB research council (from 01/01/2015 to 31/12/2015) and a personal FWO SB grant (from 01/01/2016 to 31/12/2019).

1 Time to harmonize Dengue nomenclature and classification

Cuypers L, **Libin P***, Simmonds P, Nowé A, Muñoz-Jordán J, Alcantara L, Vandamme AM, Santiago G, Theys K., "Time to Harmonize Dengue Nomenclature and Classification.", Viruses, vol. 10, issue 10, 2018. [Peer reviewed, 2-yearly JCR impact factor 2018: 3.811] (* denotes equal contribution)*

Dengue virus (DENV) is estimated to cause 390 million infections per year worldwide. A quarter of these infections manifest clinically and are associated with a morbidity and mortality that put a significant burden on the affected regions. Reports of increased frequency, intensity, and extended geographical range of outbreaks highlight the ongoing global viral spread. Persistent transmission in endemic areas and the emergence in territories formerly devoid of transmission have shaped DENV's current genetic diversity and divergence. This genetic layout is hierarchically organized in serotypes, genotypes, and sub-genotypic clades. While serotypes are well defined, the genotype nomenclature and classification system lack consistency, which complicates a broader analysis of their clinical and epidemiological characteristics.

We identify five key challenges: (1) Currently, there is no formal definition of a DENV genotype; (2) Two different nomenclature systems are used in parallel, which causes significant confusion; (3) A standardized classification procedure is lacking so far; (4) No formal definition of sub-genotypic clades is in place; (5) There is no consensus on how to report antigenic diversity.

Taking in to account these challenges, our aim was to raise awareness that the time is right to re-evaluate DENV genetic diversity. This effort will benefit greatly from the thousands of DENV genome sequences already available in the public domain. Advances in methodologies for DENV surveillance and guided sampling strategies, as well the potential use of DENV whole-genome sequencing in a clinical and epidemiological context, further illustrates the urgency to question the current DENV taxonomy. Here, our primary intent is to raise attention of the need for a re-evaluation by identifying challenges that affect the increasing importance of DENV genomics in understanding virus disease manifestations and epidemic spread. A re-evaluation will provide harmonization across DENV studies and guide scientists to construct tools to detect outbreaks and infer epidemiological trends.

2 Automated genotyping of arboviruses

Vagner Fonseca, **Pieter J K Libin***, Kristof Theys*, Nuno R Faria, Marcio R T Nunes, Maria I Restovic, Murilo Freire, Marta Giovanetti, Lize Cuypers, Ann Nowé, Ana Abecasis, Koen De-forche, Gilberto A Santiago, Isadora C de Siqueira, Emmanuel J San, Kaliane C B Machado, Vasco Azevedo, Ana Maria Bispo-de Filippis, Rivaldo Venâncio da Cunha, Oliver G Pybus, Anne-Mieke Vandamme, Luiz C J Alcantara, Tulio de Oliveira, "A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes", PLoS Neglected Tropical Diseases, vol. 13, issue 5, 2019. [Peer reviewed, 2-yearly JCR impact factor 2018: 4.487] (* denotes equal contribution)*

In the recent years, an increasing number of outbreaks of Dengue (DENV), Chikungunya (CHIKV) and Zika (ZIKV) viruses have been reported in Asia and the Americas. The predominant mosquito species transmitting DENV, CHIKV and ZIKV, are Aedes aegypti and Aedes Albopictus, which are widely distributed in tropical and sub-tropical regions. In the past few years, several studies have reported concurrent outbreaks of DENV, CHIKV and ZIKV in the same geographical area.

Monitoring virus genotype diversity is crucial to understand the emergence and spread of outbreaks, both aspects that are vital to develop effective prevention and treatment strategies. Both DENV and CHIKV epidemics are associated with a mortality and morbidity that puts a significant economic burden on the affected regions. While infections with ZIKV are rarely fatal, ZIKV infections may result in Guillain-Barré syndrome and congenital malformations. Genomic surveillance of epidemics at the appropriate resolution enables the identification of strains associated with greater epidemic potential or disease severity.

However, methods that consistently classify arbovirus sequences at the level of species and sub-species (i.e., serotype and/or genotype) are currently lacking. Additionally, whole genome sequences are often not available in routine clinical settings, forcing the use of shorter gene sequences to classify at viral species or sub-species level. It has however insufficiently been explored which genomic regions are most suitable for accurate classification.

A new computational method for the identification of DENV/CHIKV/ZIKV sequences, with respect to species and sub-species, is presented. The classification method was validated on a large dataset by assessing the classification performance of whole-genome sequences, partial-genome sequences and products from next-generation sequencing methods. Furthermore, the suitability of different genomic regions for virus classification was evaluated.

3 The impact of HIV-1 within-host evolution on transmission dynamics

Kristof Theys, **Libin P.***, Andrea-Clemencia Pineda-Peña,Ann Nowé,Anne-mieke Vandamme,Ana B Abecasis, "The impact of HIV-1 within-host evolution on transmission dynamics", Current Opinion in Virology, vol. 28, p 92 - 101, 2018. [Peer reviewed, 2-yearly JCR impact factor 2018: 5.4] (* denotes equal contribution)*

Despite the remarkable progress in the last three decades, Human Immunodeficiency Virus type 1 (HIV-1) remains a global health threat with almost two million new infections and one million AIDS-related deaths in 2016. HIV-1, as many RNA viruses, is characterized by high rates of evolutionary changes during the course of infection, and this potential for adaptive evolution has proven to be a cornerstone mechanism of viral escape from host immune responses and antiretroviral treatment. Yet, drug resistance rates are stabilizing or decreasing due to more efficacious and tolerable regimens, particularly in developed countries, resulting in a nearly normal life expectancy for treated HIV-1 infected individuals.

By contrast, as immune-driven viral control by means of a preventive or therapeutic vaccine remains challenging, main priorities with respect to public health now concern the formulation and implementation of efficient and effective prevention strategies. The widespread implementation of Treatment as Prevention is considered one of the most important intervention strategies to reduce the rate of HIV-1 transmission. Furthermore, methodological improvements to infer transmission networks and to model epidemics in silico enable authorities to monitor dynamics of epidemic spread and ameliorate prevention strategies. Despite these advances, the resulting worldwide decline in HIV-1 infections is considered insufficient to meet the UNAIDS visionary goal to end AIDS by 2030. Periodical reports of increasing incidence (e.g. Eastern Europe and Central Asia) and persistent challenges (e.g. late presenters, suboptimal adherence and increasing rates of drug resistance in Africa) illustrate the continuous need to optimize targeted intervention strategies.

To this end, novel perspectives on the impact of HIV-1 within-host evolutionary processes can be of high value to better understand and mitigate population-level transmission. Knowledge on the association between viral genotype and epidemic potential is crucial to formulate prevention strategies directed towards sub-epidemics that have the largest impact on the incidence rate. Furthermore, insights into the different mechanisms and virus diversity that dominate transmission and early infection will improve transmission history inference as well as vaccine design.

HIV-1 is characterized by extensive genetic diversity within and between hosts, but distinct processes and rates shape viral evolution at these levels. While within-host evolutionary

dynamics are dominated by selective forces and competitive fitness, between-host evolution is considered to be largely shaped by neutral processes and multiple HIV-1 variants co-exist at the epidemiological level, although recent findings might imply more selective action at this level than generally assumed. Circulating viral diversity at population level is an intrinsic reflection of transmission dynamics within the epidemic, but virus and host genetics additionally impact between-host dynamics. As a result, a complex interplay of multi-scale evolutionary processes exists and the selective advantage of viral traits differs at the within-host and between-host level as conflicting evolutionary forces apply. While viral spread in a population is expected to select for traits that maximize transmission efficiency, viral strategies that favour within-host fitness do not necessarily benefit epidemiological fitness, leading to a delicate evolutionary trade-off between virulence and transmission.

We discuss the connection between HIV-1 within-host evolution and transmission dynamics and highlight recent theoretical and experimental work bridging different scales. We first address the processes that shape HIV viral evolution within the host, then we focus on the bottleneck transmission event and finally on how this can be translated at the population level. We end by shedding light on how these factors can affect the epidemiological transmission studies. Understanding the link between within-host processes, transmission fitness and epidemic spread will be essential to achieve HIV-1 eradication.

4 Interactively exploring large phylogenies

***Pieter Libin**, Ewout Vanden Eynden, Francesca Incardona, Ann Nowé, Antonia Bezenchek, An-ders Sönnerborg, Anne-Mieke Vandamme, EucoHIV Study Group, Kristof Theys, Guy Baele, "PhyloGeoTool: interactively exploring large phylogenies in an epidemiological context", Bioinformatics, vol. 33, issue 24, p. 3993 - 3995, 2017. [Peer reviewed, 2-yearly JCR impact factor 2018: 4.531]*

Expanding and intensifying sequencing efforts for the management of infectious diseases along with the generation of large-scale databases of clinical and demographical information provide unprecedented opportunities for the surveillance of epidemics and outbreaks of viral pathogens. Mapping the origin and the dynamics of epidemics in space and time is becoming feasible as geo-tagged and time-stamped sequence data are now part of routine clinical care. Tracking the geographical spread and the relationship to specified characteristics for distinct virus clades (e.g., transmission risk group, tropism, drug resistance profile) can help to improve our understanding of such outbreaks. Computational and methodological advances now allow to infer phylogenies of tens of thousands of sequences and applications have been developed to visualize such large phylogenetic trees. However, efficient means to visually navigate through these large phylogenies and the annotated

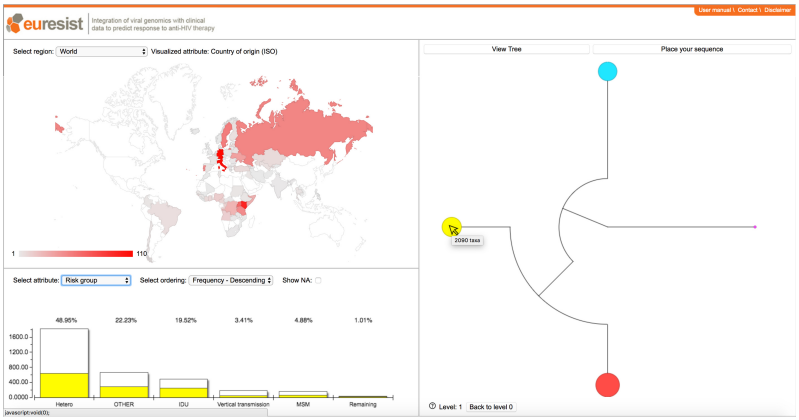


Figure 1: The PhyloGeoTool graphical user interface. The upper left panel shows the geographical distribution of the samples present in the selected cluster. The lower left panel shows the distribution for a selected trait of interest; white bars show the distribution for the entire dataset for that level of the tree, whereas the colored bars show the distribution for a specific selected cluster and are annotated by their respective percentage. The right panel shows the clustered phylogenetic tree and allows to perform phylogenetic placement

information, are currently still lacking. Furthermore, we envision the need for fast and accurate placement of novel virus sequences onto an existing phylogenetic tree, as this can provide valuable insights for outbreak detection, by relating evolutionary dynamics to epidemiological and clinical characteristics.

To advance the state-of-the-art, we present PhyloGeoTool (Figure 1), an application to interactively navigate large phylogenies and to explore associated clinical and epidemiological data. PhyloGeoTool implements an algorithm that automatically partitions a phylogeny into an optimal number of clusters, thereby recursively partitioning each identified cluster.

To demonstrate PhyloGeoTool's potential, we present a case study concerning transmitted HIV-1 drug resistance in Europe and evaluated PhyloGeoTool in the context of the Dengue virus.

5 Bayesian inference of set-point viral load transmission models

Pieter Libin*, *Laurens Hernalsteen**, *Kristof Theys, Perpetua Gomes, Ana Abecasis, Ann Nowé, "Bayesian inference of set-point viral load transmission models", Benelux Artificial Intelligence Conference, p.: 107 - 12, 2018. (* denotes equal contribution)*

When modelling HIV epidemics, it is important to incorporate set-point viral load and its heritability. As set-point viral load distributions can differ significantly amongst epidemics, it is imperative to account for the observed local variation. This can be done by using a heritability model and fitting it to a local set-point viral load distribution. However, as the fitting procedure needs to take into account the actual transmission dynamics (i.e., social network, sexual behaviour), a complex model is required. Furthermore, in order to use the estimates in subsequent modelling analyses to inform prevention policies, it is important to assess parameter robustness.

In order to fit set-point viral load models without the need to capture explicitly the transmission dynamics, we present a new protocol. Firstly, we approximate the transmission network from a phylogeny that was inferred from sequences collected in the local epidemic. Secondly, as this transmission network only comprises a single instance of the transmission network space, and our aim is to assess parameter robustness, we infer the transmission network distribution. Thirdly, we fit the parameters of the selected set-point viral load model on multiple samples from the transmission network distribution using approximate Bayesian inference.

Our new protocol enables researchers to fit set-point viral load models in their local context, and diagnose the model parameter's uncertainty. Such parameter estimates are essential to enable subsequent modelling analyses, and thus crucial to improve prevention policies.

6 Towards a phylogenetic measure to quantify HIV incidence

Libin, P.*, *Versbraegen, N.*, Abecasis, A. B., Gomes, P., Lenaerts, T., and Nowé, A., "Towards a phylogenetic measure to quantify HIV incidence.", Benelux Conference on Artificial Intelligence, 2019. (* denotes equal contribution)*

One of the cornerstones in combating the HIV pandemic is the ability to assess the current state and evolution of local HIV epidemics. This remains a complex problem, as many HIV infected individuals are unaware of their infection status, leading to parts of HIV epidemics being undiagnosed and under-reported.

An abundance of clinical data is available in the context of HIV epidemics, as upon diagnosis a number of tests are performed and the results thereof collected. One of those tests determines the specific genotype of the virus infecting a patient. To that purpose, the genetic sequence of the virus is determined. As a result, a vast amount of HIV sequences have been collected over the last decades.

We first present a method to learn epidemiological parameters from phylogenetic trees, using approximate Bayesian computation. The epidemiological parameters learned as a result of applying approximate Bayesian computation are subsequently used in epidemiological models that aim to simulate a specific epidemic. Secondly, we continue by describing the development of a tree statistic, rooted in coalescent theory, which we use to relate epidemiological parameters to a phylogenetic tree, by using the simulated epidemics. We show that the presented tree statistic enables differentiation of epidemiological parameters, while only relying on phylogenetic trees, thus enabling the construction of new methods to ascertain the epidemiological state of an HIV epidemic. By using genetic data to infer epidemic sizes, we expect to enhance our understanding of the portions of the infected population in which diagnosis rates are low. This understanding will allow for more effective health policies, through diagnosis strategies that are directed towards these particular sub-populations.

We validate our research on the HIV-1 epidemic in Portugal. We therefore first present inference of the epidemiological parameters of said epidemic by applying approximate Bayesian computation. We apply approximate Bayesian computation to fit a model that contains the epidemiological parameters in question.

Contents

Summary **5**

Samenvatting **7**

Acknowledgments **9**

Preface **11**
1 Time to harmonize Dengue nomenclature and classification 12
2 Automated genotyping of arboviruses 13
3 The impact of HIV-1 within-host evolution on transmission dynamics . . . 14
4 Interactively exploring large phylogenies 15
5 Bayesian inference of set-point viral load transmission models 17
6 Towards a phylogenetic measure to quantify HIV incidence 17

Contents **19**

Nomenclature **25**

1 Introduction **29**
1 Infectious diseases 30
2 Pandemic influenza 31
3 Studying epidemics 32
4 Studying the effect of interventions 36

4.1 Objective of the intervention . . . 37
4.2 Therapeutic and non-therapeutic interventions . . . 38
4.3 Optimal use of resources . . . 39
5 Reinforcement learning . . . 40
6 Research objectives and contributions . . . 41

2 Multi-armed bandits and reinforcement learning 45
1 Multi-armed bandit . . . 46
2 Cumulative regret . . . 48
2.1 ϵ-greedy . . . 48
2.2 Upper confidence bound . . . 50
2.3 Bayesian inference . . . 51
2.4 Thompson sampling . . . 54
3 The reinforcement learning problem . . . 59
4 Temporal difference learning . . . 62
5 Artificial neural networks . . . 64
6 Deep Q-networks . . . 66
7 Policy gradient . . . 68

3 Epidemiological models 71
1 Compartment models . . . 71
1.1 SIR model . . . 72
1.2 SEIR model . . . 76
1.3 Age-heterogeneous SIR model . . . 77
1.4 Stochastic SIR model . . . 79
2 Individual-based models . . . 82
3 Meta-population models . . . 84

4 Bayesian bandits for decision making in an influenza pandemic 85
1 Rationale and objectives . . . 86
2 Background . . . 88
2.1 Pandemic influenza and vaccine production . . . 88
2.2 Modelling influenza . . . 88
2.3 Bandits and best-arm identification . . . 89
3 Related work . . . 89
4 Epidemic bandits . . . 90

4.1 Evaluating preventive strategies with bandits . . . 90
4.2 Outcome distribution . . . 91
4.3 Epidemic fade-out threshold . . . 91
4.4 Best-arm identification with a fixed budget . . . 92
4.5 Probability of success . . . 97
5 Experiments . . . 97
5.1 Influenza model and configuration . . . 98
5.2 Formulating vaccine allocation strategies . . . 98
5.3 An influenza preventive bandit . . . 99
5.4 Outcome distributions . . . 99
5.5 Best-arm identification experiment . . . 100
6 Future work . . . 104
7 Discussion . . . 105

5 Bayesian Anytime m-top Exploration 107
1 Rationale and objectives . . . 108
2 Background: AT-LUCB . . . 110
3 Related work . . . 110
4 Boundary Focused TS . . . 111
5 Experiments . . . 113
5.1 Gaussian bandit with fixed variance . . . 114
5.2 Cartoon caption bandit . . . 116
5.3 Pandemic bandit . . . 118
5.4 Organic bandit . . . 120
5.5 Overall performance of BFTS . . . 121
6 Bayesian analysis of BFTS . . . 122
6.1 Empirical validation of the heuristics . . . 125
7 Discussion . . . 127

6 Spatial model for pandemic influenza 131
1 Rationale and objectives . . . 132
2 Related work . . . 134
3 Intra-patch age-dependent SEIR model . . . 134
3.1 Contact matrix . . . 135
3.2 Census data . . . 137
3.3 Parametrising the model with R_0 . . . 140

3.4 Stochastic trajectories . . . 141
3.5 Compartment model parameters . . . 142
4 Between-patch model . . . 142
4.1 Time scale transformation algorithm . . . 143
4.2 TSTA to model the infection of patches . . . 147
5 Model validation . . . 148
5.1 Comparison to the Eames SEIR compartment model . . . 149
5.2 2009 influenza pandemic in Great Britain . . . 153
6 Computational complexity and performance . . . 156
7 Discussion . . . 157

7 Studying school closure policies following a reinforcement learning approach 161
1 Rationale and objectives . . . 162
2 Related work . . . 163
3 Learning environment . . . 164
4 Epidemiological setting . . . 166
5 Proximal Policy Optimization and Deep Q-Networks . . . 166
6 Studying the effect of the population composition . . . 167
6.1 Establishing a ground truth . . . 168
6.2 Selecting districts . . . 168
6.3 Exhaustive policy search . . . 171
7 Evaluate PPO with respect to the ground truth . . . 173
8 Analysing the mobility network to partition . . . 175
9 Multi-agent reinforcement learning . . . 179
10 Reward function to investigate policies to shift the epidemic's peak day . . . 182
11 Discussion . . . 183

8 General discussion 189
1 Contributions . . . 189
2 Dissemination and valorisation . . . 192
3 Future work . . . 193

A Appendices 195
1 BayesGap simple regret bound for t-distributed posteriors . . . 195
2 Epidemic bandit experiments: computational details . . . 205
2.1 FluTE source . . . 205

2.2 FluTE configurations 205
2.3 Bandit implementation 205
2.4 High performance computing 205
3 Outcome (i.e., epidemic size) distributions 207
4 Bandit run success rates 208
5 P_s values for Top-two Thompson sampling 209
6 Binned distribution of P_s values for Top-two Thompson sampling 210
7 Expectation of the truncated t-distribution posterior 211
8 Empirical validation of BFTS' heuristics 213
9 Influenza model validation 216
10 Comparing PPO and DQN 218
11 Histograms for the top policies (attack rate) 219
12 PPO learning curves ($R_0 = 1.8$) 225
13 PPO learning curves ($R_0 = 2.4$) 226
14 Comparing PPO to the ground truth (attack rate) 227

Curriculum vitae **229**

Bibliography **239**

Nomenclature

K	number of arms in a multi-armed bandit (MAB), page 47
a_k	k-th arm of a MAB, page 47
r_k	reward sampled from a MAB's k-th arm, page 47
μ_k	expected reward of a MAB's k-th arm, page 47
$\boldsymbol{\mu}$	expected reward vector of a MAB, page 47
$\mathbb{E}[.]$	expectation operator, page 47
a_*	best arm of a MAB, page 47
$a^{(t)}$	the arm pulled at time t, page 47
$\mu_{a^{(t)}}$	expected reward of the arm pulled at time t, page 47
$R^{(T)}$	simple regret at time T, page 48
$J^{(T)}$	recommended arm at time T, page 48
$\hat{\mu}_k^{(t)}$	empirical reward of a MAB's k-th arm, at time t, page 48
$n_k^{(t)}$	number of times arm k was pulled at time t, page 50
$\mathcal{B}$eta	Beta distribution, page 53
$\mathcal{H}^{(t-1)}$	observed history of arm pulls and rewards until time $t-1$, page 57

$\pi(.)$	prior over the means of a MAB, posterior when conditioned on the history, page 57
$\tilde{\boldsymbol{\mu}}^{(t)}$	estimate for the means $\mu_{1..K}$, sampled from posterior $\pi(\cdot \mid \mathcal{H}^{(t-1)})$, page 57
Ψ_ρ	Thompson sampling ranking operator, returns the ρ-ranked arm, page 57
$\mathcal{S}$	state space of a Markov Decision Process (MDP), page 61
$\mathbf{s}$	state in a MDP, page 61
$\mathcal{A}$	MDP's action space, page 61
$\mathbf{a}$	action in a MDP, page 61
T	MDP's transition probability, page 61
d	MDP's discount factor, page 61
R	MDP's reward function, page 61
p	policy to interact with an MDP, page 61
V^p	value function in terms of an MDP and a policy p, page 62
Q^p	Q function in terms of an MDP and a policy p, page 62
p^*	optimal policy to interact with an MDP, page 62
α	reinforcement learning learning rate, page 63
R_0	basic reproductive number, page 74
$S(t)$	susceptible individuals in the SEIR model at time t, page 76
$E(t)$	exposed individuals in the SEIR model at time t, page 76
$I(t)$	infected individuals in the SEIR model at time t, page 76
$R(t)$	recovered individuals in the SEIR model at time t, page 76
β	probability of infection in the SEIR model, page 76
χ	contact rate in the SEIR model, page 76
ζ	latency rate in the SEIR model, page 76
γ	recovery rate in the SEIR model, page 76

M	contact matrix in the age-dependent SEIR model, page 78
$\mathcal{W}$	Wiener process, page 81
$\mathcal{E}$	outcome distribution for an epidemiological model, page 90
p_{ext}	probability of epidemic extinction, page 92
T_0	probability of epidemic extinction threshold, page 92
$\mathcal{T}$	Student's t-distribution, page 94
$S_k^{(t)}$	sum of squares of arm k at time t, page 94
$\mathbb{V}[.]$	variance operator, page 94
$\mathcal{U}$	uniform distribution, page 101
$\mathcal{B}$er	Bernoulli distribution, page 112
$\mathcal{J}^{(t)}$	the set of recommended arms at time t, page 114
$\mathcal{N}_{[a,b]}$	Gaussian distribution, truncated on the interval $[a, b]$, page 115
$\mathcal{N}$	Gaussian distribution, page 116
$\mathbb{S}^D$	unit simplex, page 138
$\mathcal{P}$	set of patches in the meta-population model, page 142
$\mathcal{M}$	mobility matrix, page 142
$\perp\!\!\!\perp$	independence operator, page 144
$\mathcal{P}$	Homogeneous Poisson process, page 144
$\mathcal{P}^{\lambda(t)}$	Non-Homogeneous Poisson process with rate $\lambda(t)$ at time t, page 144
$\inf$	infimum, page 146
$\mathcal{b}^{(t)}$	school closure budget at time t, page 165
$\mathbb{R}$	real numbers, page 165

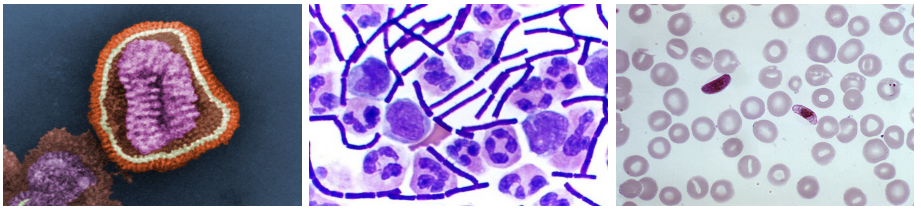
1 | Introduction

> La dernière chose qu'on trouve en faisant un ouvrage est de savoir celle qu'il faut mettre la première.
>
> Blaise Pascal, *Pensées*

In this dissertation, we contribute to the decision making process that aims for strategies to mitigate epidemics of infectious diseases. To this end, we follow a reinforcement learning approach to develop new methods and techniques to investigate mitigation policies in epidemiological models. In this chapter, we aim to situate this multi-disciplinary research project in the different related research domains. We start by providing background on infectious diseases and the challenges they impose. We continue by presenting more details on pandemic influenza, an infectious disease that has caused several devastating epidemics throughout human history, and the pathogen on which most of the experiments in this dissertation will focus. Next, we situate our work in the broad field of research disciplines that study epidemics. Subsequently, we introduce the concepts that are key to study intervention strategies aimed at the mitigation of epidemics, and we introduce the machine learning technique reinforcement learning, on which our methodological approach will be based upon. Finally, we provide a detailed description of our research objectives and highlight the contributions that will be presented in this dissertation.

1 Infectious diseases

Infectious diseases have been around since the dawn of time, with infectious agents, i.e., pathogens, causing disease in plants, animals and humans. A pathogen is an infectious micro-organism, including viruses (e.g., influenza), bacteria (e.g., anthrax), protozoa (e.g., malaria parasite) and prions [Alberts et al., 2002], as shown in Figure 1.1.



(a) Digitally-colourized electron microscopic image of an influenza virion.

(b) Cerebrospinal fluid showing anthrax bacilli (purple rods).

(c) Blood smear that reveals a *Plasmodium falciparum* parasite.

Figure 1.1: Images showing a virion (a), bacilla (b) and protozoa (c). Figure (a) and (c) were authored by Frederick Murphy and downloaded from the CDC Public Health Image Library. Figure (b) was authored by Jernigan et al. [2001].

We as humans are most familiar with pathogens through the burden they put on us and our economy, either by influencing our agricultural processes (i.e., pathogens that impact our live stock or crops) or by affecting our health directly. The latter has severely impacted human history. An example, that possibly speaks most to the imagination, is the bubonic plague, significantly dubbed the black death[1]. The plague is caused by the bacterium *Yersinia pestis*, and was responsible for a decimation of the world population in the 14th century [Parkhill et al., 2001]. In Europe alone, it is estimated that the plague killed 30% to 60% of the population [Spyrou et al., 2019]. Evidently, times have changed since the 14th century, as we now know that the plague is caused by a bacterium[2] and is spread via flea bites, enabling us to take the necessary precautions to avoid infection.

Furthermore, thanks to the development of antibiotics[3], people infected with *Yersinia pestis* can be treated effectively, thereby significantly reducing the mortality rate [Prentice and Rahalison, 2007].

[1] Because of the black colouration of the patient's skin due to subcutaneous haemorrhages.

[2] Germ theory, the theory that pathogens (i.e., germs) can lead to disease, was only popularized in the 19th century by Louis Pasteur.

[3] Although antibiotic resistance is a growing concern, also for *Yersinia pestis* [Galimand et al., 2006].

Yet, knowing a pathogen and how it spreads does not guarantee that we can avoid major pandemics. This was demonstrated at the beginning of the twentieth century (1918-1920), when an influenza strain commonly referred to as "Spanish flu" swiftly spread throughout the world, killing millions of people [Taubenberger and Morens, 2006].

Nevertheless, we have come a far way. We have a good scientific understanding of how pathogens spread, and this understanding continues to grow every day. We have prevention strategies ranging from vaccines to bacterial and antiviral treatments [Behbehani, 1983; Clardy et al., 2009; Marseille et al., 2002], ever-improving ways to chart ongoing epidemics [Grubaugh et al., 2019] and sophisticated means to reason about epidemics and their control [Diekmann et al., 2012].

Nonetheless, in recent history, we found ourselves defenceless against a new virus, i.e., the human immunodeficiency virus (HIV). Only quite recently, the development of anti-retroviral drugs enabled us to turn the tide [Moore and Chaisson, 1999]. Mosquito-borne infections, not the least malaria [Chen et al., 2018], but also the Dengue and Zika virus [Messina et al., 2019; Petersen et al., 2016], continue to plague the peoples in the tropics. The Ebola virus keeps sparking new epidemics that are as devastating as they prove hard to control [Dudas et al., 2017]. Moreover, we see the insurgence of pathogens that we considered to be eliminated, e.g., the measles are resurfacing throughout the Western world [Holzmann and Wiedermann, 2019]. And to conclude this depressing enumeration, the author is writing this introduction in lockdown, as we attempt to contain the ongoing COVID-19 pandemic [Zhu et al., 2020].

All this indicates that many challenges remain, some of which we will try to address in this dissertation.

2 Pandemic influenza

In this dissertation, we will develop new methods to help decision makers to reduce the burden of infectious diseases. While these methods are applicable to a wide range of pathogens, we will perform our experiments in the context of pandemic influenza. To this end, we will provide some background on influenza pandemics, and explain how these relate to seasonal influenza epidemics.

Seasonal influenza is a respiratory illness that is caused by the influenza A or B virus. Seasonal influenza epidemics occur every year during the fall/winter seasons, i.e., at different times in the Northern and Southern hemisphere [Nelson et al., 2007]. Such epidemics are responsible for the deaths of half a million people each year, and cause a significant economic burden [Stöhr, 2002]. To reduce this burden, prior to each season, a vaccine is developed against a set of influenza virus strains, i.e., a tetravalent or quadrivalent vaccine [Belshe, 2010]. This set of influenza strains is updated every year, by a global program

that is operated by the World Health Organization, which generates and interprets surveillance data to predict which virus strains will be most likely to appear in the next influenza season [Klingen et al., 2018].

Contrary to seasonal influenza, influenza pandemics occur less frequently. The influenza virus type A has the potential to cause pandemics, due to its high genetic variability. When discussing genetic variation in the context of the type A influenza virus, two gene segments are most important, i.e., hemagglutinin, a surface antigen which enables the virus to enter cells, and neuraminidase, a surface antigen which cleaves the newly formed virus [Nicholls, 2006]. As these segments are targeted by the host's immune system, they are susceptible to selective pressure, which is referred to as genetic drift [Lewis et al., 2016]. Next to genetic drift, antigenic shift, the process of reassortment of one of the hemagglutinin and one of the neuraminidase subtypes, can induce a significant change in the antigenic properties of the virus [Treanor, 2004]. A strain of the influenza A virus is typically referred to by the combination of these subtypes of the hemagglutinin (H) and neuraminidase (N) regions, e.g., H1N1, H2N2, H5N1, H3N2. This genetic diversity can result in viruses to which no (or little) immunity in the human population exists, and thus have the potential to cause a large pandemic [Nicholls, 2006]. Therefore, the outcome of an influenza pandemic can be much more severe than that of seasonal epidemics, with the potential to kill millions of people worldwide [Paules and Subbarao, 2017].

In the past century, there were three influenza pandemics. The Spanish influenza of 1918 is assumed to have infected close to a third of the world's population [Spinney, 2017], and was caused by an H1N1 influenza virus. More recently, the Asian influenza pandemic of 1957, caused by an H2N2 virus, killed 2 million people [Viboud et al., 2016]. And finally, in 1958, an H3N2 pandemic was responsible for one million deaths [Nicholls, 2006]. The first pandemic emergency in the twenty-first century was the influenza pandemic of 2009, caused by an H1N1 influenza virus [Fraser et al., 2009]. This pandemic was estimated to have killed close to three hundred thousand people [Dawood et al., 2012].

As we cannot predict the virus strain that will be responsible for the next pandemic, we can only start producing an effective vaccine when the virus strain has been identified. Therefore, vaccines will be available only in limited supply at the beginning of the pandemic [WHO, 2004]. This lack of vaccine availability, renders the containment of influenza pandemics challenging. We will further discuss these challenges throughout this chapter.

3 Studying epidemics

As indicated in the previous section, understanding the processes that underlie the epidemic spread of pathogens, is key to their control. This is a highly multi-disciplinary endeavour, where the interplay between different complex systems is investigated [Bedford et al., 2019].

We now present a broad overview of the different scientific disciplines that are involved in this process, to highlight the multi-disciplinary nature of this research. In this section we focus on epidemics where humans are central, although many of these considerations can be extrapolated towards epidemics that occur in livestock [Keeling et al., 2003].

The study of the pathogen and how it infects an individual human host and causes disease therein, is conducted by microbiologists. Insights in the pathogen enable pharmacists to develop new drugs or vaccines that can be used to implement therapeutic preventive actions.

In order to understand how pathogens move between humans, it is important to study the social contact network that connects the different individuals. The properties of such networks are examined in sociology, physics and economics. Often, a population of individuals is connected through a set of overlaying networks. E.g., on the one hand, adults are connected through a commuting network in the daytime, on the other hand, children meet their peers in school. These sub-populations are thus connected by different social contact networks, yet in the evening, when families meet at home, these contact networks are overlayed [Pilosof et al., 2017] (see Figure 1.2).

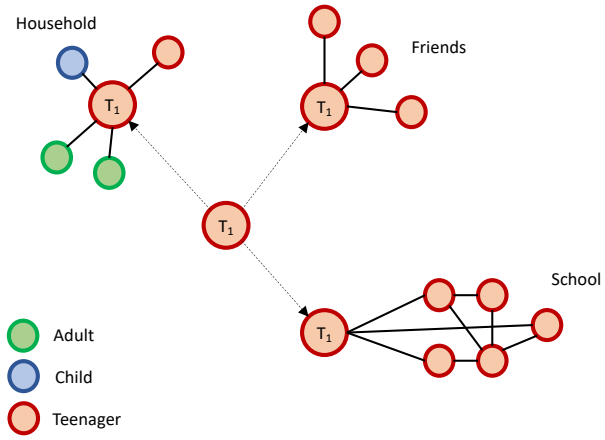


Figure 1.2: Different social networks from the perspective of a teenager T_1: household network (i.e., parents and siblings), friends network and school network. This figure is inspired by the work of Glass et al. [2006].

Furthermore, such social contact networks are often dynamic in nature. E.g., in a sexual contact network, for certain individuals, the partnerships in the network will change regularly. Moreover, the structure of such networks depends on the social context. E.g., while

friendship networks within a school can be captured using an exponential random graph model [Potter et al., 2012], sexual contact networks are scale-free [Liljeros et al., 2001].

For many pathogens, there is another animal species involved in the transmission process. For such pathogens, we make the distinction between zoonotic and vector-borne diseases. A disease is zoonotic, when the pathogen resides in another animal, through which it is able to infect humans. Such infections can lead to epidemics within the human population, if the virus is fit to be transmitted along human contact networks, as is for example the case for the Ebola virus [Pigott et al., 2014]. As another example, it has been shown for several influenza pandemics, that an avian influenza virus, or a combination between an avian and a human influenza virus, was at the source of the epidemic [Belshe, 2005]. Also the new SARS-CoV-2 virus, that is causing the ongoing COVID-19 pandemic, has a zoonotic origin Andersen et al. [2020]. An infectious disease is vector-borne if there is an animal (i.e., vector) that carries the pathogen from one human host to another. This vector is typically unaffected by the pathogen. A major group of pathogen vectors are arthropods, i.a., ticks, mosquitos, fleas, mites. Vector species impose an additional transmission network that complements human mobility, e.g., mosquitos can travel between houses.

This signifies the importance to study the ecology of animals responsible for zoonotic or vector transmissions, in order to get a full understanding of how these intermediate species affect the spread of pathogens in human populations. This includes studying how animal species, responsible for zoonotic transmission, come into contact with humans [George et al., 2011] and investigating how arthropod populations can be controlled [Moreira et al., 2009].

The ecology of animal species in general, and vector species in particular, is also affected by climatological changes. For example, the mosquito species *Aedes Albopictus*[4] was introduced in Southern Europe over the last decades [Akiner et al., 2016; Aranda et al., 2018], and is expected to settle in Northern Europe in the next decades due to climatological changes [Caminade et al., 2012; Fischer et al., 2014].

To understand how an epidemic unravels and to study its trajectory afterwards, it is important to chart the epidemic progress in great detail, and analyse the observed time series in a statistically sound framework.

The most fundamental statistic to record is incidence, i.e., the number of new cases per unit time, over time. Medical doctors will be able to report such cases when a patient is diagnosed[5], which implies that only (sufficiently) symptomatic cases will be considered. While the symptomatic incidence is a key statistic to monitor and study epidemics, for many

[4] *Aedes Albopictus* is a vector that is capable of transmitting different viruses, i.a., Zika virus, Dengue virus, Chikungunya virus.

[5] As many infectious diseases share symptoms, diagnostic laboratory kits are necessary to make unambiguous diagnoses.

pathogens (i.a., influenza [Leung et al., 2015], Dengue virus [Duong et al., 2015], Zika virus [Aubry et al., 2017]), there is a significant proportion of the infected population that will not develop any symptoms, but is still able to generate new infections. In order to investigate the proportion of asymptomatic infections in a population, the seroprevalence (i.e., the number of individuals in a population that test positive for an infectious disease based on a blood serum test) in a population can be assessed [Staras et al., 2006]. Furthermore, for pathogens with a prolonged incubation period (e.g., patients infected with HIV or Hepatitis C virus will only develop symptoms after months or years) a proactive diagnosis policy will be necessary [Boyer and Marcellin, 2000; Nakagawa et al., 2012].

When a diagnosis is confirmed, the genetic code of the pathogen that causes the infection can be isolated. This is especially interesting for fast-evolving organisms (e.g., RNA viruses, such as influenza, Dengue virus and HIV), from which the evolutionary relationship between the different samples obtained in an epidemic can be inferred [Kühnert et al., 2011].

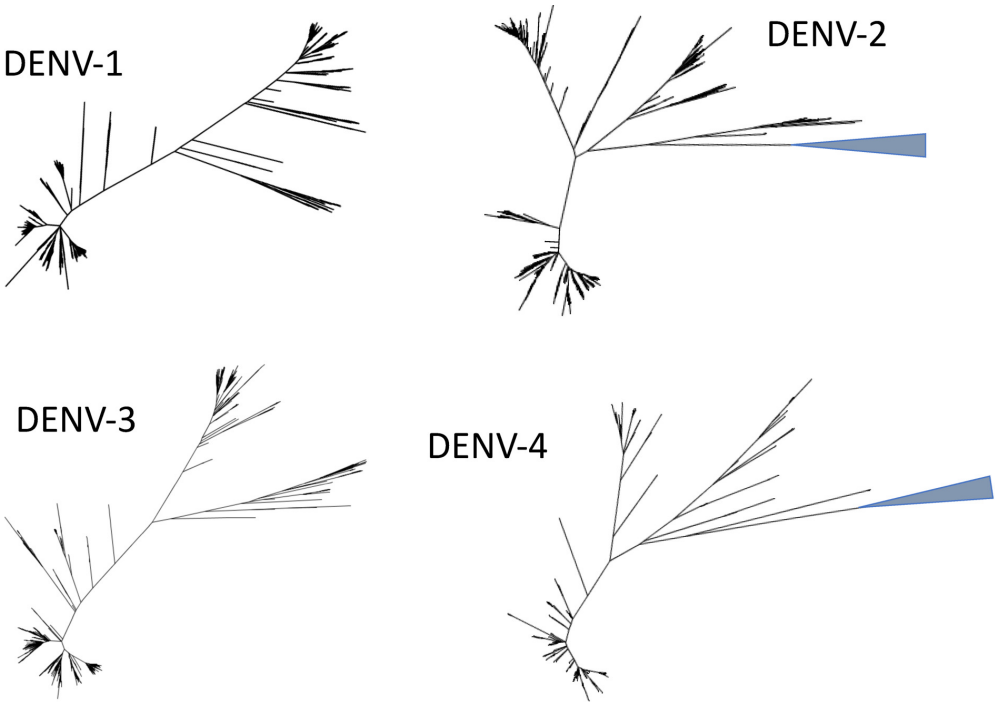


Figure 1.3: Phylogenetic trees of the four Dengue serotypes [Cuypers et al., 2018]. This figure demonstrates the vast genetic diversity within each of the Dengue serotypes.

These evolutionary relationships can be used to study the genetic diversity of the pathogen, which enables epidemiologists to classify the virus into genotypes (see Figure 1.3), which is essential to study pathogens in their broader context [Geretti, 2006; Hemelaar et al., 2006]. Furthermore, through phylodynamical analysis, these relationships can also be used to make inferences about epidemiological trends [Drummond et al., 2005] and the geographical origin and distributions of different strains of a pathogen [Holmes, 2004; Lemey et al., 2009].

In addition to charting ongoing or past epidemics, it is important to create models of infectious diseases that capture the right level of complexity to investigate the research question at hand. This field of research is grounded in mathematics, statistics and computer science. Such models range from elementary mathematical models (i.e., compartment models), over which formal reasoning is possible, to detailed mechanistic models (i.e., individual based models), that enable fine-grained models at the level of each individual. In this dissertation, we will use epidemiological models to investigate the efficiency of prevention strategies.

Aforesaid models will expose a set of parameters with some meaningful epidemiological interpretation. For example, the most elementary of models, i.e., the SIR model (see Chapter 3 for more details), will just consider two parameters that concern the rate of transmission and the rate of recovery. By fitting such models to data that was recorded during an epidemic, we can infer the value (or distribution) of such parameters, providing us with insight in the epidemiological process.

The parameters exposed in these models can also be used to explore potential scenarios, e.g., what the impact would be of an influenza virus that exhibits a particularly high infectiousness. By including parameters in the model that signify the impact of certain preventive measures, we can investigate the effect of interventions. The main objective of this dissertation is the use of epidemiological models to investigate and optimize prevention strategies. We will discuss this topic in the next section.

4 Studying the effect of interventions

In this dissertation, the overall goal is to develop methods that can be used to learn optimal intervention strategies, to reduce the burden caused by epidemics. We will now discuss different elements that are central to the study of interventions. First, we consider the exact objectives that we wish to optimize. Next, we enumerate different therapeutic and non-therapeutic intervention options. And finally, we discuss the optimal use of public health interventions.

4.1 Objective of the intervention

When discussing epidemic mitigation strategies, our goal is to minimize the pathogen's burden on a population. However, this objective can be interpreted in different ways.

The most intuitive objective is to minimize the cumulative incidence:

$$I_{\text{cum}} = \sum_{t=1}^{T} I_t, \tag{1.1}$$

where I_t signifies the number of newly infected individuals at time t, and T denotes the end of the epidemic. This objective directly reflects the impact of a intervention on the epidemiological process.

Another common objective is to minimize the mortality count over the course of the epidemic. Given the total incidence I_{cum} at time T, mortality can be formulated as a proportion f (i.e., fatality) of this quantity:

$$m = f \cdot I_{\text{cum}} \tag{1.2}$$

However, the probability that an infection is fatal is rarely uniformly distributed across the population. For many pathogens, the elderly will be more vulnerable to die upon infection, which is for example the case for seasonal influenza [Thompson et al., 2003]. We thus need to take into account the age-dependent cumulative infection count:

$$I_{\text{cum}}(a) = \sum_{t=1}^{T} I_t(a), \tag{1.3}$$

where $I_{\text{cum}}(a)$ and $I_t(a)$ are functions in terms of age a. Mortality can be defined in terms of survival probability function $s(a)$ that expresses the probability to die when infected:

$$m = \int_0^{100} (1 - s(a)) \cdot I_{\text{cum}}(a) da, \tag{1.4}$$

when we consider ages between $0 - 100$ years.

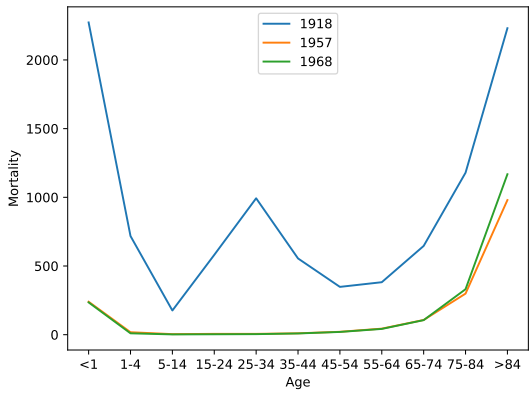


Figure 1.4: Mortality curves for three pandemics (i.e., 1918, 1957, 1968) [Luk et al., 2001]. The y-axis shows the number of deaths per 100000 individuals.

In Figure 1.4, we show the mortality function for three influenza pandemics (i.e., the 1918, 1957 and 1968 pandemic), to demonstrate the difference between mortality trends between age groups.

Other interesting objectives include the quality-adjusted life-year (QALY) and the economic cost induced by the epidemic. QALY is a measure of health, where health is modelled as a function that takes both the quantity and quality of the life saved into account [Sassi, 2006]. The economic cost of an epidemic attempts to quantify the impact of illness (e.g., sick people cannot work), death (i.e., economically active people that die leave an economic gap) and the cost of the intervention [Meltzer et al., 1999].

Furthermore, there are scenarios where multiple objectives can be of interest, and we might want to investigate a combination of objectives. This complicates the analysis as this requires us to specify how the different objectives need to be weighed.

In this dissertation, we consider the optimization of cumulative incidence. We do however discuss how our methods can be extended towards age-dependent and multi-objective settings.

4.2 Therapeutic and non-therapeutic interventions

Depending on the pathogen and the stratum of the population we are targeting, different intervention options are available. These can be divided in two groups: therapeutic and non-therapeutic interventions.

Therapeutic interventions consist of pharmaceutical measures that are provided to uninfected individuals to avoid infection or to treat infected individuals to avoid severe illness and death. Furthermore, by treating infected individuals, we also reduce the chance that these individuals will generate new infections. Some examples to avoid infection in healthy individuals include: vaccination to avoid measles infection [Wolfson et al., 2007], pre-exposure prophylaxis to avoid HIV infection [Karim and Karim, 2011] and post-exposure prophylaxis to avoid infection with bacterial pathogens such as anthrax [Wein et al., 2003]. Some examples to reduce the number of newly generated infections include: lifelong anti-retroviral therapy for HIV infected individuals to reduce their infectiousness [Lima et al., 2008] and using antiviral drugs to cure Hepatitis C infected patients [Hofmann and Zeuzem, 2011].

Non-therapeutic interventions consist of non-pharmaceutical measures to reduce the number of new infections, which includes behavioural and restrictive measures. An example of a behavioural measure is to promote condom-use to reduce the spread of sexually transmitted infections [Schick et al., 2010]. Restrictive measures can be targeted towards infected individuals (e.g., restricting traditional burying of individuals that died of an Ebola infection [Pandey et al., 2014]) or to reduce the mixing in a population subgroup (e.g., school closures [Cauchemez et al., 2008]).

4.3 Optimal use of resources

Given the availability of an efficient therapeutic intervention option (e.g., vaccines) infectious diseases can be controlled effectively [Piot et al., 2019]. A prime example of this is smallpox that was declared to be globally eradicated by the World Health Organization in 1980, through the use of an effective vaccine [Breman and Arita, 1980]. Many other infectious diseases are successfully contained through the use of vaccines (i.a., tetanus [Rappuoli et al., 2014], yellow fever [Barrett, 2017] and measles [WHO et al., 2017a]). The World Health Organization estimates that between 2010 and 2015, 10 million deaths were prevented due to vaccines, and that many more individuals were guarded from disease and disability because of vaccines [WHO et al., 2017b].

However, there is a substantial amount of pathogens for which no vaccine is readily available. For example, in the case of pandemic influenza, the virus needs to be isolated before the production of a vaccine can start (see Section 2). For other viruses, such as HIV, the development of a vaccine is challenging due to the extensive genetic variability of the virus [Johnston and Fauci, 2008]. Furthermore, the development of vaccines takes time and potential vaccines need to go through extensive clinical trials in order to be used [Abbink et al., 2018]. Therefore, in case of an emerging epidemic, no vaccine will be readily available.

In the absence of an efficient vaccine, other resources need to be allotted to contain an emerging epidemic, such as non-therapeutic measures (e.g., school closures to reduce social mixing) or therapeutic measures other than vaccines (e.g., antiviral drugs as a pre-prophylactic). Such measures have severe limitations. For one, schools can not be closed indefinitely. Also, antiviral influenza medication is both expensive and should be allocated with providence, as it might cause resistance in the virus, rendering it insusceptible to the antiviral compound [Lipsitch et al., 2007]. Therefore, it is important that such scarce resources will be optimally used.

When a vaccine does become available while an epidemic is ongoing, limitations with respect to vaccine production or logistic constraints will likely induce challenges regarding its distribution [Ulmer et al., 2006]. Moreover, other concerns might hamper an efficient implementation. For one, in the 2019 Ebola epidemic, researchers had to test vaccines in a war zone [Maxmen, 2019]. Furthermore, the spread of misinformation with respect to vaccines has induced vaccine hesitancy amongst a significant part of the population [MacDonald et al., 2015]. This is one of the main reasons for the major drop in vaccine coverage that has been reported in many countries, what has been an important driver for several measles outbreaks throughout the Western world [Robert et al., 2019].

Given the various constraints and complications that we discussed, it is crucial to optimize intervention strategies in order to effectively and efficiently contain or prevent epidemics. Furthermore, intervention strategies that efficiently use the available resources, make budgets available that can be used for a wider implementation of the measures, thus increasing their impact.

In this work, we investigate the use of reinforcement learning methods to identify optimal prevention strategies, in an epidemiological model, to support policy makers with their decision making. In the next section we introduce reinforcement learning, and in Section 6, where we introduce our research objectives in detail, we motivate why reinforcement learning is a suitable methodology to approach this problem.

5 Reinforcement learning

Artificial intelligence involves the study of agents (i.e., an entity that perceives and acts [Russell and Norvig, 2016]) that exhibit machine intelligence. Machine learning is a subfield in artificial intelligence, which concerns algorithms that do not explicitly encode the rules to follow in order to solve a particular problem, but rather operate on data through pattern detection or inference. This data can either be available beforehand or become available by interacting with an environment. In the latter case, this is referred to as reinforcement learning [Sutton and Barto, 1998], where an agent learns to behave optimally in an environment. To this end, the agent interacts with the environment (e.g., a video

game), by performing actions (e.g., controlling a joystick). Each action has an effect on the environment's state, which the agent observes (e.g., the screen of the video game), together with the reward for executing that particular action (e.g., the immediate score received). By carefully observing how the state evolves and considering the rewards that follow upon executing actions, reinforcement learning algorithms will aim to optimize the long term reward (e.g., to win the video game), thereby learning a policy that behaves optimally in said environment.

Although in the previous description we assume that reinforcement learning agents observe a state upon executing an action, there are reinforcement learning settings where no state space is considered, i.e., multi-armed bandits[6] [Audibert and Bubeck, 2010]. A multi-armed bandit models a set of actions (i.e., arms) that when played return a reward according to the action's reward probability distribution. This model is thus useful to formulate decision problems where there is a set of options to be considered. The multi-armed bandit formalism can be used to deal with decision problems with various objectives, of which regret minimization (i.e., to optimize the overall reward over time) [Sutton and Barto, 1998] and best arm identification [Bubeck et al., 2009], are the most well known.

6 Research objectives and contributions

The overall aim of our research is to investigate how decision making in an epidemiological context can be improved by following a reinforcement learning approach.

Firstly, we consider the decision problem where an optimal prevention strategy needs to be identified from a set of alternatives, where we assume that these prevention strategies can be evaluated in a stochastic individual-based epidemiological model. As such models are computationally intensive, it is paramount to identify the optimal strategy using a minimal number of model evaluations. Additionally, as computational resources are limited, these resources need to be reserved, and therefore epidemiological modelling experiments need to be carefully planned. Therefore, we assume that we need to operate using a fixed number of model evaluations (i.e., budget). To this end, we formulate this decision problem as a multi-armed bandit, where the different prevention strategy options are modelled as arms. Pulling an arm evaluates the corresponding prevention strategy in a stochastic epidemiological model, upon which the output of the model (e.g., attack rate) is returned as a reward. In this framework, we attempt to find the optimal prevention strategy by applying fixed-budget best-arm identification algorithms. We use epidemiological modelling theory to

[6]Note that there are also contextual multi-armed bandits (CMAB), where the agent observes a state prior to choosing an action [Féraud et al., 2016]. As in this dissertation, we will not consider the CMAB formalism, for simplicity, we consider multi-armed bandits to be stateless.

derive knowledge about the reward distribution, which we exploit using Bayesian best-arm identification algorithms. This enables us to boost the performance of the best-arm identification procedure. We evaluate these algorithms in a realistic experimental setting that concerns the vaccine allocation in an influenza pandemic. Through experiments, we demonstrate that it is possible to identify the optimal strategy using only a limited number of model evaluations. Finally, we show that the uncertainty distribution constructed by Bayesian best-arm identification algorithms can be used to inform decision makers about the confidence of an arm recommendation.

Secondly, we reflect on our efforts to use best-arm identification to select prevention strategies, and we identify two caveats. On the one hand, we argue that simply returning the best prevention strategy can be an impediment for public health scientists, as this limits flexibility with respect to decision making. On the other hand, we recognize that deciding the budget upfront can be challenging, which is especially true when computationally intensive models are used. For such models, it is difficult to make a trade-off between the available budget and desired confidence beforehand. To alleviate these concerns, we consider to formulate the decision task as an anytime m-top exploration problem, where the objective is to recommend the top m arms after every time step. The anytime m-top exploration setting was introduced by Jun and Nowak [2016] only recently and these authors introduced a new algorithm, i.e., AT-LUCB, to interact with this bandit setting. While AT-LUCB is an interesting algorithm that remains the state-of-the-art to date, it is an Upper Confidence Bound (UCB) variant, which makes it hard to incorporate prior information about the arms' reward distribution. This is unfortunate, as we show that such information can greatly improve learning performance, by using Bayesian bandit algorithms, such as Thompson sampling. Therefore, we investigate the potential of Thompson sampling for the m-top exploration problem, and propose the first Bayesian algorithm for this setting: Boundary Focused Thompson Sampling (BFTS). We demonstrate that BFTS outperforms AT-LUCB in the benchmarks introduced by Jun and Nowak [2016], and show that BFTS significantly outperforms AT-LUCB in the context of epidemic decision making, by introducing a new and challenging benchmark problem. We further establish BFTS's potential in a bandit setting with Poisson reward distributions, to show that BFTS is able to handle skewed and high-variance (i.e., challenging) reward distributions. Next, we perform a Bayesian analysis of BFTS that provides additional insight in BFTS' exploration strategy, and confirms that this strategy is well-grounded.

The first two objectives concern a reinforcement learning approach to select the best options out of a discrete set of policies. Next, we will investigate the use of reinforcement learning techniques to learn adaptive policies that encode which action is optimal given a particular epidemic state. We will investigate the use of reinforcement learning

in the context of pandemic influenza, where we aim to study and optimize school closure policies in Great Britain. Again, we learn in an epidemiological model. However, as the state-of-the-art in reinforcement learning algorithms, and especially these algorithms that can take into account a realistic state space, are quite sample inefficient, the use of aforementioned individual-based models is unattainable. Moreover, in order to learn fine-grained intervention policies, a careful balance between model granularity and the model's computational efficiency is required.

The third objective is thus to construct a model that is sufficiently fine-grained to evaluate school closure policies, and is yet computationally efficient such that it can be used in combination with reinforcement learning algorithms. To this end, we construct a meta-population model that consists out of a set of interconnected patches. Each patch corresponds to an administrative region in Great Britain and is internally represented by a compartment model that divides the population in four different groups (i.e., susceptible, exposed, infected and recovered) and covers four different age groups (i.e., children, adolescents, adults and elderly). In order to realistically model the mixing between the different age groups, we use a contact matrix that encodes the contact frequency between each age category. By using contact matrices that differentiate between school term and school holidays we can model school closure events. To realistically model the population heterogeneity and mobility in Great Britain, we parametrize each patch in the meta-population model with the census data of the corresponding district and connect the model patches according to Great Britain's census mobility network. We evaluate the model in both a set of synthetic scenarios and by reproducing the 2009 pandemic in Great Britain.

Finally, we investigate school closure policies in this model using different reinforcement learning techniques. We start by examining individual districts, with different population compositions, for which we establish a ground truth. We then evaluate two deep reinforcement learning algorithms and study which hyper-parameters optimize their learning performance in our new epidemiological model. Based on this analysis, we evaluate and discuss the learned policies. Next, we explore how we can learn policies considering a group of districts, to investigate whether there is a collaborative advantage when implementing school closures policies. As we consider a large number of districts in the epidemiological model (i.e., 379 districts throughout England, Wales and Scotland), we need to partition the model into smaller parts, as the state-of-the-art of multi-agent reinforcement learning algorithms is only able to deal with a limited number of agents. To this end, we use the Leiden algorithm to detect communities in the census mobility network. In these groups of districts, we use state-of-the-art multi-agent reinforcement learning algorithms to examine whether there is an advantage in collaborating between districts.

This work is relevant to policy makers on two levels. Firstly, it contributes new insights with respect to school closure policies. Secondly, it proposes a new framework to study intervention strategies through the use of deep reinforcement learning.

2 | Multi-armed bandits and reinforcement learning

> Reinforcement learning is the best representative of the idea that an intelligent system must be able to learn on its own, without constant supervision.
>
> Richard Sutton

The field of machine learning is commonly categorized in three main areas: supervised learning, unsupervised learning and reinforcement learning. Both supervised and unsupervised learning algorithms start from a set of examples, i.e., instances, about the problem at stake. In the case of supervised learning, each instance is composed of an observation, and the outcome of that observation [Caruana and Niculescu-Mizil, 2006]. As an example, consider that we want to learn a model for a spam filter, then our instances would consist out of a set of e-mails, with for each e-mail an outcome (i.e., a boolean flag) that indicates whether the e-mail is spam or not. Supervised learning can be used to solve classification problems (when the outcome is categorical) and regression problems (when the outcome is quantitative). In the case of unsupervised learning [Hinton et al., 1999], observations are not required to be paired with an outcome, and the learning algorithm aims to discover patterns in the set of examples. Unsupervised learning can, for example, be used to partition a set of examples in clusters.

Reinforcement learning is quite distinct from the former areas, as the algorithms in this realm do not start from a dataset, but consider an agent that learns by interacting with

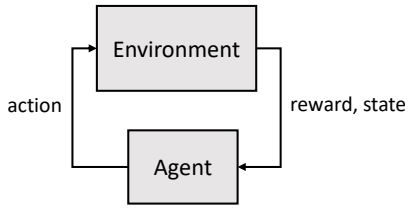


Figure 2.1: A reinforcement learning agent interacting with an environment. Upon executing an action, the agent receives feedback as a reward and can observe the state of the environment. (This figure takes inspiration from the book of Sutton and Barto [1998].)

an environment. More specifically, we consider an environment in which the agent can perform actions which may alter the environment's state. The agent can observe the state of the environment and receives a reward upon executing an action. This process is schematically visualized in Figure 2.1.

The central concept that underlies reinforcement learning, i.e., trial and error learning, finds its origin in psychology [Thorndike, 1911]. In the context of artificial intelligence, this learning scheme was first introduced by Minsky [1954] and popularized in the book of Sutton and Barto [1998]. Over the last years, a number of important milestones were achieved using the reinforcement learning framework, i.a., learning to play ATARI 2600 video games [Mnih et al., 2015], learning to play the board game Go to beat the number one human player [Silver et al., 2016] and learning to reach grandmaster performance in the real-time strategy game Starcraft II [Vinyals et al., 2019]. Each of these achievements consider a complex state space and we will further address this form of stateful reinforcement learning in Section 3. However, we will first consider reinforcement learning in a stateless setting, i.e., the multi-armed bandit as specified in the next section.

1 Multi-armed bandit

A multi-armed bandit (MAB) exposes a discrete set of actions (i.e., arms) that when executed (i.e., pulled) return a stochastic reward (see Definition 1) [Audibert and Bubeck, 2010].

Definition 1: Multi-armed bandit

The *multi-armed bandit* has K arms, where each arm a_k returns a reward r_k when it is pulled. r_k thus represents a sample from a_k's reward distribution. Each arm a_k has an expected reward, that is unknown, to which we refer as $\mu_k = \mathbb{E}[r_k]$.

The name multi-armed bandit stems from the slot machines that are found in casinos and which are also referred to as one-armed bandits. Following the notation used in Definition 1, a multi-armed bandit can thus be seen as a slot machine with K levers, where the player can at each time step choose which lever to interact with.

The multi-armed bandit is a formalism that can be used to model different problems, of which cumulative regret minimization and best-arm identification are the most well known. To introduce these settings, we will first define the bandit's best arm (see Definition 2), i.e., the arm with the highest expected reward.

Definition 2: Multi-armed bandit's best arm

For any *multi-armed bandit* (Definition 1), we have a best arm a_*, with corresponding expected reward:

$$\mu_* = \max_k \mu_k. \tag{2.1}$$

Cumulative regret is the regret we build up over time, by interacting with the multi-armed bandit. When we pull an arm $a^{(t)}$ at time t we suffer an instantaneous regret:

$$\mu_* - \mu_{a^{(t)}}. \tag{2.2}$$

Cumulative regret accumulates this instantaneous regret over time, as formally stated in Definition 3.

Definition 3: Cumulative regret

For any *multi-armed bandit* (Definition 1), with best arm μ^* (Definition 2), we have the cumulative regret over time T:

$$CR^{(T)} = \sum_{t=1}^{T} \mu_* - \mu_{a^{(t)}}. \tag{2.3}$$

Clearly, the best way to minimize cumulative regret is to always choose the best arm. However, as this best arm is a priori unknown, we will need to explore which arm is best. Although, as too much exploration is detrimental for our cumulative regret [Auer et al., 2002], we also need to exploit the arms which we believe to be good. Thus, in order to fulfil the goal to minimize the cumulative regret, the player needs to carefully balance between exploitation and exploration.

Orthogonal to the cumulative regret minimization problem, there is the best-arm identification problem, where the aim is to identify the best arm (see Definition 2). The best-arm

identification problem is an instance of the pure-exploration problem [Bubeck et al., 2009]. Best-arm identification can be approached from three angles: i.e., operate within a fixed budget [Audibert and Bubeck, 2010], return a decision when a given confidence is reached [Even-Dar et al., 2006], or to recommend the best arm after every time step (i.e., anytime recommendation) [Jamieson and Talwalkar, 2016]. In this setting, the objective is to minimize the simple regret (see Definition 4) [Bubeck et al., 2011].

Definition 4: Simple regret

Simple regret is the difference between the average reward of the best arm μ_* and the average reward $\mu_{J^{(T)}}$ of the recommended arm $J^{(T)}$ at time T:

$$R^{(T)} = \mu_* - \mu_{J^{(T)}} \tag{2.4}$$

In the next section, we will introduce two algorithms to address the cumulative regret minimization setting: the ϵ-greedy and Upper confidence bound algorithm. This will serve as a stepping stone to the Thompson sampling algorithm (Section 2.4). The Thompson sampling algorithm is an important prerequisite for Chapter 4 and Chapter 5, where we will use and construct Thompson sampling variants to address pure-exploration problems.

2 Cumulative regret

We will now describe two algorithms aimed at solving the cumulative regret problem: ϵ-greedy and Upper Confidence Bound (UCB). To provide some insight in how these algorithms work, we will run these algorithms on a simple 3-armed Bernoulli bandit[1], with expected rewards $\boldsymbol{\mu} = (.25, .5, .75)$.

2.1 ϵ-greedy

A straightforward way to address the cumulative regret bandit problem is to simply divide our budget of arm pulls between exploration and exploitation, in a fixed way. This is exactly what the ϵ-greedy algorithm does (see Algorithm 1). As ϵ will typically be small, most of the time (i.e., with probability $1 - \epsilon$) we will greedily select the arm that is currently thought to be best (i.e., exploit) and sometimes (i.e., with probability ϵ) we will choose an arm at random (i.e., explore). Note that the arms are ranked by their empirical mean $\hat{\mu}_k^{(t)}$. While this approach is not optimal, it is a intuitive algorithm that is often employed in practice due to its simplicity.

[1] A Bernoulli bandit is a multi-armed bandit (Definition 1), where each arm a_k, when played, returns a reward that is the result of a Bernoulli trial with mean μ_k.

Given: ϵ and a MAB with K arms
for $t = 1, \ldots, +\infty$ **do**
  $a_{max} = \arg\max_k \hat{\mu}_k^{(t)}$
  $\begin{cases} a^{(t)} = a_{max}, & \text{with probability } 1 - \epsilon \\ a^{(t)} = \text{random element of } \{1, ..., K\}, & \text{with probability } \epsilon \end{cases}$
  Play $a^{(t)}$ and observe its reward $r^{(t)}$
  Update $\hat{\mu}_{a^{(t)}}^{(t)}$ with $r^{(t)}$
end

Algorithm 1: ϵ-greedy

From the description of the algorithm, it is clear that we need to choose a hyper-parameter ϵ, that signifies the amount of exploration that will be necessary to solve the problem. This parameter is hard to choose beforehand as it depends on the hardness of the problem. In Figure 2.2, we demonstrate the working of this algorithm for $\epsilon = 0.01$ and $\epsilon = 0.1$. On the one hand, Figure 2.2 shows that when little exploration is used (i.e., $\epsilon = 0.01$) we are at risk of being stuck at a sub-optimal arm for a long time, thus cumulating regret quickly. On the other hand, when more exploration is used (i.e., $\epsilon = 0.1$) we will cumulate less regret at the start, but the exploration will continue even when the optimal arm has been identified, which will result in the cumulation of regret. This is shown in Figure 2.2, as the performance of ϵ-greedy with $\epsilon = 0.1$, with respect to cumulative regret, will eventually be surpassed by the algorithm with ϵ-greedy with $\epsilon = 0.01$.

This demonstration exposes two important limitations of plain ϵ-greedy. First, it keeps the amount of exploration constant over time. Secondly, ϵ-greedy's exploration scheme is inefficient, as it explores uniformly random, even when the empirical estimates of the means indicate large differences between the bandit's arms.

These limitations indicate that the amount of exploration will need to be decayed over time[2] and the uncertainty over the mean estimates should be taken into account to guide the exploration. To this end, we will now describe and demonstrate the Upper Confidence Bound (UCB) algorithm [Auer et al., 2002].

[2] In practice, ϵ-greedy variants that decay the ϵ over time are often used. However, it remains challenging to define the appropriate decay function beforehand, as this again depends on the hardness of the problem.

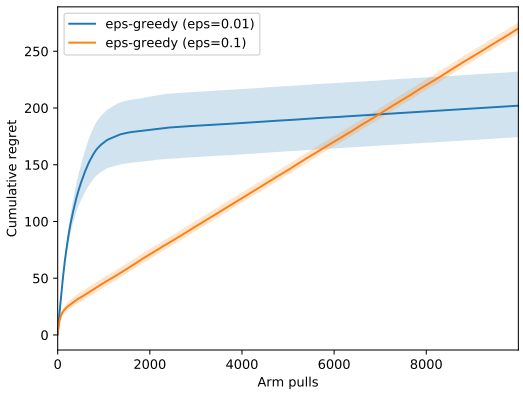


Figure 2.2: Results of running ϵ-greedy on the 3-armed Bernoulli bandit, for $\epsilon = 0.01$ and $\epsilon = 0.1$. We show the average cumulative regret with a 95% confidence interval for 100 runs.

2.2 Upper confidence bound

For each arm, UCB maintains a confidence bound that expresses the uncertainty of our estimate of the mean of this arm:

$$B_k^{(t)} = \hat{\mu}_k^{(t)} + \kappa \cdot \sqrt{\frac{\ln(t)}{n_k^{(t)}}}, \tag{2.5}$$

where $\hat{\mu}_k^{(t)}$ is the empirical mean of arm k, t is the total number of arm pulls, $n_k^{(t)}$ is the number of times arm k was pulled, and κ is a constant that modulates the amount of exploration. At each time step, the arm will be pulled that maximizes $B_k^{(t)}$, as formalized in Algorithm 2. Intuitively, this bound signifies that UCB will prefer promising arms that were played less, to balance between exploration and exploitation.

Importantly, we need to choose a hyper-parameter κ that modulates the amount of exploration. We demonstrate UCB on the three-armed Bernoulli bandit that we introduced earlier in this section, with two different values of κ (i.e., $\kappa = 1$ and $\kappa = 2$), in Figure 2.3. UCB immediately outperforms the ϵ-greedy algorithm that uses a low exploration hyper-parameter (i.e., $\epsilon = 0.01$), and quickly outperforms the ϵ-greedy algorithm that uses a higher exploration hyper-parameter (i.e., $\epsilon = 0.1$). For both exploration value $\kappa = 1$ and $\kappa = 2$, UCB demonstrates a cumulative regret curve that converges as the algorithm becomes more certain about the best arm, i.e., the behaviour we aim for.

Given: ϵ and a MAB with K arms
for $t = 1, \ldots, +\infty$ **do**
  $a^{(t)} = \arg\max_k \left[\hat{\mu}_k^{(t)} + \kappa \cdot \sqrt{\frac{\ln(t)}{n_k^{(t)}}} \right]$
  Play $a^{(t)}$ and observe its reward $r^{(t)}$
  Update $n_{a^{(t)}}$: $n_{a^{(t)}} = n_{a^{(t)}} + 1$
  Update $\hat{\mu}_{a^{(t)}}^{(t)}$ with $r^{(t)}$
end

Algorithm 2: Upper confidence bound

For this problem, the UCB exploration hyper-parameter that performs best is $\kappa = 1$. However, as this parameter depends on the hardness of the problem, it is difficult to come up with a good value upfront. Furthermore, per arm, UCB only uses the empirical mean $\hat{\mu}_k^{(t)}$ and the number of times the arm was pulled $n_k^{(t)}$, and thus disregards all other properties of the reward distribution. In many practical cases[3], prior knowledge about the reward distribution is available, even if it is only in the form of basic common knowledge or intuitions.

The Bayesian statistical framework provides a natural way to incorporate such prior knowledge, and in the context of multi-armed bandits, the Thompson sampling algorithm addresses the exploration-exploitation trade-of from a Bayesian perspective. Thompson sampling, also known as probability matching, has gained attention for its excellent performance [Scott, 2010; Chapelle and Li, 2011] and its applicability to a wide range of problems [Agrawal and Goyal, 2012; Osband et al., 2013; Russo and Van Roy, 2013; Kocák et al., 2014; Guha and Munagala, 2014]. As Thompson sampling operates by sampling from the Bayesian belief we have over the means, we first introduce the Bayesian foundations that underlie this algorithm.

2.3 Bayesian inference

To convey some intuition, we will rely on the example of tossing a coin to investigate the coin's fairness. When performing Bayesian inference, our goal is to estimate the posterior distribution of a particular hypothesis, where this hypothesis is expressed as a vector of parameters $\boldsymbol{\theta}$. To estimate the posterior, we use both the prior belief we have over the hypothesis and data points D that are collected through experimentation. Here, we want to investigate the fairness of a coin, so our hypothesis can be represented by a parameter θ that denotes the probability of tossing heads. The data D will be collected by tossing

[3]Including the work in this dissertation.

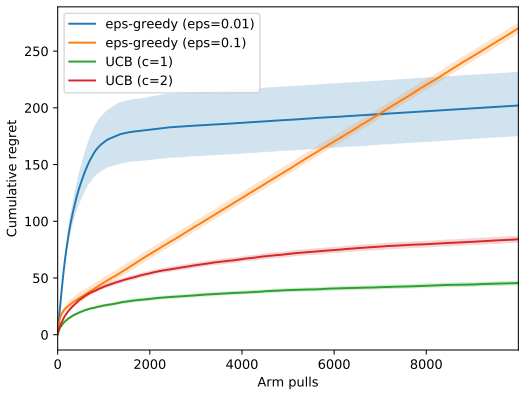


Figure 2.3: Results of running UCB ($\kappa = 1$ and $\kappa = 2$) on the 3-armed Bernoulli bandit, for reference we also show the ϵ-greedy results. We show the average cumulative regret with a 95% confidence interval for 100 runs.

the coin and observing the outcome, i.e., head or tail. Using this terminology, we express the posterior distribution (i.e., the probability of a hypothesis when we observe data) as:

$$P(\theta \mid D) \tag{2.6}$$

To compute this posterior we first introduce two concepts: i.e., the prior belief over the hypothesis,

$$P(\theta), \tag{2.7}$$

and the likelihood of observing the data given a hypothesis θ,

$$P(D \mid \theta). \tag{2.8}$$

Bayes' theorem shows that we can compute the posterior distribution as product of the prior and likelihood, normalized by the marginal likelihood:

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{\int P(D \mid \theta')P(\theta')d\theta'}. \tag{2.9}$$

As our coin fairness evaluation requires coin tosses (i.e., Bernoulli experiments), we will use a Bernoulli likelihood function. Consider an ordered list of data points $\mathbf{x}$, for each

x_i we compute the likelihood as:

$$P(x \mid \theta) = \begin{cases} \theta & \text{if } x \text{ is head} \\ 1 - \theta & \text{if } x \text{ is tail} \end{cases} \tag{2.10}$$

As we consider the coin tosses as independent, the likelihood over $\mathbf{x}$ consists out of the product of the likelihoods for each x_i:

$$\begin{aligned} P(\mathbf{x} \mid \theta) &= \prod_i P(x_i \mid \theta) \\ &= \theta^{h} \cdot (1 - \theta)^{t}, \end{aligned} \tag{2.11}$$

where h is the number of observed heads and t is the number of observed tails.

When the posterior distribution is in the same probability distribution family as the prior probability distribution, the prior and posterior are called conjugate distributions [Schlaifer and Raiffa, 1961]. For Bernoulli experiments, the usual conjugate prior is the **beta *distribution*** [Robert, 2007]:

$$P(\theta) = \mathcal{B}\text{eta}(\theta \mid \alpha_0, \beta_0), \tag{2.12}$$

with,

$$\mathcal{B}\text{eta}(\theta \mid \alpha_0, \beta_0) = \frac{\theta^{\alpha_0 - 1}(1 - \theta)^{\beta_0 - 1}}{B(\alpha_0, \beta_0)}, \tag{2.13}$$

where $B(.,.)$ is the **beta *function*** and α_0 and β_0 are hyper-parameters that allow us to specify prior knowledge. For now, we will use a uniform prior that corresponds to the beta distribution with $\alpha_0 = 1$ and $\beta_0 = 1$.

Conjugate priors are convenient, as they allow us to specify a closed-formed expression to update the prior when data is observed [Robert, 2007]. This avoids that we need to numerically approximate the posterior, e.g., using Markov Chain Monte Carlo [Hastings, 1970].

For our Bernoulli experiments, given the prior in Equation 2.12 and the likelihood in Equation 2.11, we can compute this posterior using Bayes' theorem:

$$\begin{aligned}
P(\theta \mid \mathbf{x}) &= \frac{P(\theta)P(\mathbf{x} \mid \theta)}{\int P(\mathbf{x} \mid \theta')P(\theta')d\theta'} \\
&= P(\theta)P(\mathbf{x} \mid \theta) \cdot \frac{1}{\int P(\mathbf{x} \mid \theta')P(\theta')d\theta'} \\
&= \mathcal{B}\text{eta}(\theta \mid \alpha_0, \beta_0) \cdot \theta^{h}(1-\theta)^{t} \cdot \frac{1}{\int_0^1 \mathcal{B}\text{eta}(\alpha_0, \beta_0) \cdot \theta'^{h}(1-\theta')^{t}d\theta'} \\
&= \frac{\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1}}{B(\alpha_0, \beta_0)} \cdot \theta^{h}(1-\theta)^{t} \cdot \frac{1}{\int_0^1 \frac{\theta'^{\alpha_0-1}(1-\theta')^{\beta_0-1}}{B(\alpha_0,\beta_0)} \cdot \theta'^{h}(1-\theta')^{t}d\theta'} \\
&= \frac{\theta^{\alpha_0-1}(1-\theta)^{\beta_0-1} \cdot \theta^{h}(1-\theta)^{t}}{\int_0^1 \theta'^{\alpha_0-1}(1-\theta')^{\beta_0-1} \cdot \theta'^{h}(1-\theta')^{t}d\theta'} \\
&= \frac{\theta^{\alpha_0-1+h}(1-\theta)^{\beta_0-1+t}}{\int_0^1 \theta'^{\alpha_0-1+h}(1-\theta')^{\beta_0-1+t}d\theta'} \\
&= \frac{\theta^{\alpha_0-1+h}(1-\theta)^{\beta_0-1+t}}{B(\alpha_0+h, \beta_0+t)} \\
&= \mathcal{B}\text{eta}(\theta \mid \alpha_0+h, \beta_0+t)
\end{aligned} \tag{2.14}$$

This derivation shows that we have a $\mathcal{B}$eta posterior that keeps track of heads and tails.

To demonstrate this posterior distribution, we will conduct two experiments, one with a fair coin (i.e., $\theta_f = .5$) and one with a biased coin (i.e., $\theta_b = .7$). We show the evolution of the posterior in Figure 2.4 for the fair coin, and in Figure 2.5 for the biased coin. For both experiments we start with a uniform $\mathcal{B}$eta prior. After observing the outcome of only a few coin tosses, we see that the posterior indicates either fairness (in Figure 2.4) or bias (in Figure 2.5), albeit with a significant amount of uncertainty. By adding more observations the uncertainty goes down, and after 500 observations the evidence for fairness or bias is quite strong, supported by the narrow posterior probability distribution.

2.4 Thompson sampling

Thompson sampling [Thompson, 1933] operates by maintaining a Bayesian belief over the means of the bandit's arms. We thus need to impose a prior belief over the expected rewards (i.e., a prior probability distribution), which we can update to a posterior belief upon observing rewards (i.e., a posterior probability distribution). At each time step, Thompson sampling is given a sample from the bandit posterior, thereby obtaining a

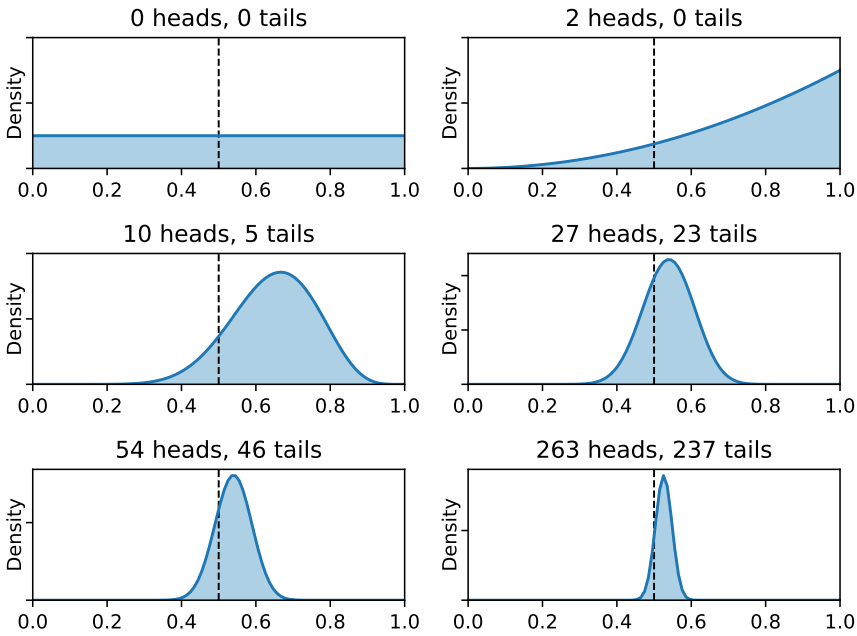


Figure 2.4: Posterior probability distributions for the experiment with a fair coin, where the x-axis represents θ. For each sub-figure, the header mentions the number of heads and tails that were observed. The vertical dotted line represents a fair coin, i.e., $\theta = 0.5$.

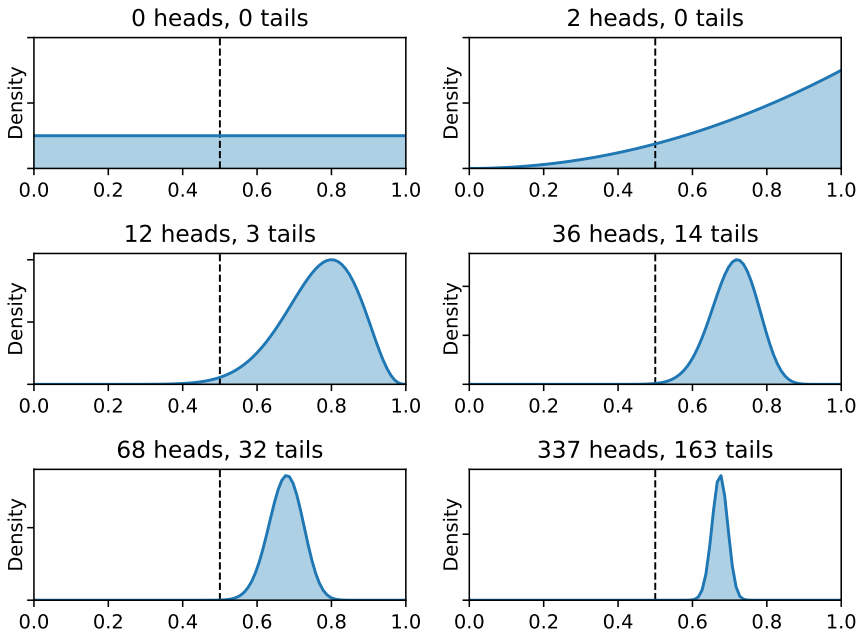


Figure 2.5: Posterior probability distributions for the experiment with a biased coin, where the x-axis represents θ. For each sub-figure, the header mentions the number of heads and tails that were observed. The vertical dotted line represents a fair coin, i.e., $\theta = 0.5$.

sample for each arm. This list of samples is then ranked, upon which the arm that is associated with the highest ranked posterior sample will be played and the observed reward will be used to update the bandit's posterior.

We will now formally present the Thompson sampling algorithm (Algorithm 3). To this end, we first define the posterior distribution over the means of the bandit (Definition 5).

Definition 5: Multi-armed bandit posterior

Consider a stochastic multi-armed bandit, for which our prior belief over the means is given by a distribution $\pi(.)$. Provided the observed history of arm pulls and rewards, until time $t-1$,

$$\mathcal{H}^{(t-1)} = \left\{a^{(i)}, r^{(i)}\right\}_{i=1}^{t-1}, \tag{2.15}$$

we have this posterior over the means of the bandit:

$$\pi(\cdot \mid \mathcal{H}^{(t-1)}), \tag{2.16}$$

given by $\pi(.)$ conditioned on the history of arm pulls and observed rewards $\mathcal{H}^{(t-1)}$.

At each time step t we sample an estimate $\tilde{\boldsymbol{\mu}}^{(t)}$ for the means $\mu_{1..K}$ from the posterior $\pi(\cdot \mid \mathcal{H}^{(t-1)})$. To order the elements that comprise $\tilde{\boldsymbol{\mu}}^{(t)}$, we use a ranking operator (Definition 6).

Definition 6: Thompson sampling ranking operator

For an estimate $\tilde{\boldsymbol{\mu}}^{(t)}$ of the means $\mu_{1..K}$, sampled from the posterior $\pi(\cdot \mid \mathcal{H}^{(t-1)})$ (Definition 5), we define the operator $\Psi_\rho(\tilde{\boldsymbol{\mu}}^{(t)})$ that denotes the ρ-ordered arm.

Using this ranking operator we can express the arm that is associated with the highest ranked posterior sample as:

$$\Psi_1(\tilde{\boldsymbol{\mu}}^{(t)}), \tag{2.17}$$

and complete the description of the algorithm.

Thompson sampling thus balances the exploration-exploitation problem by taking into account the uncertainty of the posterior. At the start, as little reward observations are available, a high level of uncertainty will result in a close to uniform exploration. The more certain we become about which arms are best, the more we will play them, and eventually we will focus almost exclusively on these arms. Note however, as long as we use

Given: An MAB with K arms, a prior $\pi(.)$ and history $\mathcal{H}^{(0)} = \emptyset$
for $t = 1, \ldots, +\infty$ **do**
  $\tilde{\boldsymbol{\mu}}^{(t)} \sim \pi(\cdot \mid \mathcal{H}^{(t-1)})$
  $a^{(t)} = \Psi_1(\tilde{\boldsymbol{\mu}}^{(t)})$
  $r^{(t)} \leftarrow$ Pull arm $a^{(t)}$
  $\mathcal{H}^{(t)} \leftarrow \mathcal{H}^{(t-1)} \cup \{a^{(t)}, r^{(t)}\}$
end

Algorithm 3: Thompson sampling

a posterior with infinite support, there will always be a low probability to play the arms that are believed to be sub-optimal, and we will never stop exploring entirely.

Thomson sampling will consider any available intuition that can be provided to quantify the uncertainty of the arm's means, ranging from an uninformative prior that only specifies the family of the reward distribution (see Chapter 4), to a prior that specifies the family of the reward distribution and its variance (see Chapter 5), to a prior that specifies dependencies between arms [Gopalan et al., 2014]. Such prior knowledge can result in a significant learning advantage.

From a Bayesian perspective, Thompson sampling thus is a remarkably intuitive algorithm, that balances the exploitation-exploration trade-off by sampling from its belief over the bandit. Furthermore, as this belief is formalized as a posterior distribution, we can reason over the uncertainty of the decision problem at hand. To demonstrate this process, we will use Thompson sampling to solve the Bernoulli bandit setting that we introduced at the start of this section, and investigate how the posterior evolves for one particular run.

To use Thompson sampling, we first need to choose an appropriate prior. In this example, we will use an uninformative prior, i.e., the Jeffreys prior, that is conjugate with respect to the Bernoulli likelihood. For the Bernoulli likelihood, the Jeffreys prior is a $\mathcal{B}$eta distribution with $\alpha_0 = \beta_0 = 0.5$ [Lunn et al., 2012].

From Figure 2.6, it is clear that Thompson sampling significantly outperforms both ϵ-greedy and UCB. To gain some more insight in Thompson sampling's operation, we will now examine how the posterior evolves for one particular run. To this end, we will show the bandit's posterior distribution at different time steps, as depicted in Figure 2.7, for one run.

We show the evolution of the posterior distributions in Figure 2.7. We start with the Jeffreys prior (Figure 2.7, panel a), and after 10 observations, we observe that the posterior associated with the best arm (green curve) tends towards the right i.e., a higher mean value (Figure 2.7, panel b). After 50 observations, the green posterior turns into a bell shaped distribution (Figure 2.7, panel c), which is further tightened (Figure 2.7, panel d and e), until a distribution shape that is close to a delta-peak is observed (Figure 2.7, panel f).

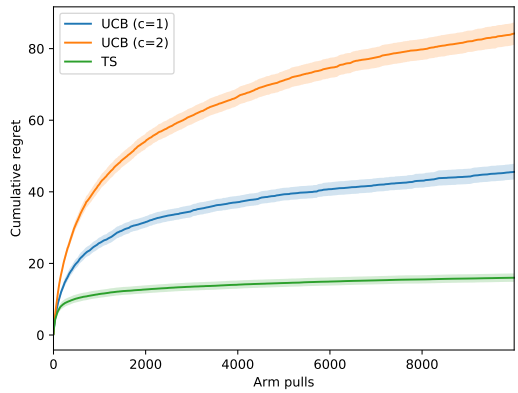


Figure 2.6: Results of running Thompson sampling on the 3-armed Bernoulli bandit, as reference, also the UCB results are included. We show the average cumulative regret with a 95% confidence interval for 100 runs.

In this dissertation we will not consider cumulative regret in a bandit setting, as we aim to solve pure exploration problems, i.e. best-arm identification (Chapter 4) and m-top exploration (Chapter 5). For these settings we will investigate different Thompson sampling variants that all operate by sampling from the posterior distribution, i.e., Top-two Thompson sampling for the best-arm identification setting and a new algorithm Boundary Focused Thompson sampling for the m-top exploration setting. For clarity and brevity, we will refer to the basic Thompson sampling algorithm that aims to optimize cumulative regret, as described in this section, as vanilla Thompson sampling.

3 The reinforcement learning problem

As stated earlier, reinforcement learning[4] concerns the process of an agent (i.e., an entity that perceives and acts [Russell and Norvig, 2016]) that learns to behave optimally by interacting with an environment. Each time the agent executes an action, it receives a reward from the environment. The environment, in which the agent acts, changes due to actions performed by the agent and external factors. While interacting with the

[4]For this introduction on the reinforcement learning problem, we took inspiration from the Sutton book [Sutton and Barto, 1998], a Phd thesis [Brys, 2016], and the video series on reinforcement learning by David Silver.

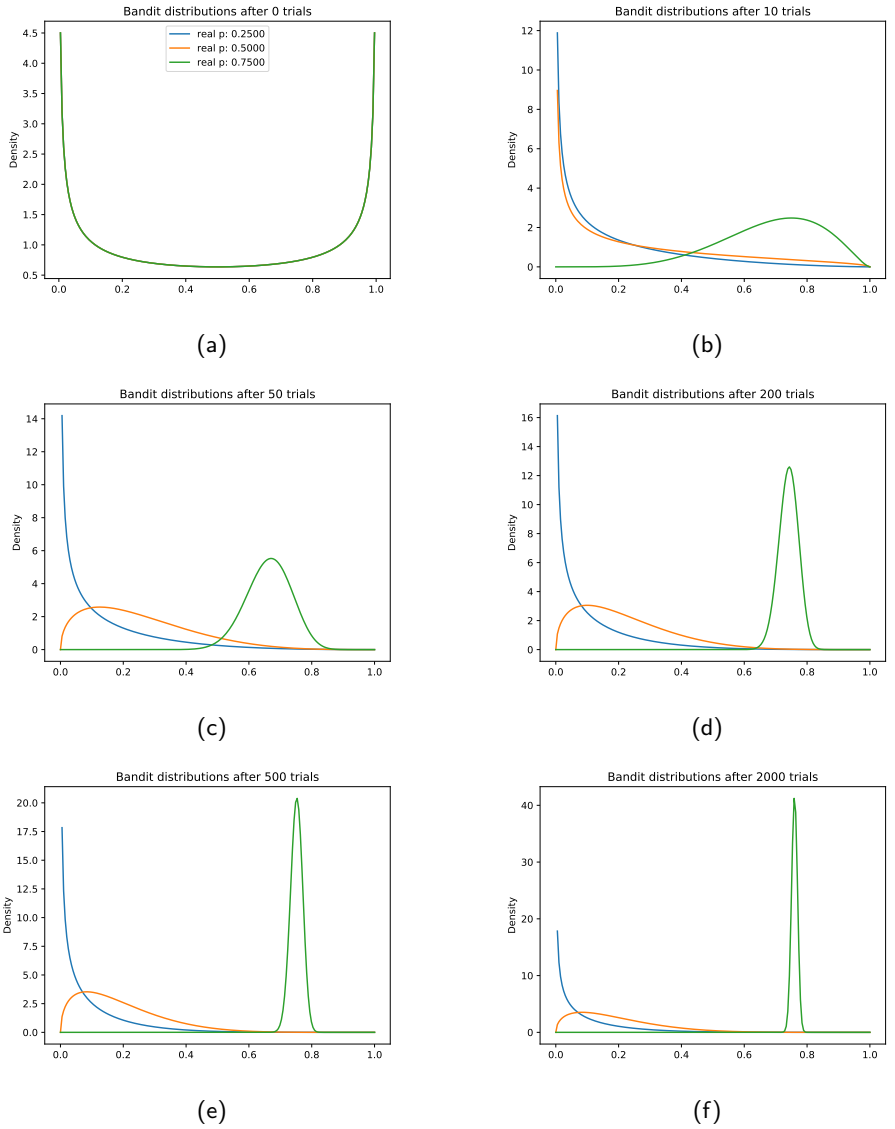


Figure 2.7: Posteriors (probability density function) of a Bernoulli bandit while running Thompson sampling. Each Figure represents the probability density function of the $\mathcal{B}$eta posterior after a specific number of trials, as specified in the Figure's title.

environment, the agent perceives the state of the environment, and needs to decide which action to choose, such that the agent maximizes the reward it accumulates. This process can be formalized as a Markov Decision Process (Definition 7).

Definition 7: Markov Decision Process

A Markov Decision Process corresponds to a tuple $\langle \mathcal{S}, \mathcal{A}, T, \gamma, R \rangle$, where

- $\mathcal{S}$ is the set of possible states the environment can take upon
- $\mathcal{A}$ is the set of possible actions an agent can take
- $T(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ signifies the transition probability to go from state $\mathbf{s}$ to state $\mathbf{s}'$ by taking an action $\mathbf{a}$
- γ is the discount factor that modulates the importance of future rewards
- $R(\mathbf{s}, \mathbf{a}, \mathbf{s}')$ is the reward function that specifies which reward the agents receives upon choosing an action $\mathbf{a}$ in state $\mathbf{s}$

Transitioning from one state $\mathbf{s}$ to another state $\mathbf{s}'$ is assumed to adhere to the Markov property, i.e., the transition only depends on the previous state $\mathbf{s}$. Note that $\mathcal{S}$ and $\mathcal{A}$ can be infinite sets.

An agent acts in a Markov Decision Process environment (Definition 7), by following a policy (Definition 8).

Definition 8: Policy

Given a Markov Decision Process $\langle \mathcal{S}, \mathcal{A}, T, \gamma, R \rangle$, an agent follows a policy p, that expresses the probability to take action $\mathbf{a}$ when in state $\mathbf{s}$:

$$p : \mathcal{S} \times \mathcal{A} \to [0, 1] \tag{2.18}$$

The goal is to learn a policy (Definition 8) that maximizes the return, i.e., the discounted sum of rewards (Definition 9).

Definition 9: Return

The return, or the discounted sum of rewards, starting from time step t, is defined as:

$$\sum_{i=0}^{\infty} \gamma^i R(\mathbf{s}_{t+i}, \mathbf{a}_{t+i}, \mathbf{s}_{t+i+1}), \tag{2.19}$$

where γ is the discount factor.

Based on the return (Definition 9), we define the value function (Definition 10), i.e., the value of being in state $\mathbf{s}$ when following policy p.

Definition 10: Value function

We define the value of being in state $\mathbf{s}$, following a policy p, as:

$$V^p(\mathbf{s}) = \mathop{\mathbb{E}}_{p,T}\left[\sum_{i=0}^{\infty} d^i R(\mathbf{s}_{t+i}, \mathbf{a}_{t+i}, \mathbf{s}_{t+i+1}) \mid \mathbf{s}_t = \mathbf{s}\right] \tag{2.20}$$

Analogously, we can define the state-action quality function (Definition 11), i.e., the quality of taking action $\mathbf{a}$ when in state $\mathbf{s}$ and subsequently following policy p.

Definition 11: State-action quality function (Q-function)

We define the value of being in state $\mathbf{s}$, when taking action $\mathbf{a}$, while following a policy p, as:

$$Q^p(\mathbf{s}, \mathbf{a}) = \mathop{\mathbb{E}}_{p,T}\left[\sum_{i=0}^{\infty} d^i R(\mathbf{s}_{t+i}, \mathbf{a}_{t+i}, \mathbf{s}_{t+i+1}) \mid \mathbf{s}_t = \mathbf{s}, \mathbf{a}_t = \mathbf{a}\right] \tag{2.21}$$

Based on Definition 11, we have the optimal policy:

$$p^*(\mathbf{s}, \mathbf{a}) = \arg\max_p Q^p(\mathbf{s}, \mathbf{a}) \tag{2.22}$$

When the MDP's transition and reward function are fully known, dynamic programming techniques can be used to solve a reinforcement learning problem [Sutton and Barto, 1998]. However, when this is not the case, reinforcement learning algorithms are necessary to solve this decision problem. In the next two sections, we consider two types of reinforcement learning algorithms: temporal difference learning (see Section 4) and policy gradient algorithms (see Section 7) .

4 Temporal difference learning

In temporal difference learning algorithms, an estimate of the quality function (Definition 11) is maintained, and this function is updated based on the observed rewards. A

well-known temporal difference algorithm is Q-learning [Watkins, 1989]. Q-learning keeps an estimate $\hat{Q}$ of the optimal Q-function Q^* (Definition 11), which is updated each time the agent interacts with the environment and observes a tuple $\langle \mathbf{s}, \mathbf{a}, \mathbf{r}, \mathbf{s}' \rangle$:

$$\hat{Q}(\mathbf{s}, \mathbf{a}) \leftarrow \hat{Q}(\mathbf{s}, \mathbf{a}) + \alpha \left[\mathbf{r} + \gamma \max_{\mathbf{a}'} \hat{Q}(\mathbf{s}', \mathbf{a}') - \hat{Q}(\mathbf{s}, \mathbf{a}) \right], \tag{2.23}$$

where $0 \leq \alpha \leq 1$ is the learning rate. In Algorithm 4, we show the complete Q-learning algorithm, as presented by Sutton and Barto [1998]. Watkins [1989] has shown that Q-learning converges to the optimal Q-values.

Given: a learning rate α
$\forall \mathbf{s} \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$: initialize $\hat{Q}(\mathbf{s}, \mathbf{a})$
for *each episode* e **do**
  Initialize $\mathbf{s}_0$
  for *each step in* e*:* $i = 0, 1, \ldots$ **do**
    Choose $\mathbf{a}_i$ from $\mathbf{s}_i$ using a policy derived from $\hat{Q}(\mathbf{s}, \mathbf{a})$
    Take action $\mathbf{a}_i$ and observe reward $\mathbf{r}_{i+1}$ and state $\mathbf{s}_{i+1}$
    $\hat{Q}(\mathbf{s}_i, \mathbf{a}_i) \leftarrow \hat{Q}(\mathbf{s}_i, \mathbf{a}_i) + \alpha \left[\mathbf{r}_{i+1} + \gamma \max_{\mathbf{a}'} \hat{Q}(\mathbf{s}_{i+1}, \mathbf{a}') - \hat{Q}(\mathbf{s}_i, \mathbf{a}_i) \right]$
    $\mathbf{s}_i \leftarrow \mathbf{s}_{i+1}$
  end
end

Algorithm 4: Q-learning [Sutton and Barto, 1998]

From an estimated Q-function $\hat{Q}$, we can derive a policy:

$$p(\mathbf{s}) = \arg\max_{\mathbf{a}} \hat{Q}(\mathbf{s}, \mathbf{a}). \tag{2.24}$$

When the $\hat{Q}$ estimate, used to construct this policy, converged to the optimal Q-function Q^* this policy will also be optimal.

Note that traditional Q-learning maintains a tabular estimate of the value function, i.e., a Q-value for each combination of state-action pairs. It is clear that this approach will not scale towards large or continuous state or action spaces. To make this feasible, we need to approximate the Q-function. In this dissertation, we use a variant of Q-learning called Deep Q-networks (see Section 6) that approximates the Q-function using an artificial neural network (see Section 5).

5 Artificial neural networks

An artificial neural network is a function approximator that consists out of a set of interconnected units (i.e., neurons), where each connection has a strength (i.e., weight). These units are part of different layers: an input and output layer that correspond to the interface of the function we approximate, and a set of hidden layers, that connect this input and output layer. To evaluate an artificial neural network, the values of each layer are propagated to the next layer. This propagation depends on the weights of the connections between the layers, and the activation function that is associated with each unit. We show an example of an artificial neural network in Figure 2.8 with an input layer, a hidden layer and an output layer. Both the input and output layer are fully connected[5] with the hidden layer.

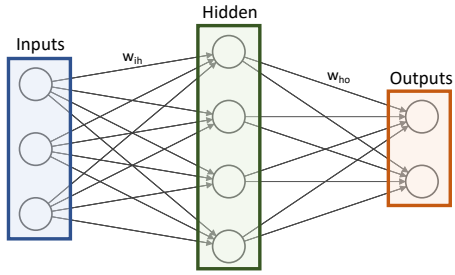


Figure 2.8: An artificial neural network with an input (blue), output (orange) and hidden (green) layer. The input and hidden layer are fully connected with connection strengths w_{ih} and the hidden and output layer are fully connected with connection strengths w_{ho}.

As the activation functions are typically chosen a priori, the weights comprise the model parameters in an artificial neural network, and can be fitted to data, through a method called back-propagation. Back-propagation uses a variant of gradient descent [Rumelhart et al., 1986], an optimization algorithm that can be used to find a local minimum of a function, by using the function's gradient with respect to its parameters [Bertsekas, 1997].

Formally [Mitchell et al., 1997], we can compute the output o_j computed by unit j,

$$o_j = \phi_j \left(\sum_i w_{ij} x_{ij} \right), \tag{2.25}$$

where w_{ji} is the weight associated with the i-th input to unit j, x_{ji} is i-th input to unit j, and ϕ_j is the activation function associated with unit j.

[5] Two layers l_1 and l_2 are fully connected when each unit in layer l_1 is connected to each unit in l_2.

Well known activation functions are the perceptron function [Rosenblatt, 1958] and the sigmoid function [Mitchell et al., 1997]. The perceptron function,

$$\phi(z) = \begin{cases} +1, & \text{if } z > 0 \\ -1, & \text{if } z \leq 0, \end{cases} \tag{2.26}$$

gives a binary output ($+1$ or -1), depending on the value of z, as shown in the top left panel of Figure 2.9. The perceptron function is not differentiable in the whole domain, which complicates its use in combination with gradient descent. To overcome this complication, the sigmoid function was introduced in the context of artificial neural networks [Rumelhart et al., 1986],

$$\phi(z) = \frac{1}{1 - \exp(-z)}. \tag{2.27}$$

The sigmoid function is a differentiable function that is monotonically increasing and constrained by a pair of horizontal asymptotes. We show the sigmoid activation function in the top right panel of Figure 2.9.

Another popular activation function is the hyperbolic tangent function,

$$\phi(z) = \frac{\exp(z) - \exp(-z)}{\exp(z) + \exp(-z)}, \tag{2.28}$$

where the output of this function is bounded by the same values (i.e., -1 and 1) as the perceptron function. We show the hyperbolic tangent activation function in the bottom left panel of Figure 2.9. Finally, we introduce the more recently proposed rectified linear activation function [Nair and Hinton, 2010],

$$\phi(z) = \begin{cases} z, & \text{if } z \geq 0 \\ 0, & \text{if } z < 0, \end{cases}, \tag{2.29}$$

which we show in the bottom right panel of Figure 2.9. Both the hyperbolic tangent activation function and rectified linear activation function are used in the context of deep artificial neural networks, a recent learning methodology that obtained remarkable results in a wide range of machine learning tasks [LeCun et al., 2015], including reinforcement learning [Arulkumaran et al., 2017].

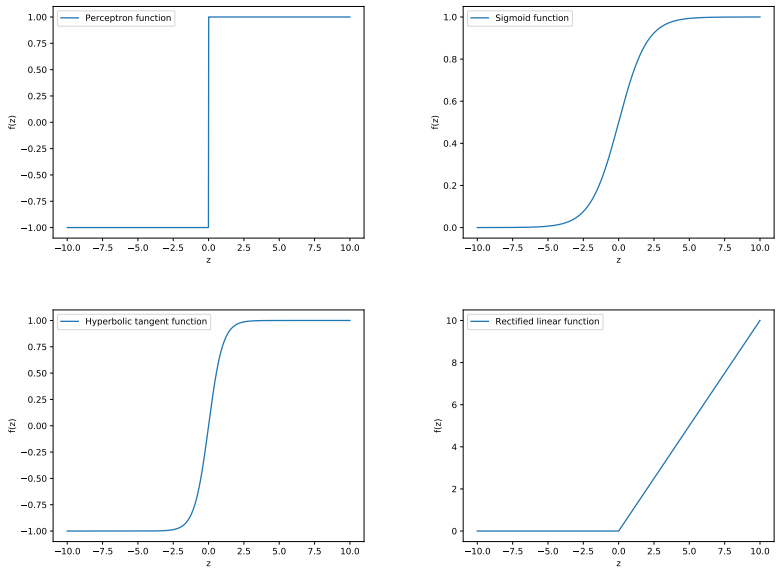


Figure 2.9: We show four activation functions: the perceptron function (top left), the sigmoid function (top right), the hyperbolic tangent function (bottom left), and the rectified linear activation function (bottom right).

6 Deep Q-networks

The deep Q-networks algorithm (DQN) is a Q-learning variant that approximates the Q-function[6], rather than representing the Q-values for each state-action pair explicitly [Mnih et al., 2015]. To do this, a parametrized representation of the Q-function,

$$Q\left(\mathbf{s}_t, \mathbf{a}_t \mid \boldsymbol{\theta}\right), \tag{2.30}$$

is maintained in the form of a neural network[7]. This neural network accepts the environment's state as an input an has an output neuron for each action, and this Q-network is thus analogous to the Q-table in tabular Q-learning. Similar to tabular Q-learning, we select actions by applying an operator on the set of the output neurons, e.g., a greedy or an ϵ-greedy selection operator. We show an example of a DQN neural network in Figure 2.10.

[6]For this introduction on deep Q-networks, we took inspiration from the DQN paper by DeepMind [Mnih et al., 2015] and the video by Olivier Sigaud.

[7]DQN typically uses a rectified linear activation function for the hidden units.

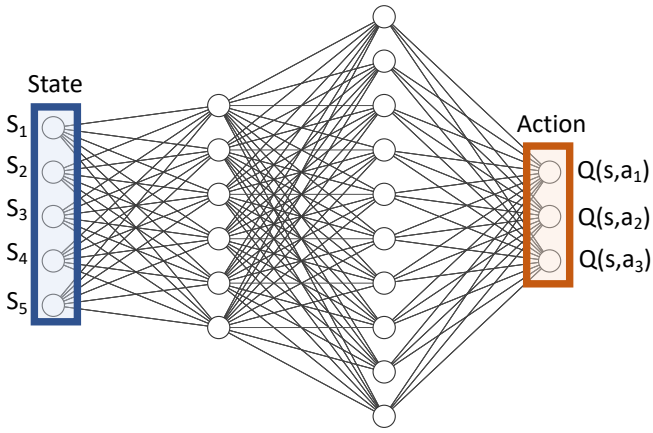


Figure 2.10: We show an example deep Q-network, where the input layer accepts the states of the environment (blue), and the output layer represents a Q-value for each of the different actions (orange).

To learn this Q-network, we need to minimize the temporal difference error, given a batch of samples, i.e., a mini-batch:

$$\{\mathbf{s}_i, \mathbf{a}_i, \mathbf{r}_i, \mathbf{s}_{i+1}\}_{i=0}^{N}. \tag{2.31}$$

From a supervised learning perspective, we could use the mean-squared error objective function:

$$L(\mathbf{s}, \mathbf{a}) = \frac{1}{N} \sum_i \left(y_i - Q(\mathbf{s}_i, \mathbf{a}_i \mid \boldsymbol{\theta})\right)^2, \tag{2.32}$$

where,

$$y_i = \mathbf{r}_i + \acute{d} \max_{\mathbf{a}} Q(\mathbf{s}_{i+1}, \mathbf{a} \mid \boldsymbol{\theta}). \tag{2.33}$$

However, optimizing this objective function is not stable, as the target y_i is also a function of Q, and having a moving target breaks supervised learning methods. To solve this problem, DQN maintains an additional target network Q', next to the Q-network. Equation 2.32 will be used to update the Q-network, however, the target y_i will be computed using the target network. As Q' is only updated periodically, the target values will be fixed most of the time, thereby stabilizing the learning procedure. Furthermore, supervised learning methods assume that samples will be identically and independently distributed, which will

not be the case for samples collected by a reinforcement learning agent, as these samples will be correlated. To this end, a replay buffer is used, in which the collected samples are stored, using a sliding window approach. From this buffer, DQN randomly samples mini-batches (Equation 2.31) to be used to train the network, such that the samples in the mini-batch will be independently distributed.

DQN has led to an important breakthrough, as it was used to learn to play simple video games (Atari 2600) from screen playback, thereby reaching human level performance [Mnih et al., 2015]. In this dissertation, we will evaluate the use of DQN, to search for optimal school closure policies, in Chapter 7.

7 Policy gradient

Policy gradient methods[8] parametrize a policy directly, instead of learning the Q-function, which is beneficial for decision problems with a state or action space that is large or continuous. We consider a stochastic parametrized policy $p_{\boldsymbol{\theta}}$, and we aim to optimize the expectation of the policy gradient,

$$\mathop{\mathbb{E}}_{p_{\boldsymbol{\theta}},T}\left[\sum_{t=0}^{\infty}\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\hat{A}_t\right], \tag{2.34}$$

where $\hat{A}_t$ is the advantage estimate:

$$\hat{A}_t = \left[\sum_{i=0}^{\infty} d^i R\left(\mathbf{s}_{t+i}, \mathbf{a}_{t+i}, \mathbf{s}_{t+i+1}\right)\right] - V^p\left(\mathbf{s}_t\right), \tag{2.35}$$

i.e., the difference between the return (Definition 9) and the value (Definition 10). Both $p_{\boldsymbol{\theta}}$ and $V^p\left(\mathbf{s}_t\right)$ are modelled using distinct artificial neural networks, and we can estimate the expectation in Equation 2.34, by collecting a set of trajectories Y and computing a sample mean $\hat{g}$ from it:

$$\hat{g} = \frac{1}{|Y|}\sum_{y\in Y}\sum_{t}\nabla_{\boldsymbol{\theta}}\log p_{\boldsymbol{\theta}}\left(\mathbf{a}_t \mid \mathbf{s}_t\right)\hat{A}_t \tag{2.36}$$

We present a sketch of the vanilla policy gradient algorithm, as presented by Williams [1992], in Algorithm 5.

[8] For this introduction on policy gradient reinforcement learning, we took inspiration from different sources: the Sutton book Sutton and Barto [1998], the papers by Williams [1992], Schulman et al. [2015], Schulman et al. [2017] and the video on Proximal Policy Optimization by *Arxiv Insights*.

Given: a learning rate $\alpha(.)$
Initialize policy parameter $\boldsymbol{\theta}$ and value function V
for $i = 1, 2 \ \ldots$ **do**
  Collect a set of trajectories Y using the current policy $p_{\boldsymbol{\theta}_i}$
  for *each time step* t *and each trajectory* $y \in Y$ **do**
    Compute the advantage $\hat{A}_t$ (Equation 2.35)
  end
  Re-fit $V^p(s_t)$
  Compute the policy gradient estimate $\hat{g}_i$ (Equation 2.36)
  Update the policy: $\boldsymbol{\theta}_{i+1} = \boldsymbol{\theta}_i + \alpha(i)\,\hat{g}_i$
end

Algorithm 5: Vanilla policy gradient [Williams, 1992]

While vanilla policy gradient is an elegant reinforcement learning technique, its use in real-world applications remains challenging, as these methods are quite sensitive to the choice of the learning rate. When the learning rate is too big, this may result in a significant performance drop, and when the learning rate is too small, the learning agent will progress slowly.

One successful approach to address this problem, is the trust region policy optimization (TRPO) algorithm [Schulman et al., 2015], where the idea is not to move too far from the old policy when updating the new policy. This is done by maximizing this objective function,

$$L^{\mathsf{TRPO}}(\boldsymbol{\theta}) = \mathbb{E}\left[\frac{p_{\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}_{\mathsf{old}}}}\hat{A}_t\right], \tag{2.37}$$

with a Kullback-Leibler (KL) divergence constraint [Kullback and Leibler, 1951],

$$D_{\mathsf{KL}}\left(p_{\boldsymbol{\theta}} \mid p_{\boldsymbol{\theta}_{\mathsf{old}}}\right) \leq \delta, \tag{2.38}$$

that avoids that the updated policy $p_{\boldsymbol{\theta}}$ will move too far from the current policy $p_{\boldsymbol{\theta}_{\mathsf{old}}}$.

This additional constraint complicates the optimization procedure, and Proximal Policy Optimization (PPO) attempts to approximate this hard constraint with a penalty in the objective function [Schulman et al., 2017]. The main idea is to clip the policy probability ratio:

$$r_t(\boldsymbol{\theta}) = \frac{p_{\boldsymbol{\theta}}}{p_{\boldsymbol{\theta}_{\mathsf{old}}}}, \tag{2.39}$$

resulting in an objective function:

$$L^{\mathsf{CLIP}}(\boldsymbol{\theta}) = \mathbb{E}\left[\min\left(r_t(\boldsymbol{\theta}), \mathsf{clip}\left(r_t(\boldsymbol{\theta}), 1-\epsilon, 1+\epsilon\right)\right)\hat{A}_t\right]. \tag{2.40}$$

PPO approximates the policy network and value network with two distinct neural networks.

In this dissertation, we will use the Proximal Policy Optimization (PPO) algorithm, to search for optimal school closure policies (Chapter 7).

3 | Epidemiological models

A smart model is a good model.

Tyra Banks

In order to study the impact of mitigation policies a pertinent epidemiological model structure needs to be defined, which depends on the pathogen, the social contact network, ecological aspects (e.g., vector ecology) and the character of the investigated mitigation policy. Furthermore, an important decision with respect to model structure is its granularity. Compartment models, on the one hand, divide the population into discrete homogeneous states (i.e., compartments) and describe the transition rates from one state to another [Diekmann et al., 2012]. Individual-based models, on the other hand, explicitly represent all individuals and their connections, and simulate the spread of a pathogen among this network [Willem et al., 2017]. In between these extremes in model space, all kinds of meta-population models can be constructed, as shown in Figure 3.1. These distinct model structures are important to address different aspects of public health inquiries. In this chapter, we enumerate three important model structures that we will use in this dissertation: compartment models, individual-based models and meta-population models.

1 Compartment models

In Section 1.1, we will present the SIR model, one of the most fundamental epidemiological models, that was introduced by Kermack and McKendrick [1927]. We will extend the

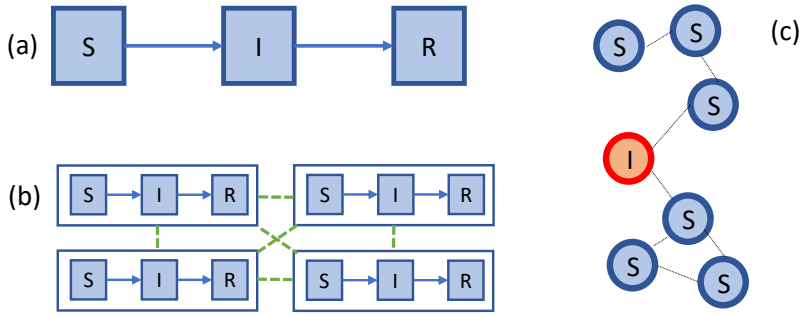


Figure 3.1: Two extremes in model space: (a) an SIR compartment model, (c) and individual-based model. We show a meta-population model (b), that is situated between the models in (a) and (c), in terms of complexity.

SIR model, firstly, by adding a compartment (Section 1.2), and secondly, by adding age heterogeneity (Section 1.3). We will start by describing the models in term of ordinary differential equations which implies a deterministic evaluation. In Section 1.4, we will discuss how these models can be evaluated stochastically.

1.1 SIR model

Compartment models partition the population into a finite set of states (i.e., compartments) between which communication is possible. For a pathogen, to which a patient can obtain immunity after being infected (e.g., pandemic influenza), we can partition the population in three groups: individuals that are susceptible to infection (i.e., susceptibles), individuals that are infected (i.e., infected), and individuals that recovered and obtained immunity (i.e., recovered) (see Figure 3.2). This model, referred to as the SIR model (i.e., abbreviation of Susceptibles-Infected-Recovered), was introduced by Kermack and McKendrick [1927]. Communication between the different compartments occurs when susceptibles become infected, as modulated by a transmission rate $\beta\chi$ (with β the probability of infection and χ the contact rate), and when infected individuals recover, as modulated by a recovery rate γ (see Figure 3.2).

The SIR model applies to epidemics that suddenly cause an outbreak and have a relatively short duration, such that the births and deaths of the hosts can be ignored. Examples of such epidemics are seasonal influenza (within one season), pandemic influenza and the Ebola virus.

The SIR model can be formalized as a system of ordinary differential equations, as shown in Definition 12.

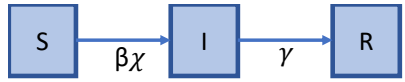


Figure 3.2: An SIR model with three compartments, i.e., susceptibles (S), infected (I) and recovered (R) between which communication is modulated with transmission rate $\beta\chi$ and a recovery rate γ.

Definition 12: SIR model

$$\begin{aligned}\dot{S}(t) &= -\beta\chi S(t)\frac{I(t)}{N(t)}\\ \dot{I}(t) &= \beta\chi S(t)\frac{I(t)}{N(t)} - \gamma I(t)\\ \dot{R}(t) &= \gamma I(t),\end{aligned} \tag{3.1}$$

where, β is the probability of infection when a contact takes place, χ is the contact rate and γ is the recovery rate, with,

$$S(0) > 0, I(0) > 0, R(0) = 0, \tag{3.2}$$

and the total population N is constant:

$$N(t) = S(t) + I(t) + R(t). \tag{3.3}$$

Definition 12 shows that each susceptible person comes into contact with χ people per day (i.e., the contact rate), of which a fraction $\frac{I(t)}{N(t)}$ is infectious. Each contact has a probability to transmit the infection β that depends on the pathogen and the route of transmission. Thus, $\beta\chi S\frac{I(t)}{N(t)}$ is the amount of susceptibles, per unit time, that move out of the susceptible into the infected compartment. After being infected for a certain period, individuals will recover at a rate γ. Thus, $\gamma I(t)$ is the amount of infected individuals, per unit time, that move out of the infected compartment into the recovered compartment.

Furthermore, Definition 12 reveals three main assumptions that underlie this model. Firstly, the population is closed, i.e., no births, deaths or migrations are modelled. Secondly, we assume that the entire population mixes homogeneously (e.g., spatial homogeneity and age homogeneity). Thirdly, during the initial phase of the epidemic, the number of infected individuals will grow exponentially.

A key epidemic parameter is the basic reproductive number (Definition 13), which we can intuitively derive from the SIR model.

Definition 13: Basic reproductive number

The basic reproductive number, R_0, is the number of infections that is, on average, generated by one single infected individual that is placed in an otherwise fully susceptible population.

The basic reproductive number is an important parameter, as it signifies the initial rate of spread of the epidemic [Keeling and Rohani, 2011]. This means that R_0 is, on average, the threshold for the epidemic to succeed ($R_0 > 1$) or die out ($R_0 \leq 1$). For the SIR model, we can derive R_0 by considering that there will be an epidemic if and only if,

$$\dot{I}(t) > 0. \tag{3.4}$$

When we substitute $\dot{I}(t)$ with its definition (Definition 12), we have:

$$\begin{aligned} 0 < \dot{I}(t) \\ &= \beta\chi S(t)\frac{I(t)}{N(t)} - \gamma I(t) \\ &\leq \beta\chi N(t)\frac{I(t)}{N(t)} - \gamma I(t) \\ &= I(t)(\beta\chi - \gamma). \end{aligned} \tag{3.5}$$

From Equation 3.5, we derive that the relative removal rate,

$$\frac{\gamma}{\beta\chi}, \tag{3.6}$$

should be less than 1 to allow for the disease to spread [Keeling and Rohani, 2011]. As R_0 is the inverse of the relative removal rate [Keeling and Rohani, 2011], we have:

$$R_0 = \frac{\beta\chi}{\gamma}. \tag{3.7}$$

To demonstrate the SIR model, we show an example with a transmission rate $\beta\chi = 0.2$ and recovery rate $\gamma = 0.1$ in Figure 3.3.

While the SIR model is a basic model that assumes homogeneity in the population, it can be easily be extended in various ways to accommodate more complex modelling inquiries. There are two common ways to do this. On the one hand (i), we can add compartments to include additional features or complexity. On the other hand (ii), we can repeat a basic model to represent heterogeneity in the population. We will now provide examples for both approaches.

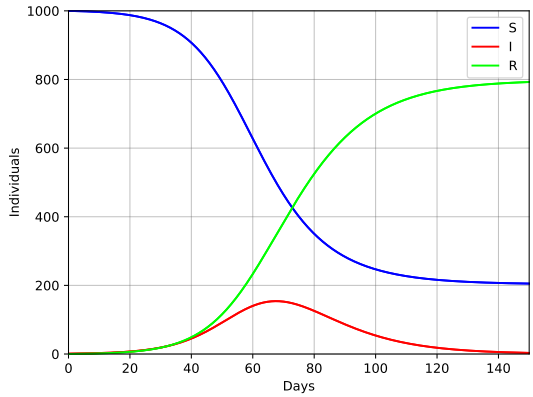


Figure 3.3: The output of a SIR model with transmission rate $\beta\chi = 0.2$, recovery rate $\gamma = 0.1$, population count $N = 1000$ and $I(0) = 1$, for each of the three state variables.

i For example, the SIR model makes the assumption that when a susceptible individual is infected, this individual becomes infectious instantaneously. However, for many pathogens, infected individuals experience a latent period prior to becoming infectious, e.g., pandemic influenza [Mills et al., 2004], Ebola virus [Lekone and Finkenstädt, 2006] and Dengue virus [Marini et al., 2019]. This can be mitigated by adding an exposed compartment in between the susceptible and infected compartment, which extend the SIR into a SEIR model (see Section 1.2). Other extensions are possible, such as adding an additional infected compartment to account for super-spreading events, e.g., to the 2013-2014 Ebola epidemic [Volz and Siveroni, 2018]. Another example, where the aim is to model an arbovirus epidemic, is to use the S(E)IR model, to model the epidemic in the human host, in conjunction with a SEI model, to model infection in the mosquito vector [Huber et al., 2018].

ii For example, to add age heterogeneity, we can consider a set of discrete age groups, and repeat the S(E)IR model for each of the age groups (see Section 1.3). A similar approach can also be used to incorporate heterogeneity with respect to transmission intensity, where a discrete set of transmission intensities is considered (e.g., hotspot vs non-hotspot), and the S(E)IR model is repeated for each of these traits [Azman and Lessler, 2015].

We defined the SIR model in terms of a system of ordinary differential equations (see Definition 12), which implies a deterministic evaluation of the system. However, for predictions, stochastic models are preferred, as they to account for stochastic variation and allow us to quantify uncertainty [King et al., 2015]. Furthermore, it is necessary to take the stochasticity of the epidemic process into account to evaluate preventive strategies [Germann et al., 2006]. To this end, in Section 1.4, we will discuss how the SIR model and other compartment models can be evaluated stochastically.

1.2 SEIR model

For many pathogens, infected individuals experience a latent phase prior to becoming infectious. To address this, the SEIR model adds an exposed (E) compartment to the SIR model and a new transition to move from the exposed to the infected compartment, which we will refer to as ζ, i.e., the latency rate.

The SEIR model can again be formalized as a system of ordinary differential equations, as shown in Definition 14.

Definition 14: SEIR model

$$\begin{aligned}
\dot{S}(t) &= -\beta\chi S(t)\frac{I(t)}{N(t)} \\
\dot{E}(t) &= \beta\chi S(t)\frac{I(t)}{N(t)} - \zeta E(t) \\
\dot{I}(t) &= \zeta E(t) - \gamma I(t) \\
\dot{R}(t) &= \gamma I(t).
\end{aligned} \tag{3.8}$$

where β is the probability of infection when a contact takes place, χ the contact rate, γ is the recovery rate and ζ is the latency rate, with,

$$S(0) > 0, E(0) > 0, I(0) \geq 0, \text{ and} R(0) = 0, \tag{3.9}$$

and the total population N is constant:

$$N(t) = S(t) + E(t) + I(t) + R(t). \tag{3.10}$$

To demonstrate the SEIR model, we show an example with a transmission and recovery rate as in Figure 3.3 (i.e., $\beta\chi = 0.2, \gamma = 0.1$) and a latency rate $\zeta = 1$ in Figure 3.4.

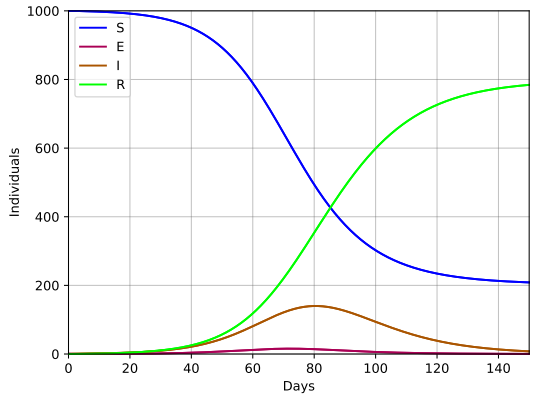


Figure 3.4: The output of a SEIR model with transmission rate $\beta\chi = 0.2$, recovery rate $\gamma = 0.1$, latency rate $\zeta = 1$, population count $N = 1000$ and $E(0) = 1, I(0) = 0$, for each of the four state variables.

1.3 Age-heterogeneous SIR model

The SIR model assumes that all individuals in the population mix homogeneously. This is a strong assumption that is untenable when school closure or vaccine allocation policies need to be evaluated, as such allocation schemes need to take into account age-dependent mixing. To incorporate age-dependent mixing in the SIR model, we can consider a set of n disjoint age groups and maintain a SIR model SIR_i for each age group i. The different age-specific SIR models are then connected to model age-dependent mixing between the different age groups.

We formalize this age-heterogeneous SIR model as a system of ordinary differential equations, as shown in Definition 15.

Definition 15: Age-heterogeneous SIR model

For each age group i, out of a set of n disjoint age groups, we have:

$$\begin{aligned}\dot{S}_i(t) &= -\beta S_i(t) \sum_{j=0}^{n} M_{ij} \frac{I_j(t)}{N_j(t)} \\ \dot{I}_i(t) &= \beta S_i(t) \sum_{j=0}^{n} M_{ij} \frac{I_j(t)}{N_j(t)} - \gamma I_i(t) \\ \dot{R}_i(t) &= \gamma I_i(t),\end{aligned} \tag{3.11}$$

where β is the probability of infection when a contact takes place, γ is the recovery rate and M_{ij} is the average frequency of contacts that an individual in age group i has with an individual in age group j, with,

$$S_i(0) > 0, E_i(0) > 0, I_i(0) \geq 0, \text{ and } R_i(0) = 0, \tag{3.12}$$

and the total population N_i is constant:

$$N_i(t) = S_i(t) + E_i(t) + I_i(t) + R_i(t). \tag{3.13}$$

Compared to Definition 12, we have a separate SIR model for each of the age groups. Furthermore, for each SIR_i, we have a term that consists of the fraction of infected in each of the age groups j,

$$\frac{I_j(t)}{N_j(t)}, \tag{3.14}$$

weighed by the average mixing frequency between age group i and j,

$$M_{ij}. \tag{3.15}$$

From this definition it is clear that we need information on the mixing between the different age groups, and this is typically recorded in a contact matrix M, which can be established by conducting surveys [Mossong et al., 2008].

To demonstrate the age-heterogeneous SIR model, we will use a model that considers two age groups (i.e., children and adults). This model structure is schematically depicted in Figure 3.5.

We use the population census and contact matrix presented in Sherry Tower's lectures[1].

[1] http://sherrytowers.com/2012/12/11/sir-model-with-age-classes/

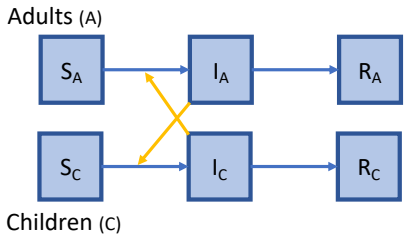


Figure 3.5: We depict an age-heterogeneous SIR model that considers two age groups (i.e., adults and children). This model consists out of two SIR models, one for each age group, that are connected to represent mixing between the age groups (yellow arrows).

The population census data, indexed with C for children and A for adults:

$$\begin{aligned} N_C &= 1500000 \\ N_A &= 3500000 \\ S_C(0) &= N_C - 1 \\ S_A(0) &= N_A - 1 \\ I_C(0) &= I_A(0) = 1, \end{aligned} \tag{3.16}$$

and the contact matrix:

$$M = \begin{array}{c} \\ C \\ A \end{array} \begin{array}{c} \begin{array}{cc} C & A \end{array} \\ \begin{bmatrix} 18 & 9 \\ 3 & 12 \end{bmatrix} \end{array}. \tag{3.17}$$

Note that the contact matrix M should be reciprocal, such that we have:

$$N_i M_{ij} = N_j M_{ji}, \tag{3.18}$$

which is the case for our example matrix in Equation 3.17.

We demonstrate this age-heterogeneous SIR model with the above contact matrix and population census in Figure 3.6.

In Chapter 6, we will create a new meta-population multi-patch model where each patch consists out of an age-heterogeneous SEIR model with four different age classes: children, adolescents, adults and elderly.

1.4 Stochastic SIR model

Different approaches exist to sample trajectories from a compartment model.

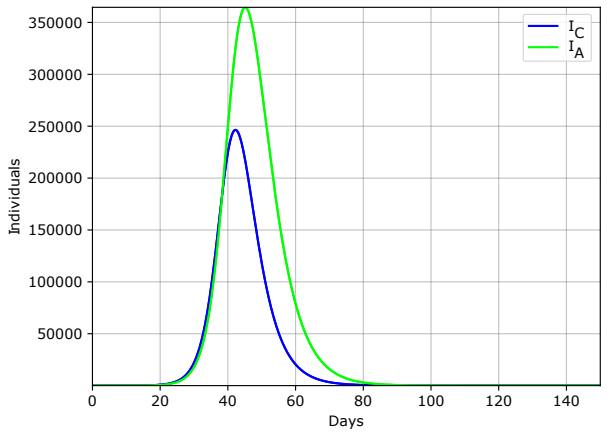


Figure 3.6: The output of a SIR model with transmission rate $\beta = 0.033$, recovery rate $\gamma = \frac{1}{3}$, census data as specified in Equation 3.16 and contact matrix as specified in Equation 3.17. We show the infection curve for the children (I_C) and adults (I_A).

An important method is the Gillespie algorithm [Gillespie, 1977], which considers the epidemic environment as a chemical process. As such, individuals are considered to be the reactants of the type of their respective compartment and the transition between compartments (e.g., infection or recovery) is considered a chemical reaction. The Gillespie algorithm allows us to generate an exact stochastic trajectory of this chemical equation. The algorithm uses Monte Carlo sampling to determine which reaction should take place next and at which time, where the probability to choose a reaction is proportional to the number of substrate molecules (i.e., available individuals), and the time interval is an exponential random variable parametrized with the total reaction time. While the original algorithm [Gillespie, 1977] assumes constant rates, the algorithm was extended towards time-dependent propensities and reaction delays [Cai, 2007; Anderson, 2007]. The ability to sample exact trajectories comes with a computational cost, and thus several adaptations of the Gillespie algorithm exist: adaptations that allow to sample exact trajectories [Gibson and Bruck, 2000; Anderson, 2007; Cai, 2007; Slepoy et al., 2008] and adaptations that generate approximative trajectories [Cao et al., 2006; Anderson, 2008].

Another major method is the use of stochastic differential equations. In order to obtain stochastic trajectories from a compartment model, we transform the system of ordinary differential equations (ODEs) to a system of stochastic differential equations (SDEs), using the transformation procedure presented by Allen et al. [2008]. This procedure considers the compartments and transitions of the original ODE and adds noise terms for each

transition in the ODE. For a transition $(X \to Y)$ between compartment X and Y with rate $\xi_X(t)$, we have a noise term

$$\sqrt{\xi_X(t)X(t)}\dot{\mathcal{W}}_{(X\to Y)}(t), \tag{3.19}$$

where,

$$\mathcal{W}_{(X\to Y)}(t), \tag{3.20}$$

is a Wiener process. This term is subtracted in the differential equation of the outgoing compartment and added in the differential equation of the receiving compartment.

As an example, we will apply this procedure to the SIR model specified in Section 1.1, which results in this system of SDEs:

$$\begin{aligned}
\dot{S} &= -\beta\chi S\frac{I}{N} - \sqrt{\beta\chi S\frac{I}{N}}\cdot\dot{\mathcal{W}}_{(S\to I)} \\
\dot{I} &= \beta\chi S\frac{I}{N} + \sqrt{\beta\chi S\frac{I}{N}}\cdot\dot{\mathcal{W}}_{(S\to I)} - \gamma I - \sqrt{\gamma I}\cdot\dot{\mathcal{W}}_{(I\to R)} \\
\dot{R} &= \gamma I + \sqrt{\gamma I}\cdot\dot{\mathcal{W}}_{(I\to R)}
\end{aligned}. \tag{3.21}$$

This system of SDEs can be simulated to obtain stochastic trajectories, for instance by use of the Euler-Maruyama approximation method [Allen et al., 2008; Rasmussen et al., 2011]. Using Euler's method, for each compartment we consider each deterministic term and multiply it by Δt and each stochastic term and multiply it by $\sqrt{\Delta t}$. For the SDE system in Equation 3.21, this renders this system of simulation equations:

$$\begin{aligned}
\Delta S &= -\Delta t\cdot\beta\chi S\frac{I}{N} - \sqrt{\Delta t}\sqrt{\beta\chi S\frac{I}{N}}\cdot\mathcal{N}(0,1) \\
\Delta I &= \Delta t\cdot\beta\chi S\frac{I}{N} + \sqrt{\Delta t}\sqrt{\beta\chi S\frac{I}{N}}\cdot\mathcal{N}(0,1) - \Delta t\cdot\gamma I - \sqrt{\Delta t}\sqrt{\gamma I}\cdot\mathcal{N}(0,1) \\
\Delta R &= \Delta t\cdot\gamma I + \sqrt{\Delta t}\sqrt{\gamma I}\cdot\mathcal{N}(0,1)
\end{aligned}. \tag{3.22}$$

To demonstrate this process, in Figure 3.7, we show a number of stochastic trajectories for the SIR model example that was introduced in Section 1.1 and the age-heterogeneous SIR model that was introduced in Section 1.3.

In the meta-population multi-patch model that we will introduce in Chapter 6, each patch is represented by a compartment model. We will evaluate these compartment models stochastically, by using stochastic differential equations, that consider time-dependent rates.

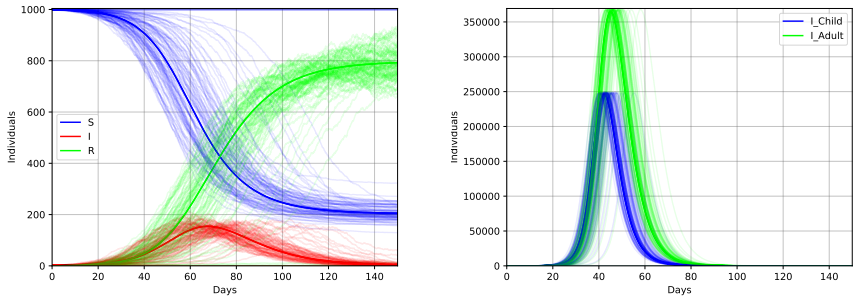


Figure 3.7: We evaluate the SIR model (Section 1.1, left panel) and the age-heterogeneous SIR model (Section 1.3, right panel) stochastically and show 100 stochastic trajectories for each of the models.

2 Individual-based models

While compartment models group individuals together based on common properties (e.g., infection or age group), an individual-based model will explicitly represent all individuals and their properties. In this model individuals are connected, either in a static or a dynamic fashion, and the spread of the epidemic is simulated among this network.

The most fundamental individual-based model will thus represent each of the individuals and maintain their infection status, i.e., susceptible (S), infected (I), recovered (R). Given this fundamental example, these individuals can be connected using a static network, e.g., an Erdős–Rényi random graph, which has a binomial degree distribution. This set-up is depicted schematically in Figure 3.8.

While this simple example only models one property of the individual, many properties of the individual and its environment can be represented. Individual properties include properties related to the infection progress (e.g., the level of infectiousness for influenza [Chao et al., 2010], or the viral load for HIV [Herbeck et al., 2014]), properties related to prevention (e.g., vaccination status [Chao et al., 2012], or condom use status [Kasaie et al., 2018]), and properties related to network formation (e.g., location of the work place [Chao et al., 2010], or the individual's capacity to have simultaneous sexual relationships [Schmid and Kretzschmar, 2012]). Environmental properties include ecological properties (e.g., other animals to model the complex life-cycle of the *Trypanosoma brucei* parasite[2] [Alderton et al., 2016]), and properties with respect to human settlement (e.g.,

[2]The *Trypanosoma brucei* parasite causes Human African trypanosomiasis, i.e., sleeping sickness

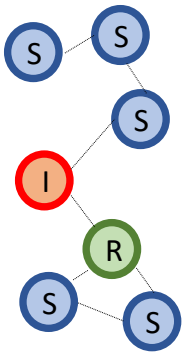


Figure 3.8: We depict a simple individual-based model where we model each individual and their infection state (i.e., S, I, R). The individuals are connected by a static network, over which the epidemic can be simulated.

urban versus rural agglomeration [Singh et al., 2019], or human mobility (e.g., Yan et al. [2017]). Furthermore, the network structure that connects the individuals can be static or dynamic. For static networks, the choice of the network structure depends on the route of transmission (e.g., sexual contact networks are scale-free [Liljeros et al., 2001]). Moreover, networks can overlap, e.g., on the one hand, in the daytime, adults are connected with co-workers and children go to school, on the other hand, in the evening, adults meet with their family [Glass et al., 2006]. For truly dynamic networks, there needs to be a mechanism in place to break and form ties between individuals. An example of this is the method presented by Schmid and Kretzschmar [2012] that forms sexual networks by breaking and forming ties based on the individual's capacity to form simultaneous sexual relationships.

From this description it is clear that individual-based models allow us to model epidemics on a fine-grained level. However, there are three important caveats. Firstly, the amount of detail that is incorporated in the model is directly proportional to the model's computational burden [Chao et al., 2010]. Secondly, to inform such a model, it is necessary to have knowledge on the statistical distribution of the different properties. For some properties, such as for example population density, this information is readily available thanks to geographical information systems (GIS). For other properties, such as to model zoonosis, it remains difficult to find the appropriate data to parametrize such models. Thirdly, due to the extensive model output, it can be difficult to obtain insight in the key determinants of this output [Ball et al., 2015].

In this dissertation, we will use an individual-based model for pandemic influenza, FluTE. We will introduce this model in Chapter 4.

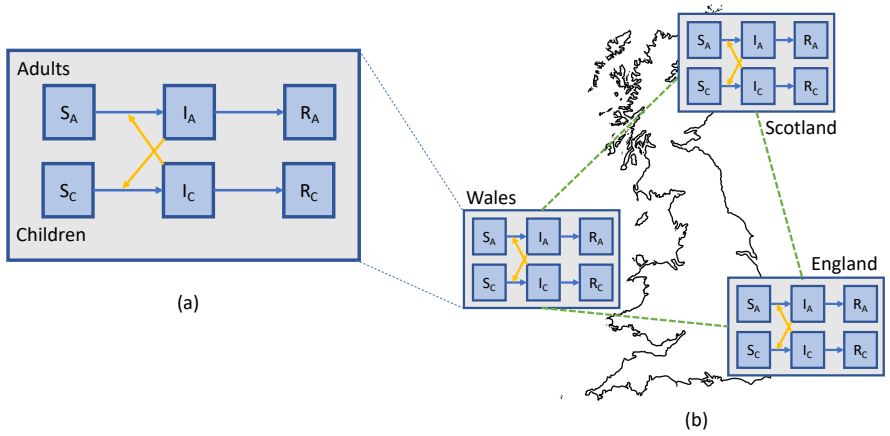


Figure 3.9: We depict the age-dependent SIR model from Section 1.3 (a), and use it in a meta-population model that has patches for the three countries in Great Britain: England, Scotland, Wales.

3 Meta-population models

On the one hand, compartment models are computationally efficient to evaluate, yet they need to divide the population in coarse compartments. On the other hand, individual-based models allow for fine-grained modelling, yet require a lot of computation. To balance between these trade-offs meta-population models are frequently used, especially to enable the modelling of epidemic processes in a spatially explicit context. Meta-population models were first introduced in the context of ecology [Hanski et al., 1999], to model sub-populations that can be separated geographically. In Figure 3.9, we show an example of a simple meta-population model with a patch for each of the countries in Great Britain.

In this dissertation (Chapter 6) we will create a new meta-population multi-patch model, where each patch corresponds to an administrative region in Great Britain. Each patch is internally represented by an age-dependent stochastic SEIR model.

4 | Bayesian bandits for decision making in an influenza pandemic

> If there is not folly in the world, then the world itself is folly. You must understand that mistakes are not always regrets.
>
> Paul Tobin *Bandette, Volume 1: Presto!*

Pandemic influenza has the epidemic potential to kill millions of people. While various preventive measures exist (i.a., vaccination and school closures), deciding on strategies that lead to their most effective and efficient use remains challenging. To this end, individual-based epidemiological models are essential to assist decision makers in determining the best strategy to curb epidemic spread. However, individual-based models are computationally intensive and it is therefore pivotal to identify the optimal strategy using a minimal amount of model evaluations. Additionally, as epidemiological modelling experiments need to be planned, a computational budget needs to be specified a priori. Consequently, we present a new sampling technique to optimize the evaluation of preventive strategies using fixed budget best-arm identification algorithms. We use epidemiological modelling theory to derive knowledge about the reward distribution which we exploit by using Bayesian best-arm identification algorithms that can incorporate this prior knowledge (i.e., Top-two Thompson

sampling and BayesGap). We evaluate these algorithms in a realistic experimental setting and demonstrate that it is possible to identify the optimal strategy using only a limited number of model evaluations, i.e., 2-to-3 times faster compared to the uniform sampling method, the predominant technique used for epidemiological decision making in the literature. Finally, we contribute and evaluate a statistic for Top-two Thompson sampling to inform the decision makers about the confidence of an arm recommendation.

The work presented in this chapter was published in the ECML and AAMAS proceedings.

Libin, P., Verstraeten, T., Roijers, D. M., Grujic, J., Theys, K., Lemey, P., and Nowé, A. (2018, September). Bayesian best-arm identification for selecting influenza mitigation strategies. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 456-471). Springer, Cham.

Libin, P., Verstraeten, T., Theys, K., Roijers, D. M., Vrancx, P., and Nowé, A. (2017, May). Efficient evaluation of influenza mitigation strategies using preventive bandits. In International Conference on Autonomous Agents and Multiagent Systems (pp. 67-85). Springer, Cham.

1 Rationale and objectives

The influenza virus is responsible for the deaths of half of a million people each year. In addition, seasonal influenza epidemics cause a significant economic burden. While transmission is primarily local, a newly emerging variant may spread to pandemic proportions in a fully susceptible host population [Paules and Subbarao, 2017]. Pandemic influenza occurs less frequently than seasonal influenza but the outcome with respect to morbidity and mortality can be much more severe, potentially killing millions of people worldwide [Paules and Subbarao, 2017]. Therefore, it is essential to study mitigation strategies to control influenza pandemics (more details on pandemic influenza in Section 2 in Chapter 1).

For influenza, different preventive measures exist: i.a., vaccination, social measures (e.g., school closures and travel restrictions) and antiviral drugs. However, the efficiency of strategies greatly depends on the availability of preventive compounds, as well as on the characteristics of the targeted epidemic. Furthermore, governments typically have limited resources to implement such measures. Therefore, it remains challenging to formulate public health strategies that make effective and efficient use of these preventive measures within the existing resource constraints.

Epidemiological models (i.e., compartment models and individual-based models) are essential to study the effects of preventive measures *in silico* [Basta et al., 2009; Germann et al., 2006]. While individual-based models are usually associated with a greater model

complexity and computational cost than compartment models, they allow for a more accurate evaluation of preventive strategies [Eubank et al., 2006]. To capitalize on these advantages and make it feasible to employ individual-based models, it is essential to use the available computational resources as efficiently as possible.

In the literature, a set of possible preventive strategies is typically evaluated by simulating each of the strategies an equal number of times [Fumanelli et al., 2016; Ferguson et al., 2005; Chao et al., 2012]. However, this approach is inefficient to identify the optimal preventive strategy, as a large proportion of computational resources will be used to explore sub-optimal strategies. Furthermore, a consensus on the required number of model evaluations per strategy is currently lacking [Willem et al., 2014] and we show that this number depends on the *hardness* of the evaluation problem. For this reason, we propose to combine individual-based epidemiological models with *multi-armed bandits*. Additionally, we recognize that epidemiological modelling experiments need to be planned and that a computational budget needs to be specified a priori. Therefore, we present a novel approach where we formulate the evaluation of preventive strategies as a *best-arm identification* problem using a *fixed budget* of model evaluations.

As running an individual-based model is computationally intensive (i.e., minutes to hours, depending on the complexity of the model), minimizing the number of required model evaluations reduces the total time required to evaluate a given set of preventive strategies. This renders the use of individual-based models attainable in studies where it would otherwise not be computationally feasible. Additionally, reducing the number of model evaluations will free up computational resources in studies that already use individual-based models, capacitating researchers to explore a larger set of model scenarios. This is important, as considering a wider range of scenarios increases the confidence about the overall utility of preventive strategies [Wu et al., 2006].

In this chapter, we contribute a novel technique to evaluate preventive strategies as a fixed budget best-arm identification problem, i.e., epidemic bandits. We employ epidemiological modelling theory to derive assumptions about the reward distribution and exploit this knowledge using Bayesian algorithms. This new technique enables decision makers to obtain recommendations in a reduced number of model evaluations. We evaluate the technique in an experimental setting, where we aim to find the best vaccine allocation strategy in a realistic simulation environment that models an influenza pandemic on a large social network. Finally, we contribute and evaluate a statistic to inform the decision makers about the confidence of a particular recommendation.

2 Background

2.1 Pandemic influenza and vaccine production

The primary preventive strategy to mitigate seasonal influenza[1] is to produce vaccine prior to the epidemic, anticipating the virus strains that are expected to circulate. This vaccine pool is used to inoculate the population before the start of the epidemic. While seasonal influenza may have a restricted susceptible population due to vaccination and pre-existing immunity, a newly emerging strain can become pandemic by spreading rapidly among naive human hosts worldwide [Paules and Subbarao, 2017].

While it is possible to stockpile vaccines to prepare for seasonal influenza, this is not the case for influenza pandemics, as the vaccine should be specifically tailored to the virus that is the source of the pandemic. Therefore, before an appropriate vaccine can be produced, the responsible virus needs to be identified. Hence, vaccines will be available only in limited supply at the beginning of the pandemic [WHO, 2004]. In addition, production problems can result in vaccine shortages [Enserink, 2004]. When the number of vaccine doses is limited, it is imperative to identify an optimal vaccine allocation strategy [Medlock and Galvani, 2009].

2.2 Modelling influenza

There is a long tradition of using individual-based models to study influenza epidemics [Basta et al., 2009; Germann et al., 2006; Fumanelli et al., 2016], as they allow for a more fine-grained evaluation of preventive strategies. A state-of-the-art individual-based model that has been the driver for many high impact research efforts [Basta et al., 2009; Germann et al., 2006; Halloran et al., 2002], is FluTE [Chao et al., 2010].

FluTE implements a contact model where the population is divided into communities of households [Chao et al., 2010]. The population is organized in a hierarchy of social mixing groups where the contact intensity is inversely proportional with the size of the group (e.g., closer contact between members of a household than between colleagues). Additionally, FluTE implements an individual disease progression model that associates different disease stages with different levels of infectiousness. FluTE supports the evaluation of preventive strategies through the simulation of therapeutic interventions (i.e., vaccines, antiviral compounds) and non-therapeutic interventions (i.e., school closure, case isolation, household quarantine).

[1] More background on seasonal and pandemic influenza in Section 2 of Chapter 1.

2.3 Bandits and best-arm identification

We define the multi-armed bandit in Section 1 of Chapter 2 (Definition 1). To remind the reader, a multi-armed bandit has K arms, where each arm a_k returns a reward r_k when it is pulled.

Our objective is to recommend the best arm a_* (i.e., the arm with the highest average reward μ_*, as specified in Definition 2), after a fixed number of arm pulls. This is referred to as the fixed budget best-arm identification problem [Audibert and Bubeck, 2010], an instance of the pure-exploration problem [Bubeck et al., 2009]. For a given budget T, the objective is to minimize the *simple regret* (Definition 4), To remind the reader, the simple regret considers the average reward of the best arm a_* and the recommended arm $J^{(T)}$, i.e.,

$$\mu_* - \mu_{J^{(T)}} \tag{4.1}$$

Simple regret is inversely proportional to the probability of recommending the correct arm a_* [Kaufmann et al., 2016].

3 Related work

As we established that a computational budget needs to be specified a priori, our problem setting matches the fixed budget best-arm identification setting. This differs from settings that attempt to identify the best arm with a predefined confidence: i.e., racing strategies [Even-Dar et al., 2006], strategies that exploit the confidence bound of the arms' means [Kaufmann and Kalyanakrishnan, 2013] and more recently fixed confidence best-arm identification algorithms [Garivier and Kaufmann, 2016].

We selected Bayesian fixed budget best-arm identification algorithms, as we aim to incorporate prior knowledge about the arms' reward distributions and use the arms' posteriors to define a statistic to support policy makers with their decisions. We refer to [Kaufmann et al., 2016; Hoffman et al., 2014], for a broader overview of the state-of-the-art with respect to (Bayesian) best-arm identification algorithms.

Best-arm identification algorithms have been used in a large set of application domains: i.a., evaluation of response surfaces, the initialization of hyper-parameters and traffic congestion.

While other algorithms exist to rank or select bandit arms, e.g. [Powell and Ryzhov, 2012], best-arm identification is best approached using adaptive sampling methods [Jennison et al., 1982], as the ones we study in this paper. Moreover, the use of best-arm identification methods clears the way for interesting future work with respect to evaluating preventive strategies while considering multiple objectives.

4 Epidemic bandits

We formulate the evaluation of preventive strategies as a multi-armed bandit problem with the aim of identifying the best arm using a fixed budget of model evaluations. The presented method is generic with respect to the type of epidemic that is modelled (i.e., pathogen, contact network, preventive strategies). The method is evaluated in the context of pandemic influenza in the next section.

4.1 Evaluating preventive strategies with bandits

First, we provide a formal definition of the epidemic model we consider.

Definition 16: Stochastic epidemiological model

A *stochastic epidemiological model* $\mathcal{E}$ is defined in terms of a model configuration $c \in \mathcal{C}$ and can be used to evaluate a preventive strategy p.

The result of a model evaluation is referred to as the *model outcome*. Evaluating the model $\mathcal{E}$ thus results in a sample of the model's *outcome distribution*:

$$\text{outcome} \sim \mathcal{E}(c, p) \tag{4.2}$$

The model outcome can be any statistic relevant to the decision maker, e.g., prevalence, proportion of symptomatic individuals, morbidity, mortality, societal cost.

Note that a model configuration $c \in \mathcal{C}$ describes the complete model environment, i.e., both aspects inherent to the model (e.g., FluTE's mixing model) and options that the modeller can provide (e.g., population statistics, vaccine properties).

Our objective is to find the optimal preventive strategy (i.e., the strategy that minimizes the expected outcome) from a set of alternative strategies

$$\{p_1, ..., p_K\}, \tag{4.3}$$

for a particular configuration

$$c_0 \in \mathcal{C}, \tag{4.4}$$

of a stochastic epidemiological model, where c_0 corresponds to the context of the studied epidemic. To this end, we consider a multi-armed bandit with

$$|\{p_1, ..., p_K\}| \tag{4.5}$$

arms. Pulling arm p_k corresponds to evaluating p_k by running a simulation in the epidemiological model $\mathcal{E}(c_0, p_k)$. The bandit thus has preventive strategies as arms with

reward distributions corresponding to the outcome distribution of a stochastic epidemiological model $\mathcal{E}(c_0, p_k)$. While the parameters of the reward distribution are known (i.e., the parameters of the epidemiological model), it is intractable to determine the optimal reward analytically. Hence, we must learn about the outcome distribution via interaction with the epidemiological model. In this work, we consider prevention strategies of equal financial cost, which is a realistic assumption, as governments typically operate within budget constraints.

4.2 Outcome distribution

As previously defined, the reward distribution associated with a bandit's arm corresponds to the outcome distribution of the epidemiological model that is evaluated when pulling that arm. Therefore, employing insights from epidemiological modelling theory allows us to specify prior knowledge about the reward distribution.

It is well known that a disease outbreak has two possible outcomes: either it is able to spread beyond a local context and becomes a fully established epidemic or it fades out [Watts et al., 2005]. Most stochastic epidemiological models reflect this reality and hence its epidemic size distribution is bimodal [Watts et al., 2005]. When evaluating preventive strategies, the objective is to determine the preventive strategy that is most suitable to mitigate an established epidemic. As in practice we can only observe and act on established epidemics, epidemics that faded out in simulation would bias this evaluation. Consequently, it is necessary to focus on the mode of the distribution that is associated with the established epidemic. Therefore we censor (i.e., discard) the epidemic sizes that correspond to the faded epidemic. The size distribution that remains (i.e., the one that corresponds with the established epidemic) is approximately Gaussian [Britton, 2010].

In this study, we consider a scaled epidemic size distribution, i.e., the proportion of symptomatic infections. Hence we can assume bimodality of the full size distribution and an approximately Gaussian size distribution of the established epidemic. We verified experimentally that these assumptions hold for all the reward distributions that we observed in our experiments (see Section 5).

To censor the size distribution, we use a threshold that represents the number of infectious individuals that are required to ensure an outbreak will only fade out with a low probability.

4.3 Epidemic fade-out threshold

For heterogeneous host populations (i.e., a population with a significant variance among individual transmission rates, as is the case for influenza epidemics [Dorigatti et al., 2012; Fraser et al., 2011]), the number of secondary infections can be accurately modelled using

a negative binomial *offspring distribution* NB(R_0, ι) [Lloyd-Smith et al., 2005], where R_0 is the basic reproductive number (Definition 13) and ι is a dispersion parameter that specifies the extent of heterogeneity.

Definition 17: Probability of epidemic extinction

The probability of epidemic extinction p_{ext} can be computed by solving $g(s) = s$, where $g(s)$ is the probability generating function of the offspring distribution [Lloyd-Smith et al., 2005]. For an epidemic where individuals are targeted with preventive measures (i.e., vaccination in our case), we obtain the following probability generating function

$$g(s) = \text{pop}_c + (1 - \text{pop}_c)\big(1 + \frac{R_0}{\iota}(1 - s)\big)^{-\iota} \tag{4.6}$$

where pop_c signifies the random proportion of controlled individuals [Lloyd-Smith et al., 2005]. From p_{ext} we can compute a threshold T_0 to limit the probability of extinction to a cutoff ℓ [Hartfield and Alizon, 2013]:

$$\big(p_{\text{ext}}\big)^{T_0} = \ell \tag{4.7}$$

4.4 Best-arm identification with a fixed budget

Our objective is to identify the best preventive strategy (i.e., the strategy that minimizes the expected outcome) out of a set of preventive strategies, for a particular configuration $c_0 \in \mathcal{C}$ using a fixed budget T of model evaluations. To find the best prevention strategy, it suffices to focus on the mean of the outcome distribution, as it is approximately Gaussian with an unknown yet small variance [Britton, 2010], as we confirm in our experiments (see Figure 4.1).

Successive rejects

Successive Rejects was the first algorithm to solve the best-arm identification in a fixed budget setting [Audibert and Bubeck, 2010]. For a K-armed bandit, Successive Rejects operates in $(K - 1)$ phases. At the end of each phase, the arm with the lowest average reward is discarded. Thus, at the end of phase $(K - 1)$ only one arm survives, and this arm is recommended. At phase $f \in \{1, \ldots, K - 1\}$, each arm that is still available is

played $m_f - m_{f-1}$ times, where:

$$\begin{aligned} m_0 &= 0 \\ m_f &= \left\lceil \frac{T-K}{K+1-f} \frac{1}{\overline{\log}(K)} \right\rceil, \end{aligned} \tag{4.8}$$

with,

$$\overline{\log}(K) = \frac{1}{2} + \sum_{k=2}^{K} \frac{1}{k}. \tag{4.9}$$

Bayesian best-arm identification

Successive Rejects serves as a useful baseline, however, it has no support to incorporate any prior knowledge. Bayesian best-arm identification algorithms on the other hand, are able to take into account such knowledge by defining an appropriate prior on the arms' reward distribution. As we will show, such prior knowledge can increase the best-arm identification accuracy. Additionally, at the time an arm is recommended, the posteriors contain valuable information that can be used to formulate a variety of statistics helpful to assist decision makers. We consider two state-of-the-art Bayesian algorithms: BayesGap [Hoffman et al., 2014] and Top-two Thompson sampling [Russo, 2016]. For Top-two Thompson sampling, we derive a statistic based on the posteriors to inform the decision makers about the confidence of an arm recommendation: the probability of success.

Prior and posterior of the reward distribution

As we established in the previous section, each arm of our bandit has a reward distribution that is approximately Gaussian with unknown mean and variance. For the purpose of genericity, we assume an uninformative Jeffreys prior

$$(\sigma_k)^{-3}, \tag{4.10}$$

on

$$(\mu_k^{(t)}, \sigma_k^2), \tag{4.11}$$

which leads to the following posterior on $\mu_k^{(t)}$ at the $n_k^{(t)}$'th pull [Honda and Takemura, 2014]:

$$\pi_{\mathcal{T}}^{(t)} = \sqrt{\frac{(n_k^{(t)})^2}{S_k^{(t)}}}(\mu_k^{(t)} - \hat{\mu}_k^{(t)}) \mid \hat{\mu}_k^{(t)}, S_k^{(t)} \sim \mathcal{T}_{n_k^{(t)}} \tag{4.12}$$

where $\hat{\mu}_k^{(t)}$ is the emperical reward mean, $S_k^{(t)}$ is the sum of squares

$$S_k^{(t)} = \sum_{m=1}^{n_k^{(t)}} (r_{k,m} - \hat{\mu}_k^{(t)})^2 \tag{4.13}$$

and $\mathcal{T}_{n_k^{(t)}}$ is the standard student t-distribution with $n_k^{(t)}$ degrees of freedom.

BayesGap

BayesGap is a gap-based Bayesian algorithm [Hoffman et al., 2014]. The algorithm requires that for each arm a_k, a high-probability upper bound $U_k^{(t)}$ and lower bound $L_k^{(t)}$ is defined on the posterior of $\mu_k^{(t)}$ at each time step t. Using these bounds, the gap quantity

$$B_k^{(t)} = \max_{l \neq k} U_l^{(t)} - L_k^{(t)} \tag{4.14}$$

is defined for each arm a_k. $B_k^{(t)}$ represents an upper bound on the simple regret (as defined in Section 2.3). At each step t of the algorithm, the arm $\mathcal{G}^{(t)}$ that minimizes the gap quantity $B_k^{(t)}$ is compared to the arm $\mathcal{g}^{(t)}$ that maximizes the upper bound $U_k^{(t)}$. From $\mathcal{G}^{(t)}$ and $\mathcal{g}^{(t)}$, the arm with the highest confidence diameter

$$U_k^{(t)} - L_k^{(t)}, \tag{4.15}$$

is pulled. The reward that results from this pull is observed and used to update a_k's posterior. When the budget is consumed, the arm

$$J^{(T)} = \mathcal{G}(\operatorname*{argmin}_{t \leq T} B_{\mathcal{G}^{(t)}}^{(t)}) \tag{4.16}$$

is recommended. This is the arm that minimizes the simple regret bound over all times $t \leq T$.

In order to use BayesGap to evaluate preventive strategies, we contribute problem-specific bounds. Given our posteriors (Equation 4.12), we define

$$\begin{aligned} U_k^{(t)} &= \mathbb{E}\left[\pi_{\mathcal{T}}^{(t)}\right]_k + \kappa \sqrt{\mathbb{V}\left[\pi_{\mathcal{T}}^{(t)}\right]_k} \\ L_k^{(t)} &= \mathbb{E}\left[\pi_{\mathcal{T}}^{(t)}\right]_k - \kappa \sqrt{\mathbb{V}\left[\pi_{\mathcal{T}}^{(t)}\right]_k} \end{aligned} \tag{4.17}$$

where $\mathbb{E}\left[\pi_{\mathcal{T}}^{(t)}\right]_k$ is the mean, and $\mathbb{V}\left[\pi_{\mathcal{T}}^{(t)}\right]_k$ is the variance of the posterior of arm a_k at time step t, and κ is the exploration coefficient.

The amount of exploration that is feasible given a particular bandit, is proportional to the available budget, and inversely proportional to the bandit's complexity [Hoffman et al., 2014]. This complexity can be modelled taking into account the game's hardness [Audibert and Bubeck, 2010] and the variance of the rewards. We use the hardness quantity defined in [Hoffman et al., 2014] (Definition 18).

Definition 18: ϵ-hardness

$$H_\epsilon = \sum_k H_{k,\epsilon}^{-2}, \tag{4.18}$$

with arm-dependent hardness,

$$H_{k,\epsilon} = \max(\frac{1}{2}(\Delta_k + \epsilon), \epsilon), \tag{4.19}$$

where,

$$\Delta_k = \max_{l \neq k}(\mu_l - \mu_k) \tag{4.20}$$

Considering the budget T, hardness H_ϵ and a generalized reward variance σ_G^2 over all arms, we have exploration term:

$$\kappa = \sqrt{\frac{T - 3K}{4H_\epsilon \sigma_G^2}} \tag{4.21}$$

In Appendix 1, we formally prove that using these bounds results in a probability of simple regret that asymptotically reaches the exponential lower bound of [Hoffman et al., 2014].

As both H_ϵ and σ_G^2 are unknown, in order to compute κ, these quantities need to be estimated. Firstly, we estimate H_ϵ's upper bound $\hat{H}_\epsilon$ by estimating Δ_k as follows

$$\hat{\Delta}_k = \max_{1 \leq l < K; l \neq k} \delta_l - \delta_k, \tag{4.22}$$

where,

$$\delta_m = \mathbb{E}\left[\pi_{\mathcal{T}}^{(t)}\right]_m - 3 \cdot \sqrt{\mathbb{V}\left[\pi_{\mathcal{T}}^{(t)}\right]_m}, \tag{4.23}$$

as in [Hoffman et al., 2014], where $\mathbb{E}\left[\pi_{\mathcal{T}}^{(t)}\right]_k$ is the mean, and $\mathbb{V}\left[\pi_{\mathcal{T}}^{(t)}\right]_k$ is the variance of the posterior of arm a_k at time step t. Secondly, for σ_G^2 we need a measure of variance

that is representative for the reward distribution of all arms. To this end, when the arms are initialized, we observe their sample variance s_k^2, and compute their average $\bar{s}_G^2$.

As our bounds depend on the variance $\mathbb{V}\left[\pi_{\mathcal{T}}^{(t)}\right]_k$ of the t-distributed posterior, each arm's posterior needs to be initialized 3 times (i.e., by pulling the arm) to ensure that $\mathbb{V}\left[\pi_{\mathcal{T}}^{(t)}\right]_k$ is defined, this initialization also ensures proper posteriors [Honda and Takemura, 2014].

Top-two Thompson sampling

Top-two Thompson sampling is a reformulation of the (vanilla) Thompson sampling algorithm, such that it can be used in a pure-exploration context [Russo, 2016]. Vanilla Thompson sampling[2] operates directly on the arms' posterior of the reward distribution's mean $\mu_k^{(t)}$. At each time step, vanilla Thompson sampling obtains one sample for each arm's posterior. The arm with the highest sample is pulled, and its reward is subsequently used to update that arm's posterior. While this approach has been proven highly successful to minimize cumulative regret [Chapelle and Li, 2011; Honda and Takemura, 2014], as it balances the exploration-exploitation trade-off, it is sub-optimal to identify the best arm [Bubeck et al., 2009]. To adapt Thompson sampling to minimize simple regret, Top-two Thompson sampling increases the amount of exploration. To this end, an exploration probability w needs to be specified. At each time step, one sample is obtained for each arm's posterior. The arm a_{top} with the highest sample is only pulled with probability w. With probability $1-w$ we repeat sampling from the posteriors until we find an arm $a_{\text{top-2}}$ that has the highest posterior sample and where $a_{\text{top}} \neq a_{\text{top-2}}$. When the arm $a_{\text{top-2}}$ is found, it is pulled and the observed reward is used to update the posterior of the pulled arm. When the available budget is consumed, the arm with the highest average reward is recommended.

As Top-two Thompson sampling only requires samples from the arms' posteriors, we can use the t-distributed posteriors from Equation 4.12 as is. To avoid improper posteriors, each arm needs to be initialized 2 times [Honda and Takemura, 2014].

Reward censoring

As specified in the previous subsection, the reward distribution is censored. We observe each reward, but only consider it to update the arm's value when it exceeds the threshold T_0 (i.e., when we receive a sample from the mode of the epidemic that represents the established epidemic).

[2] More details on Thompson sampling in Section 2.4 of Chapter 2.

4.5 Probability of success

The probability that an arm recommendation is correct presents a useful confidence statistic to support policy makers with their decisions.

As Top-two Thompson sampling recommends the arm with the highest average reward, and we assume that the arm's reward distributions are independent, the probability of success is:

$$\begin{aligned} P(\mu_J = \max_{1 \leq k \leq K} \mu_k) &= P(\cap_{k \neq J}^{K} (\mu_k \leq \mu_J)) \\ &= \int_{x \in \mathbb{R}} P(\cap_{k \neq J}^{K} (\mu_k \leq x)) P(\mu_J = x) dx \\ &= \int_{x \in \mathbb{R}} \big[\prod_{k \neq J}^{K} P(\mu_k \leq x) \big] P(\mu_J = x) dx \\ &= \int_{x \in \mathbb{R}} \big[\prod_{k \neq J}^{K} F_{\mu_k}(x) \big] f_{\mu_J}(x) dx \end{aligned}$$

where μ_J is the random variable that represents the mean of the recommended arm's reward distribution, f_{μ_J} is the recommended arm's posterior probability density function and F_{μ_k} is the other arms' cumulative density function. As this integral cannot be computed analytically, we estimate it using Gaussian quadrature.

It is important to note that, while aiming for generality, we made some conservative assumptions: the reward distributions are approximated as Gaussian and the uninformative Jeffreys prior is used. These assumptions imply that the derived probability of success will be an under-estimator for the actual recommendation success, which is confirmed in our experiments.

5 Experiments

We composed and performed an experiment in the context of pandemic influenza, where we analyse the mitigation strategy to vaccinate a population when only a limited number of vaccine doses is available (details about the rationale behind this scenario in Section 2.1). In our experiment, we accommodate a realistic setting to evaluate vaccine allocation, where we consider a large and realistic social network and a wide range of R_0 values.

We consider the scenario when a pandemic is emerging in a particular geographical region and vaccines becomes available, albeit in a limited number of doses. When the number of vaccine doses is limited, it is imperative to identify an optimal vaccine allocation strategy

[Medlock and Galvani, 2009]. In our experiment, we explore the allocation of vaccines over five different age groups, that can be easily approached by health policy officials: pre-school children, school-age children, young adults, older adults and the elderly, as proposed in [Chao et al., 2010].

5.1 Influenza model and configuration

The epidemiological model used in the experiments is the FluTE stochastic individual-based model. In our experiment we consider the population of Seattle (United States) that includes 560,000 individuals [Chao et al., 2010]. This population is realistic both with respect to the number of individuals and its community structure, and provides an adequate setting for the validation of vaccine strategies [Willem et al., 2014].

At the first day of the simulated epidemic, 10 random individuals are seeded with an infection. The epidemic is simulated for 180 days, during which time no more infections are seeded. Thus, all new infections established during the run time of the simulation, result from the mixing between infectious and susceptible individuals. We assume no pre-existing immunity towards the circulating virus variant. We choose the number of vaccine doses to allocate to be approximately 4.5% of the population size [Medlock and Galvani, 2009].

We perform our experiment for a set of R_0 values within the range of 1.4 to 2.4, in steps of 0.2. This range is considered representative for the epidemic potential of influenza pandemics [Basta et al., 2009; Medlock and Galvani, 2009]. We refer to this set of R_0 values as $\mathcal{R}_0$, i.e., $\mathcal{R}_0 = \{1.4, 1.6, 1.8, 2.0, 2.2, 2.4\}$.

Note that the setting described in this subsection, in conjunction with a particular R_0 value, corresponds to a model configuration (i.e., $c_0 \in \mathcal{C}$).

The computational complexity of FluTE simulations depends both on the size of the susceptible population and the proportion of the population that becomes infected. For the population of Seattle, the simulation run time was up to $11\frac{1}{2}$ minutes (median of $10\frac{1}{2}$ minutes, standard deviation of 6 seconds), on state-of-the-art hardware (details Appendix 2.4). We present further details on the computational processes that underlie our experiments in Appendix 2.

5.2 Formulating vaccine allocation strategies

We consider 5 age groups to which vaccine doses can be allocated: pre-school children (i.e., 0-4 years old), school-age children (i.e., 5-18 years old), young adults (i.e., 19-29 years old), older adults (i.e., 30-64 years old) and the elderly (i.e., $>$ 65 years old) [Chao et al., 2010]. An allocation scheme can be encoded as a Boolean 5-tuple, where each position in the tuple corresponds to the respective age group. The Boolean value at

a particular position in the tuple denotes whether vaccines should be allocated to the respective age group. When vaccines are to be allocated to a particular age group, this is done proportional to the size of the population that is part of this age group [Medlock and Galvani, 2009]. To decide on the best vaccine allocation strategy, we enumerate all possible combinations of this tuple.

5.3 An influenza preventive bandit

We define the bandit that we use in our experiments. The *influenza preventive bandit* has exactly 32 arms. Each arm a_k is associated with the allocation strategy for which the integer encoding is k.

In this bandit settings, we aim to find the prevention strategy that minimizes the symptomatic attack rate (Definition 19).

> **Definition 19: Attack rate**
>
> The attack rate quantifies the proportion of the population that was infected, and is determined at the end of an outbreak. When only symptomatic infections are considered, this quantity is referred to as symptomatic attack rate.

Given a model configuration $c_0 \in \mathcal{C}$ (Definition 16), when an arm a_k is pulled, FluTE is invoked with model context c_0 and the vaccine allocation strategy p_k (Definition 16) associated with the arm a_k. When FluTE finishes, it outputs the proportion of the population that experienced a symptomatic infection p_I, from which the reward (i.e., symptomatic attack rate, see Definition 19),

$$r_k = 1 - p_I, \tag{4.24}$$

is computed.

5.4 Outcome distributions

To establish a proxy for the ground truth concerning the outcome distributions of the 32 considered preventive strategies, all strategies were evaluated 1000 times, for each of the R_0 values in $\mathcal{R}_0$. We will use this ground truth as a reference to validate the correctness of the recommendations obtained throughout our experiments.

$\mathcal{R}_0$ presents us with an interesting evaluation problem. To demonstrate this, we visualize the outcome distribution for $R_0 = 1.4$ and for $R_0 = 2.4$ in Figure 4.1 (the outcome distributions for the other R_0 values are shown in Appendix 3). Firstly, we observe that for different values of R_0, the distances between top arms' means differ. Additionally, outcome

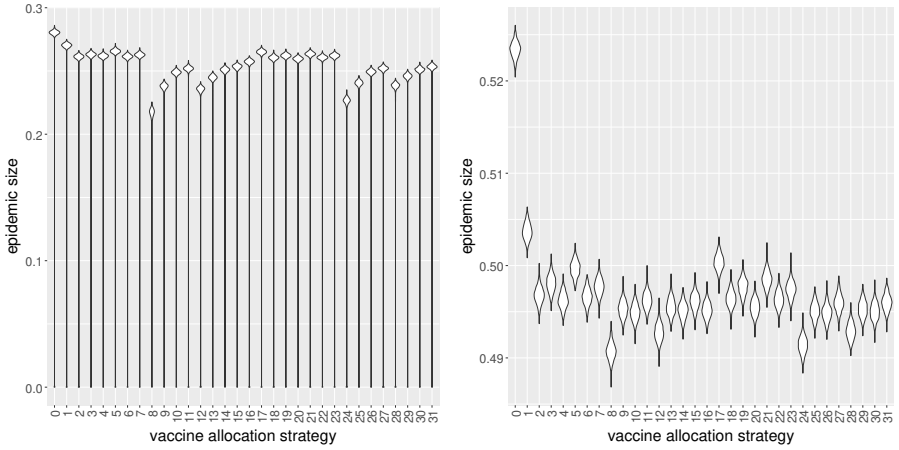


Figure 4.1: Violin plot that depicts the density of the outcome distribution (i.e., epidemic size) for 32 vaccine allocation strategies (left panel $R_o = 1.4$, right panel $R_o = 2.4$).

distribution variances vary over the set of R_0 values in $\mathcal{R}_0$. These differences produce distinct levels of evaluation hardness (see Section 4.4), and demonstrate the setting's usefulness as benchmark to evaluate preventive strategies. While we discuss the hardness of the experimental settings under consideration, it is important to state that our best-arm identification framework requires no prior knowledge on the problem's hardness. Secondly, we expect the outcome distribution to be bimodal. However, the probability to sample from the mode of the outcome distribution that represents the non-established epidemic decreases as R_0 increases [Lloyd-Smith et al., 2005]. This expectation is confirmed when we inspect Figure 4.1, the left panel shows a bimodal distribution for $R_0 = 1.4$, while the right panel shows a unimodal outcome distribution for $R_0 = 2.4$, as only samples from the established epidemic were obtained.

Our analysis identified that the best vaccine allocation strategy was $\langle 0, 1, 0, 0, 0 \rangle$ (i.e., allocate vaccine to school children, strategy 8) for all R_0 values in $\mathcal{R}_0$.

5.5 Best-arm identification experiment

To assess the performance of the different best-arm identification algorithms (i.e., Successive Rejects, BayesGap and Top-two Thompson sampling) we run each algorithm for all budgets in the range of 32 to 500. This evaluation is performed on the influenza bandit that we defined earlier. For each budget, we run the algorithms 100 times, and report the recommendation success rate. In the previous section, the optimal vaccine allocation

strategy was identified to be $\langle 0, 1, 0, 0, 0 \rangle$ (i.e., vaccine allocation strategy 8) for all R_0 in $\mathcal{R}_0$. We thus consider a recommendation to be correct when it equals this vaccine allocation strategy.

We evaluate the algorithm's performance with respect to each other and with respect to uniform sampling, the current state-of-the art to evaluate preventive strategies. The uniform sampling method pulls arm a_u for each step t of the given budget T, where a_u's index u is sampled from the uniform distribution $\mathcal{U}(1, K)$. To consider different levels of hardness, we perform this analysis for each R_0 value in $\mathcal{R}_0$.

For the Bayesian best-arm identification algorithms, the prior specifications are detailed in Section 4.4. BayesGap requires an upper and lower bound that is defined in terms of the used posteriors. In our experiments, we use upper bound $U_k^{(t)}$ and lower bound $L_k^{(t)}$ that were established in Section 4.4. Top-two Thompson sampling requires a parameter that modulates the amount of exploration w. As it is important for best-arm identification algorithms to differentiate between the top two arms, we choose $w = 0.5$, such that, in the limit, Top-two Thompson sampling will explore the top two arms uniformly.

We censor the reward distribution based on the epidemic extinction threshold T_0 (Definition 17). This threshold depends on basic reproductive number R_0 and dispersion parameter ι. R_0 is chosen explicitly for each of our experimental settings. For the dispersion parameter we choose $\iota = 0.5$, which is a conservative choice according to the literature [Dorigatti et al., 2012; Fraser et al., 2011]. We choose the probability cutoff parameter to be $\ell = 10^{-10}$.

Figure 4.2 shows recommendation success rates for each of the best-arm identification algorithms for $R_0 = 1.4$ (left panel) and $R_0 = 2.4$ (right panel). The results for the other R_0 values are visualized in Appendix 4. The results for different values of R_0 clearly indicate that our selection of best-arm identification algorithms significantly outperforms the uniform sampling method. Overall, the uniform sampling method requires more than double the amount of evaluations to achieve a similar recommendation performance. For the harder problems (e.g., setting with $R_0 = 2.4$), recommendation uncertainty remains considerable even after consuming 3 times the budget required by Top-two Thompson sampling.

All best-arm identification algorithms require an initialization phase in order to output a well-defined recommendation. Successive Rejects needs to pull each arm at least once, while Top-two Thompson sampling and BayesGap need to pull each arm respectively 2 and 3 times (details in Section 4.4). For this reason, these algorithms' performance can only be evaluated after this initialization phase. BayesGap's performance is on par with Successive Rejects, except for the hardest setting we studied (i.e., $R_0 = 2.4$). In comparison, Top-two Thompson sampling consistently outperforms Successive Rejects 30 pulls after the initialization phase.

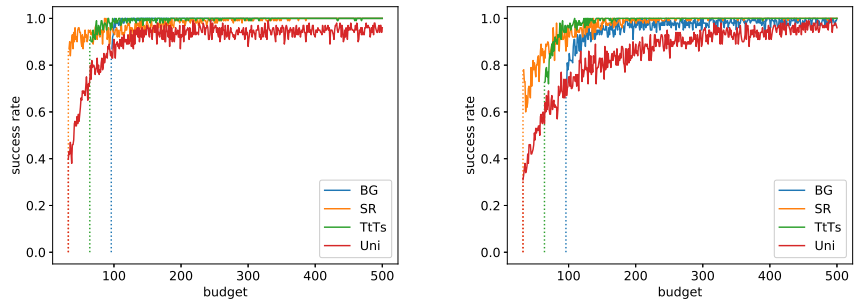


Figure 4.2: In this figure, we present the results for the experiment with $R_0 = 1.4$ (left panel) and $R_0 = 2.4$ (right panel). Each curve represents the rate of successful arm recommendations (y-axis) for a range of budgets (x-axis). A curve is shown for each of the considered algorithms: BayesGap (legend: BG), Successive Rejects (legend: SR), Top-two Thompson sampling (legend: TtTs) and Uniform sampling (legend: Uni).

Top-two Thompson sampling needs to initialize each arm's posterior with 2 pulls, i.e., double the amount of uniform sampling and Successive Rejects. However, our experiments clearly show that none of the other algorithms reach any acceptable recommendation rate using less than 64 pulls.

In Section 4 we derived a statistic to express the probability of success (P_s) concerning a recommendation made by Top-two Thompson sampling. We analyse this probability for all the Top-two Thompson sampling recommendations that were obtained in the experiment described above.

To provide some insights on how this statistic can be used to support policy makers, we show the P_s values of all Top-two Thompson sampling recommendations in Figure 4.3 for $R_0 = 1.4$ (left panel) and $R_0 = 2.4$ (right panel). This figure indicates that P_s closely follows recommendation correctness and that the uncertainty of P_s is inversely proportional to the size of the available budget. Figures for the other R_0 values are shown in Appendix 5.

Additionally, in Figure 4.4 we confirm that P_s underestimates recommendation correctness. Figures for the other R_0 values are shown in Appendix 6.

These observations show that P_s has the potential to serve as a conservative statistic to inform policy makers about the confidence of a particular recommendation, and thus can be used to define meaningful cutoffs to guide policy makers in their interpretation of the recommendation of preventive strategies.

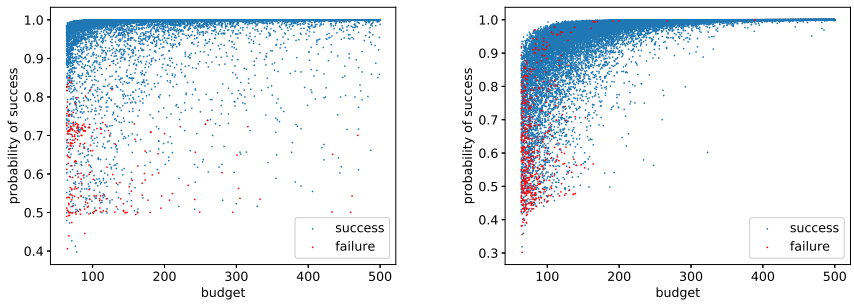


Figure 4.3: Top-two Thompson sampling was run 100 times for each budget for the experiment with $R_0 = 1.4$ (left panel) and $R_0 = 2.4$ (right panel). For each of the recommendations, P_s was computed. The P_s values are shown as a scatter plot, where each point's color reflects the correctness of the recommendation (see legend).

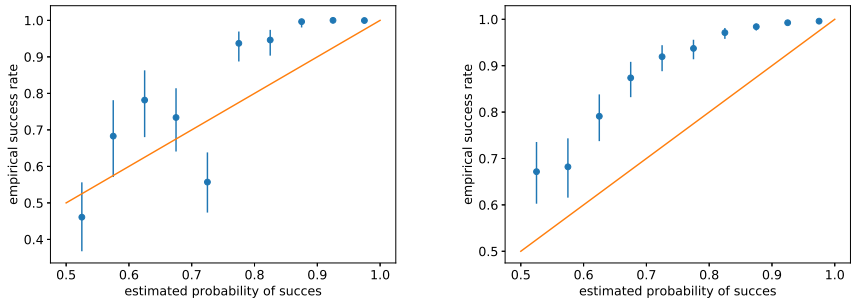


Figure 4.4: Top-two Thompson sampling was run 100 times for each budget for the experiment with $R_0 = 1.4$ (left panel) and $R_0 = 2.4$ (right panel). For each of the recommendations, P_s was computed. The P_s values were binned (i.e., 0.5 to 1 in steps of 0.05). Per bin, we thus have a set of Bernoulli trials, for which we show the empirical success rate (blue scatter) and the Clopper-Pearson confidence interval (blue confidence bounds). The orange reference line denotes perfect correlation between the empirical success rate and the estimated probability of success.

6 Future work

We identify four particular directions for future work.

Firstly, while our method is evaluated in the context of pandemic influenza, it is important to stress that it can be used to evaluate preventive strategies for other infectious diseases. Since recently, a Dengue vaccine is available [Hadinegoro et al., 2015], and the optimal allocation of this vaccine remains an important research topic [Aguiar and Stollenwerk, 2017], we recognize that Dengue epidemics are an interesting use case.

Secondly, in this paper, our preventive bandits only learn with respect to a single model outcome (i.e., the proportion of symptomatic infections). However, for many pathogens it is interesting to incorporate multiple objectives (e.g., morbidity, mortality, cost). We believe the use of *multi-objective multi-armed bandits* to be an interesting direction for future work [Roijers et al., 2013], which we explored in a bachelor project [3], to which I was advisor. In this bachelor project, we did a preliminary investigation to use the Interactive Thompson sampling algorithm [Roijers et al., 2017] in a pure-exploration setting, such that we can learn about the environment (i.e., the decision problem) and the user's preferences simultaneously.

Thirdly, we believe that bandits that can take into account delayed feedback would be of great interest. This could result in an algorithm that can evaluate different preventive strategies in parallel, alleviating the wait time between arm pulls. Recently, a delayed feedback algorithm was introduced in the best-arm identification setting [Grover et al., 2018]. An interesting direction for future work, would be to investigate a Bayesian variant of this new algorithm, and evaluate it in the context of epidemiological policy evaluation.

Fourthly, we believe it would be interesting to investigate generalized reward distributions to support a wider range of model outcomes. In this work, we focus on finding the prevention strategy that optimizes the attack rate, a normally distributed model outcome. To generalize this approach, consider that in an individual-based model we can model each individual's state transition as a Bernoulli experiment. We thus have a set of Bernoulli experiments (one for each individual) with dependent probabilities. To obtain summary statistics, we can compute the sum of this set of Bernoulli experiments, which will be distributed according to a Conway-Maxwell-binomial distribution [Kadane et al., 2016].

[3] Interactive Top Two Thompson Sampling for Multi-Objective Multi-Armed Bandits, Oktay Kavi, Vrije Universiteit Amsterdam (Defended, July 2019)

7 Discussion

We formulate the objective to select the best preventive strategy in an individual-based model as a fixed budget best-arm identification problem. We set up an experiment to evaluate this setting in the context of a realistic influenza pandemic. To assess the best arm recommendation performance of the preventive bandit, we report a success rate over 100 independent bandit runs.

We demonstrate that it is possible to efficiently identify the optimal preventive strategy using only a limited number of model evaluations, even if there is a large number of preventive strategies to consider. Compared to uniform sampling, our technique is able to recommend the best preventive strategy reducing the number of required model evaluations 2-to-3 times, when using Top-two Thompson sampling. Additionally, we defined a statistic to support policy makers with their decisions, based on the posterior information obtained during Top-two Thompson sampling. As such, we present a decision support tool to assist policy makers to mitigate epidemics. Our framework will enable the use of individual-based models in studies where it would otherwise be computationally too prohibitive, and allow researchers to explore a wider variety of model scenarios.

However, there are several limitations associated with the technique we presented in this chapter.

Firstly, simply returning the single best prevention strategy can be an impediment for public health scientists. On the one hand, this implies that public health scientists can only offer a take-it-or-leave-it option to government officials, rather than a set of options that can be evaluated within the government's political and legal framework. On the other hand, it can be beneficial to have a set of policy options available that optimize a particular basic requirement (e.g., attack rate), enabling experts to inspect this small set of alternatives in greater detail. As an example, a set of policies might all have a similar effect reducing the attack rate, but might differ with respect to logistic efforts. While such scenarios can be approached from a multi-objective perspective, this complicates the analysis, as the different objectives need to be weighed, and determining a fitting a weight vector can be challenging. Therefore, it is more practical to find the top set of prevention strategies, and investigate this set with respect to additional constraints. Moreover, from a health economics perspective, a set of optimal policies can be used to negotiate a fair cost with the producers of pharmaceutical supplies[4].

Secondly, in this chapter, we opt for an approach that assumes a fixed computational budget, that needs to be specified a priori. We motivate this choice by the fact that computational experiments need to be planned. We do however recognize that deciding

[4]Personal communication with Prof. Dr Philippe Beutels, Centre for Health Economics Research & Modelling Infectious Diseases, University of Antwerp

the budget upfront can be challenging. This is especially the case when computationally expensive models are used, for which it is difficult to make a trade-off between the available budget and desired confidence. We argue that an anytime bandit setting can overcome these limitations, as an initial budget can still be provided, but the budget can be extended when necessary.

In the next chapter, we will therefore study the recently introduced anytime m-top exploration problem [Jun and Nowak, 2016]. We will evaluate the state-of-the-art AT-LUCB algorithm [Jun and Nowak, 2016] in the context of decision making, including a generalized benchmark based on the experiments introduced in this chapter.

As a UCB-variant, AT-LUCB is not equipped to incorporate prior knowledge with respect to the reward distribution. Having shown throughout this chapter that incorporating such knowledge can greatly improve the learning performance, we present a new algorithm inspired by Thompson sampling to solve the anytime m-top exploration problem: Boundary Focused Thompson sampling (BFTS).

5 | Bayesian Anytime m-top Exploration

Tuurlijk! Tuurlijk! It's lonely, it's lonely at the top.
Hans Teeuwen, *Dat dan weer wel*

We introduce Boundary Focused Thompson sampling (BFTS), a new Bayesian algorithm to solve the anytime m-top exploration problem, where the objective is to identify the m best arms in a multi-armed bandit. We consider a set of existing benchmark problems and introduce two new environments, inspired by real world decision problems. The first new environment extends the pandemic influenza control problem from Chapter 4. To demonstrate that our method can be used in other complex decision domains, we introduce a second environment that considers an insect control decision problem for organic agriculture. This environment considers a Poisson bandit reward distribution, which is particularly hard to learn. For both the existing and newly introduced benchmarks, we experimentally show that BFTS consistently outperforms AT-LUCB, the current state-of-the-art algorithm. Finally, we analyse BFTS using Bayesian principles, to provide more insights in our algorithm's exploration strategy.

The work presented in this chapter was accepted at the ICTAI-2019 conference and is currently in press.

Libin, P., Verstraeten, T., Roijers, D. M., Wang, W., Theys, K., and Nowé, A., "Thompson sampling for m-top Exploration", International Conference on Tools with Artificial Intelligence, p. 1414-1420, 2019.

1 Rationale and objectives

The *multi-armed bandit* has K arms, as specified in Definition 1 (Chapter 2). To remind the reader, when an arm a_k is pulled, a reward r_k is drawn from that arm's reward distribution. For each arm a_k, we have the expected reward $\mu_k = \mathbb{E}[r_k]$. Our aim is to solve the m-top exploration problem ($m < K$), where the objective is to identify the m best arms, with respect to the expected reward μ_k of the arms [Bechhofer, 1958]. Formally, we have $\mu_1 \geq \ldots \geq \mu_m \geq \mu_{m+1} \geq \ldots \geq \mu_K$, and the objective is to identify the set $\{\mu_1, \ldots, \mu_m\}$.

Most commonly, the m-top exploration problem is studied in a fixed confidence or fixed budget setting. On the one hand, fixed confidence algorithms attempt to recommend the m best arms with probability $1-\delta$ using a minimal number of arm pulls, where δ is a failure probability that needs to be chosen up front [Gabillon et al., 2012; Even-Dar et al., 2002; Kalyanakrishnan et al., 2012; Karnin et al., 2013; Jamieson et al., 2014]. On the other hand, the goal for fixed budget algorithms is to recommend the top m arms, within a given budget of arm pulls [Gabillon et al., 2012; Audibert and Bubeck, 2010; Karnin et al., 2013; Bubeck et al., 2013; Hoffman et al., 2014]. Recently, a third setting was introduced, where the top m arms are to be recommended after every time step [Jun and Nowak, 2016]. This setting, referred to as anytime explore-m, is more challenging than the fixed confidence and fixed budget setting, but offers a more realistic framework [Jun and Nowak, 2016].

An example of an m-top exploration problem presented in [Jun and Nowak, 2016] is a crowd-sourcing task, i.e., the New Yorker cartoon caption contest [Jamieson et al., 2015]. In this application, the aim is to collect ratings for the captions submitted for each week's cartoon, and to identify the top-m captions at a requested time. In a crowd sourcing application, the sampling budget corresponds to the number of ratings that are obtained. Therefore, as this budget is unknown a priori, the fixed-budget setting cannot be used. Moreover, the fixed-confidence setting is not applicable either, as this setting requires that an unlimited stream of samples is available until a certain confidence threshold has been reached. The crowd sourcing application is thus a natural fit for the anytime explore-m problem.

Apart from this example, there is a great potential for the anytime m-top exploration bandit to support decision makers with complex societal challenges, such as epidemics of infectious diseases, as presented in the previous chapter. These decisions making processes are often guided by intricate simulation models, to evaluate a set of alternative policies

that can be modelled as bandit arms. By formulating the decision problem as an m-top exploration bandit problem, a learning agent can select the m policies for which it expects the highest utility, enabling the experts to inspect this small set of alternatives. The anytime component provides the decision makers with flexibility regarding the time at which a decision is made. This is especially important when computationally intensive models are used, for which it is difficult to make a trade-off between the available budget and desired confidence.

In addition to introducing the m-top exploration problem, a new algorithm is presented in [Jun and Nowak, 2016]: AnyTime Lower and Upper Confidence Bound (AT-LUCB). This algorithm remains the state-of-the-art up until today. We discuss the algorithmic details of AT-LUCB in Section 2.

While UCB algorithms, such as AT-LUCB, permit specifying tight theoretical bounds, algorithms based on Thompson Sampling (TS) typically perform better in practice (see Section 2.4 in Chapter 2). Furthermore, TS works for any type of reward distribution, and permits the inclusion of any form of prior knowledge. This is important, as prior knowledge can be specified for many practical settings, even if it is only in the form of basic common knowledge or even intuitions, and can greatly help to improve sample-efficiency. Therefore, we investigate the potential of TS for the m-top exploration problem, and propose the first Bayesian algorithm for this setting: Boundary Focused Thompson Sampling (BFTS). BFTS is a non-parametric algorithm that focuses its exploration on the problem's decision boundary, i.e., the m^{th} and $m+1^{\text{th}}$ arm.

We empirically compare the performance of BFTS to AT-LUCB. First, we evaluate the set of benchmarks settings introduced in [Jun and Nowak, 2016], which consists out of an artificial environment (i.e., a bandit with fixed-variance Gaussian reward distributions) and a bandit that models the New York cartoon crowd sourcing task introduced earlier. Next, motivated by the experiments introduced in Chapter 4, we introduce the pandemic bandit, where the objective is to select the most promising prevention strategies in the context of pandemic influenza, using scaled Gaussian reward distributions. Furthermore, to demonstrate that BFTS has potential to be used in other complex decision domains, we introduce the organic bandit, where we aim to maximize the prevalence of certain insect species on farmland to support organic agriculture [Soulsby and Thomas, 2012]. We model this setting using Poisson reward distributions [Soulsby and Thomas, 2012]. This is a particularly hard problem, as for Poisson distributions the variance is equal to the mean and subsequently there is a large variance among the top arms, complicating the m-top exploration. We show that BFTS consistently outperforms AT-LUCB for all of the investigated environments, with a particularly large improvement in performance on the pandemic and organic bandit.

In Section 6, we perform a Bayesian analysis of BFTS. While this analysis does not result in a bound on the simple regret, it does provide additional insight in BFTS' exploration strategy and confirms that this strategy is well-grounded.

2 Background: AT-LUCB

AT-LUCB repeatedly invokes the fixed-confidence LUCB algorithm [Kalyanakrishnan et al., 2012], with a decaying failure parameter, $\delta_s = \delta_1 \alpha^{s-1}$, for each LUCB stage s, where δ_1 and α are parameters of the AT-LUCB algorithm. At each time step t, AT-LUCB returns the empirical m-top arms.

To provide more insight in AT-LUCB's exploration strategy, we discuss details on AT-LUCB's exploration bound [Jun and Nowak, 2016]. Note that this bound was constructed following the assumption that reward distributions are sub-Gaussian with means in the interval $[0, 1]$.

At each stage, LUCB depends on upper confidence bound $U_k^{(t)}$ and lower confidence bound $L_k^{(t)}$, where:

$$\begin{aligned} U_k^{(t)}(\delta_s) &= \hat{\mu}_k^{(t)} + \kappa(n_k^{(t)}, t, \delta_s) \\ L_k^{(t)}(\delta_s) &= \hat{\mu}_k^{(t)} - \kappa(n_k^{(t)}, t, \delta_s), \end{aligned} \tag{5.1}$$

with,

$$\kappa(n_k^{(t)}, t, \delta_s) := \sqrt{\frac{1}{2n_k^{(t)}} \ln\left(\frac{5}{4}\frac{K \cdot t^4}{\delta_s}\right)}, \tag{5.2}$$

where $\hat{\mu}_k^{(t)}$ is the empirical mean for arm a_k at time t, K is the number of arms, $n_k^{(t)}$ is the amount of times arm a_k was pulled at time t and δ_s is the confidence parameter at stage s.

From this confidence bound definition, it is clear that the empirical mean is the only reward distribution statistic used by AT-LUCB. We expect that such a confidence bound will be sub-optimal with respect to reward distributions with complex higher-order statistics, such as skewness or high variance. In Section 5, we demonstrate that this is the case in our experiments with the organic bandit, with Poisson distributed rewards.

3 Related work

The anytime explore-m setting is a generalization of the anytime best-arm identification setting [Bubeck et al., 2009]. As introduced earlier, this setting is related to the fixed

confidence [Gabillon et al., 2012; Even-Dar et al., 2002; Kalyanakrishnan et al., 2012; Karnin et al., 2013; Jamieson et al., 2014] and fixed budget [Gabillon et al., 2012; Audibert and Bubeck, 2010; Karnin et al., 2013; Bubeck et al., 2013; Hoffman et al., 2014] explore-m algorithms.

As we stated in Section 1, the anytime explore-m setting was only recently introduced. To our best knowledge, the AT-LUCB algorithm remains the state-of-the-art algorithm. In [Jun and Nowak, 2016], another algorithm called DSAR is presented in addition to AT-LUCB. DSAR repeatedly invokes the fixed budget m-top algorithm Successive Accept and Reject (SAR) [Bubeck et al., 2013] where the budget is doubled upon each invocation. It is experimentally shown in [Jun and Nowak, 2016] that AT-LUCB consistently outperforms DSAR, and DSAR is deemed unsuitable for anytime purposes due to fluctuations in its performance (i.e., stagnation or even decrease) when the algorithm changes from one stage to the next. We therefore did not consider the DSAR algorithm in our experiments.

Bayesian exploration methods have been used in the context of best-arm identification, i.a., BayesGap, Top-Two Thompson sampling, Ordered statistic Thompson sampling, and the Top-Two Expected Improvement algorithm. BayesGap is a gap-based Bayesian algorithm [Hoffman et al., 2014] and requires that for each arm, a high-probability upper and lower bound is defined on the posterior of the arms' means at each time step t. These bounds are used to establish a gap quantity that the algorithm attempts to minimize. Top-Two Thompson sampling [Russo, 2016] uses a variant of TS that adds a re-sampling step in order to increase exploration. Ordered statistic Thompson sampling [Mellor, 2014] ranks the samples from TS and pulls any arm randomly according to a rank distribution to add extra exploration. The Top-Two Expected Improvement algorithm enhances the Expected Improvement algorithm, by randomizing which of the two top arms to sample [Qin et al., 2017].

4 Boundary Focused TS

In this section, we propose our anytime m-top algorithm Boundary Focused Thompson sampling (BFTS). The purpose of the algorithm is to recommend the top m arms at each time step. In other words, the algorithm would perform perfectly if it recommends the true top m arms at each time step.

Consider a stochastic multi-armed bandit for which our prior belief over the means is given by a distribution $\pi(.)$. Inspired by TS, at each time step t we sample an estimate $\tilde{\boldsymbol{\mu}}^{(t)}$ for the means $\mu_{1..K}$ from $\pi(\cdot \mid \mathcal{H}^{(t-1)})$, i.e., the posterior over the means, given by $\pi(.)$ conditioned on the history of arm pulls and observed rewards $\mathcal{H}^{(t-1)}$ (Definition 5). Consequently, we order the samples that comprise $\tilde{\boldsymbol{\mu}}^{(t)}$, and use the ranking operator (Definition 6), $\Psi_\rho(\tilde{\boldsymbol{\mu}}^{(t)})$, to denote the ρ ordered arm. In the case of vanilla TS [Thompson,

1933], where the objective is to minimize cumulative regret, we would always play top arm $\Psi_1(\tilde{\boldsymbol{\mu}}^{(t)})$. However, for the anytime m-top bandit problem, where the objective is to return the top m arms at any time[1], we need to focus the exploration on the decision boundary, to decrease the uncertainty about arm $a_m^{(t)}$ and $a_{m+1}^{(t)}$. We focus on *both sides* of the decision boundary, as in a pure exploration setting, it is equally important to gain information about the arms with the potential to be optimal and sub-optimal.

To implement the intuition of focussing on the decision boundary, at each time step t we play the arm ordered $\Psi_m(\tilde{\boldsymbol{\mu}}^{(t)})$ or $\Psi_{m+1}(\tilde{\boldsymbol{\mu}}^{(t)})$ with equal probability. To do this, we use a Bernoulli experiment, as formalized in Algorithm 6. The reward $r^{(t)}$ of the played arm $a^{(t)}$ is observed and used to update the history $\mathcal{H}^{(t-1)}$. At the end of each step, we recommend the m-top arms based on the current belief over the bandit posterior $\pi(\cdot \mid \mathcal{H}^{(t-1)})$.

Given: $\pi(.)$ and $\mathcal{H}^{(0)} = \emptyset$
for $t = 1, \ldots, +\infty$ **do**
  $\tilde{\boldsymbol{\mu}}^{(t)} \sim \pi(\cdot \mid \mathcal{H}^{(t-1)})$
  $b \sim \mathcal{B}\text{er}(0.5)$
  $a^{(t)} = \Psi_{m+b}(\tilde{\boldsymbol{\mu}}^{(t)})$
  $r^{(t)} \leftarrow$ Pull arm $a^{(t)}$
  $\mathcal{H}^{(t)} \leftarrow \mathcal{H}^{(t-1)} \cup \{a^{(t)}, r^{(t)}\}$
  Recommend top arms based on $\pi(\cdot \mid \mathcal{H}^{(t)})$
end

Algorithm 6: Boundary Focused TS

An important observation with respect to BFTS is that the exploration is guided by sampling from the posterior, while balancing between $\Psi_m(\tilde{\boldsymbol{\mu}}^{(t)})$ and $\Psi_{m+1}(\tilde{\boldsymbol{\mu}}^{(t)})$, i.e., our belief of the decision boundary at time t. As the posterior reflects the uncertainty with respect to the bandit problem, sampling the m^{th} or $m+1^{\text{th}}$ ordered arm will initially explore all arms, when an uninformative prior is chosen. However, as the uncertainty of the outer extreme arms is reduced, BFTS will increase its focus on the arms near the decision boundary. In Figure 5.1, we visualize this process for a simple bandit setting ($K = 6$ and $m = 3$) with Gaussian posteriors.

BFTS is thus convenient for real-world applications, as its belief-based exploration can be intuitively understood and informed by its users, without the need to specify any exploration parameters that are typically hard to choose in advance. Moreover, the anytime aspect of BFTS removes the need to decide on the computational budget or desired confidence before starting the analysis.

[1]The top m arms should be recommended, but they are not expected to be ranked.

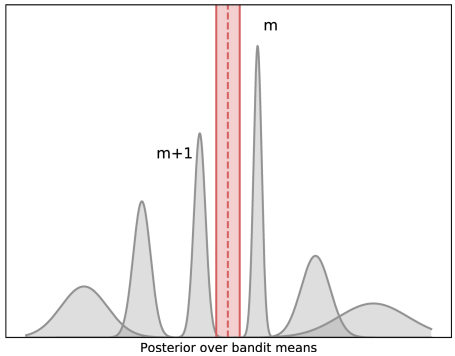


Figure 5.1: Posteriors for an artificial bandit ($K = 6, m = 3$) (gray) and BFTS' decision boundary with confidence bounds to demonstrate its uncertainty (red).

Furthermore, as the algorithm description shows, the exploration process only depends on the balancing between the m^{th} or $m + 1^{\text{th}}$ ordered arm and the choice of the prior. In Section 6 we show via a formal analysis that this exploration strategy is well-grounded.

5 Experiments

We compare the performance of BFTS to the current state-of-the-art algorithm, i.e., AT-LUCB, and uniform sampling as a baseline. AT-LUCB operates as described in Section 2, and we choose the same parameters as in [Jun and Nowak, 2016]. Uniform sampling pulls at each time step t the arm that was least sampled in the previous time steps and recommends the empirical m-top arms.

For BFTS, we recommend the m-top arms with the posterior expectation of the arms:

$$\begin{aligned} &\mu_k \sim \pi(\cdot \mid \mathcal{H}^{(t)})_k \\ &\mathbb{E}\left[\mu_k\right]. \end{aligned} \tag{5.3}$$

The use of the posterior expectation is well-grounded in our experiments, as all priors we use tend to a bell-shaped posterior, for which the expectation is a natural summary statistic.

To perform a fair and unbiased evaluation we commence with the experimental environments introduced in [Jun and Nowak, 2016]. Then, we propose two new environments, i.e., the pandemic bandit and the organic bandit. Both settings consider interesting reward distributions: the first setting has scaled Gaussian reward distributions for which the variances are not known and the second setting has Poisson reward distributions.

AT-LUCB expects sub-Gaussian reward distributions with means in the interval $[0, 1]$. We demonstrate experimentally, using the organic bandit environment with Poisson reward distributions, that AT-LUCB indeed performs poorly when this assumption is not met.

The probability of success, i.e., the probability that all of the true best arms are recommended, does not yield a useful comparison in our experiments as the considered environments are hard and it takes a large amount of samples to find the true m top arms [Jun and Nowak, 2016]. Therefore, we evaluate the algorithms' performance using two proxy statistics instead: the sum of the means of the m top arms at time t, as introduced in [Jun and Nowak, 2016],

$$\sum_{a \in \mathcal{J}^{(t)}} \mu_a, \tag{5.4}$$

and the proportion of correctly recommended arms at time t,

$$\frac{|\mathcal{J}^{(t)} \cap \mathcal{J}^{\mathsf{True}}|}{m}, \tag{5.5}$$

where $\mathcal{J}^{(t)}$ is the set of recommended arms at time t and $\mathcal{J}^{\mathsf{True}}$ is the true set of optimal arms.

All of the algorithms were run 100 times for each of the stochastic bandit environments, as such, the average of the statistics over these runs is reported. In order to justify this number of replicates, all figures include the variance of the reported statistic, which is visualized using a lighter bound around the mean curve. In every run, each algorithm was allowed to consume 15×10^4 samples (i.e., arm pulls), a sufficient amount to discern a clear learning curve. Note that for BFTS and uniform sampling only one sample per time step is obtained, while for AT-LUCB two samples per time step are used. Therefore, all figures report their results in terms of the number of samples, to allow for a fair comparison.

For all BFTS experiments we consistently use Jeffreys' priors. Such priors are considered non-informative and objective, such that when data is observed, the posteriors are not influenced by the prior's hyper-parameters [Jaynes, 1968].

5.1 Gaussian bandit with fixed variance

The first set of benchmark environments introduced in [Jun and Nowak, 2016] concerns Gaussian reward distributions with fixed variance $\sigma^2 = 0.25$ and means in the interval $[0, 1]$. The environment defines a bandit with 1000 arms.

The benchmark includes two instances, one where the gap between means is increased linearly (Equation 5.6) and one where the gap is increased polynomially (Equation 5.7).

$$\forall k : \mu_k = .9(\frac{n-i}{n-1}) \tag{5.6}$$

$$\mu_1 = .9, \forall k \geq 2 : \mu_k = .9(1 - \sqrt{i/n}) \tag{5.7}$$

In this environment, as each arm a_k has a reward distribution $\mathcal{N}(\mu, \sigma^2)$ with known variance, we have a conjugate prior for the mean that is Gaussian with hyper-parameters μ_0 and σ_0^2. As the means are in $[0, 1]$, we choose this Gaussian prior to be truncated on said interval. We consider $\sigma_0^2 \to \infty$, which results in a uniform prior over μ:

$$\lim_{\sigma_0 \to +\infty} \mathcal{N}_{[0,1]}(\mu \mid \mu_0, \sigma_0^2). \tag{5.8}$$

To evaluate this limit, we consider that a truncated normal distribution can be expressed in terms of its probability density function $g(\mu)$:

$$\begin{aligned}
\mathcal{N}_{[0,1]}(\mu \mid \mu_0, \sigma_0^2) =& \frac{g(\mu)}{\int_{-\infty}^{1} g(\mu')d\mu' - \int_{-\infty}^{0} g(\mu')d\mu'} \\
=& \frac{g(\mu)}{\int_{0}^{1} g(\mu')d\mu'} \\
\overset{(1)}{=}& \frac{\exp\left(-\frac{(\mu-\mu_0)^2}{2{\sigma_0}^2}\right)}{\int_{0}^{1} \exp\left(-\frac{(\mu'-\mu_0)^2}{2{\sigma_0}^2}\right) d\mu'} \\
\overset{(2)}{=}& \frac{1}{\exp\left(\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \int_{0}^{1} \exp\left(-\frac{(\mu'-\mu_0)^2}{2{\sigma_0}^2}\right) d\mu'} \\
\overset{(3)}{=}& \frac{1}{\int_{0}^{1} \exp\left(\frac{(\mu-\mu_0)^2}{2\sigma_0^2}\right) \exp\left(-\frac{(\mu'-\mu_0)^2}{2{\sigma_0}^2}\right) d\mu'} \\
=& \frac{1}{\int_{0}^{1} \exp\left(\frac{(\mu-\mu_0)^2}{2\sigma_0^2} - \frac{(\mu'-\mu_0)^2}{2\sigma_0^2}\right) d\mu'}
\end{aligned} \tag{5.9}$$

First (1), we fill in the probability density function for $g(x)$ and resolve the constant $\frac{1}{\sqrt{2\pi\sigma_0^2}}$ in the nominator and denominator. Next (2), we move the exponential in the nominator to the denominator. Finally, (3) we move the constant inside of the integral.

Based on this derivation, we can solve the limit:

$$\begin{aligned}\lim_{\sigma_0 \to +\infty} \mathcal{N}_{[0,1]}(\mu \mid \mu_0, \sigma_0^2) &= \lim_{\sigma_0 \to +\infty} \frac{1}{\int_0^1 \exp\left(\frac{(\mu-\mu_0)^2-(\mu'-\mu_0)^2}{2\sigma_0^2}\right) d\mu'} \\ &= 1\end{aligned} \tag{5.10}$$

This uniform prior corresponds to the Jeffreys prior [Robert, 2007].
Considering the non-truncated Gaussian posterior:

$$\mu \sim \mathcal{N}\left(\frac{\sigma^2}{n}\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n r_i}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)\right), \tag{5.11}$$

and the rewards $\mathbf{r} = \langle r_1, ..., r_n \rangle$ we have the truncated Gaussian posterior:

$$\mu \sim \mathcal{N}_{[0,1]}(\mu_n, \sigma_n^2), \tag{5.12}$$

with,

$$\begin{aligned}\mu_n &= \lim_{\sigma_0 \to +\infty} \frac{\sigma^2}{n}\left(\frac{\mu_0}{\sigma_0^2} + \frac{\sum_{i=1}^n r_i}{\sigma^2}\right) = \frac{\sum_{i=1}^n r_i}{n} \\ \sigma_n^2 &= \lim_{\sigma_0 \to +\infty} \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right) = \frac{\sigma^2}{n},\end{aligned} \quad . \tag{5.13}$$

The expectation of the posterior over μ, that is required for recommending the m top arms, is the mean of the truncated Gaussian in Equation 5.12.

As in [Jun and Nowak, 2016], we perform the experiment with $m = 10$ and $m = 50$, for both the linear and polynomial environment. We present the results for the linear bandit in Figure 2 ($m = 10$) and Figure 4 ($m = 50$). We present the results for the polynomial bandit in Figure 3 ($m = 10$) and Figure 5 ($m = 50$). In general, BFTS needs a short burn-in period to meet AT-LUCB's performance for both statistics, but then consistently outperforms AT-LUCB, most apparently with respect to the proportion of success' learning curve. On the one hand, for the linear Gaussian environment with $m = 10$, it takes BFTS the most time to meet AT-LUCB's performance. On the other hand, for the linear bandit with $m = 50$, BFTS takes the least iterations to meet the performance of AT-LUCB.

5.2 Cartoon caption bandit

The second benchmark environment introduced in [Jun and Nowak, 2016] concerns the New York cartoon caption contest we described in Section 1. This benchmark simulates the caption contest by setting up a bandit with 496 arms, where each arm follows a

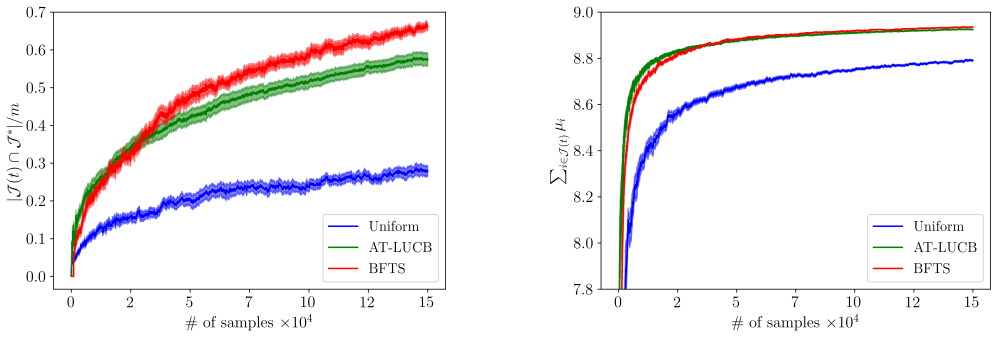


Figure 5.2: Results for the linear Gaussian benchmark with fixed variance ($m = 10$).

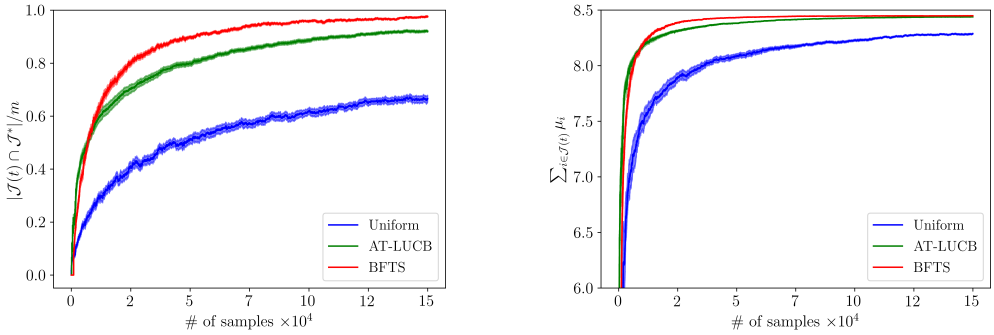


Figure 5.3: Results for the polynomial Gaussian benchmark with fixed variance ($m = 10$).

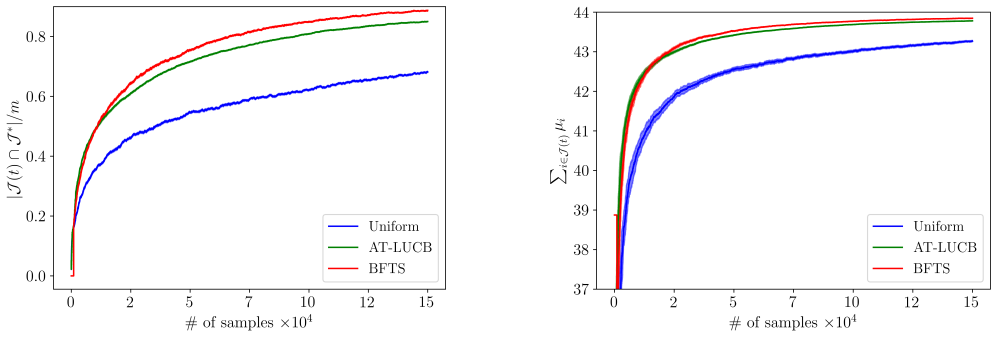


Figure 5.4: Results for the linear Gaussian benchmark with fixed variance ($m = 50$).

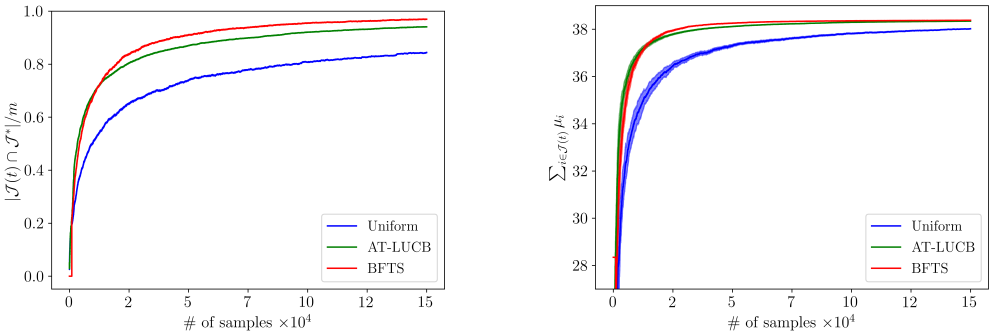


Figure 5.5: Results for the polynomial Gaussian benchmark with fixed variance ($m = 50$).

categorical distribution $\mathcal{C}\text{at}_{\mathbf{c}}(\mathbf{p})$ on three categories $\mathbf{c} = [0, 0.5, 1]$. The distribution is parametrized with an event probability vector $\mathbf{p}$. For each arm, $\mathbf{p}$ is determined using maximum likelihood estimation on the dataset used in [Jun and Nowak, 2016].

For a categorical distribution $\mathcal{C}\text{at}_{\mathbf{c}}(\mathbf{p})$, the conjugate prior is a Dirichlet distribution $\mathcal{D}\text{ir}_{\mathbf{c}}(\boldsymbol{\alpha_0})$ with prior parameter $\boldsymbol{\alpha_0}$. Given rewards $\mathbf{r} = \langle r_1, ..., r_n \rangle$, we have posterior

$$\boldsymbol{\mu} \sim \mathbf{c} \cdot \mathcal{D}\text{ir}_{\mathbf{c}}(\boldsymbol{\alpha_0} + \mathbf{f}) \tag{5.14}$$

where $\mathbf{f}$ is a vector of frequencies at which the categories occur in $\mathbf{r}$. Note that this is a proper posterior if all elements in $\boldsymbol{\alpha_0}$ are greater than zero. For the experiment we use an uninformative Jeffreys prior $\boldsymbol{\alpha_0} = \langle .5, .5, .5 \rangle$ [Tuyl, 2017]. We report the expectation over μ with respect to the Dirichlet posterior:

$$\mathbb{E}\big[\mathbf{c} \cdot \mathcal{D}\text{ir}_{\mathbf{c}}(\boldsymbol{\alpha_0} + \mathbf{f})\big]. \tag{5.15}$$

As in [Jun and Nowak, 2016], we run the caption contest bandit experiment for $m = 50$. We present the results for this experiment in Figure 5.6. BFTS needs a short burn-in to meet AT-LUCB's performance for both statistics, but then consistently outperforms AT-LUCB, most significantly with respect to the proportion of success' learning curve.

5.3 Pandemic bandit

In Chapter 4, we used best-arm identification to perform the evaluation of preventive strategies with the objective to curb epidemics (e.g., school closures, vaccine allocation). As a reminder to the reader, in this setting, we have a bandit where each arm corresponds to a prevention strategy, and when the arm is pulled, this policy is evaluated using a

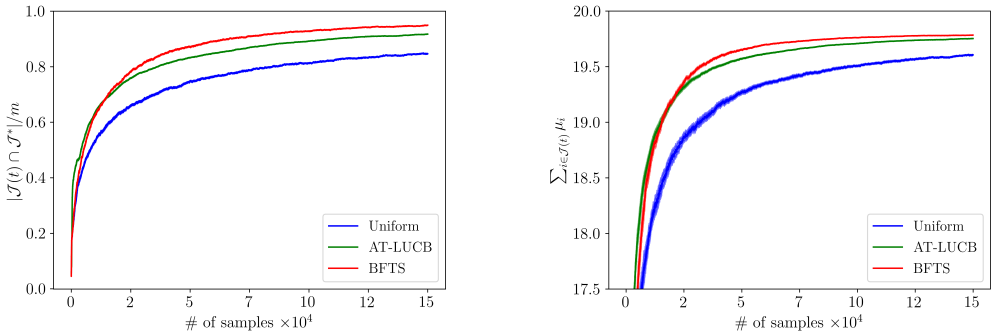


Figure 5.6: Results for the cartoon caption benchmark

stochastic simulator. When the arm is pulled, the outcome of the simulator is returned as reward. This implies that the arm's reward distribution corresponds to the outcome distribution of the policy that is associated with that arm. As we established in Chapter 4, this outcome distribution is Gaussian and the means are distributed in a sub-interval of $[0, 1]$. The range of this sub-interval is unknown, as it depends on the simulated scenario. Additionally, we know that the variance of the outcome distribution differs per prevention strategy and is unknown.

Generalizing the experiments that we introduced in Chapter 4, i.e., we set up a new benchmark, which we denote the pandemic bandit environment, where the means and variances are uniformly sampled from respectively $\mathcal{U}(0.3, 0.4)$ and $\mathcal{U}(1.06 \cdot 10^{-6}, 3.6 \cdot 10^{-2})$. The environment defines a bandit with 1000 arms.

As each arm has a Gaussian reward distribution with unknown variance, we assume an uninformative Jeffreys prior $(\sigma)^{-3}$ on (μ, σ^2) [Honda and Takemura, 2014]. Given rewards $\mathbf{r} = \langle r_1, ..., r_n \rangle$, this prior leads to the non-standardized t-distributed posterior. We truncate this prior on $[0, 1]$, given that we know that the arms' means are in this interval:

$$\mu \sim \mathcal{T}_{n,[0,1]} \left(\mu_n = \hat{\mu}, \sigma_n^2 = \frac{\sum_{i=1}^{n} (r_i - \hat{\mu})^2}{n^2} \right). \tag{5.16}$$

This posterior needs to be initialized two times for it to be proper. The expectation of the posterior over μ is derived in the Appendix 7.

We present the results for the pandemic bandit for $m = 10$, in Figure 5.7. The BFTS algorithm needs two rounds of initialization per arm, but after this initialization phase, its performance surpasses AT-LUCB quickly.

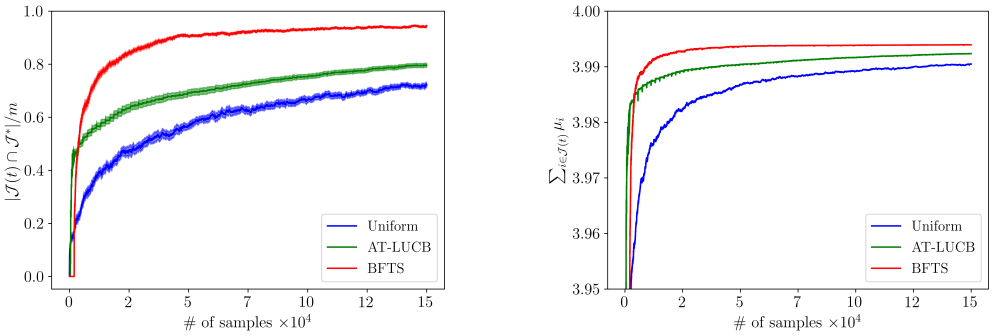


Figure 5.7: Results for the pandemic bandit benchmark

5.4 Organic bandit

Finally, we present a benchmark environment motivated by a research question that stems from organic agriculture, i.e., to investigate strategies that maximize the prevalence of certain insect species on farmland [Soulsby and Thomas, 2012],

As this prevalence distribution follows a Poisson distribution [Soulsby and Thomas, 2012], we construct a benchmark environment with Poisson reward distributions. We increase the means linearly:

$$\mu_k = \mu_{\min} + \frac{k \cdot (\mu_{\max} - \mu_{\min})}{K - 1}, \tag{5.17}$$

for $\mu_{\min} = 0.5$ and $\mu_{\max} = 5$. The environment defines a bandit with 1000 arms.

As mentioned in the Section 1, this is a particularly challenging benchmark, as for a Poisson distribution, the variance equals the mean, which complicates the m-top exploration process.

For a Poisson distribution, the conjugate Jeffreys prior is a gamma distribution [Lunn et al., 2012]:

$$\mathcal{G}\text{amma}(\alpha_0 = 0.5, \beta_0 = 0). \tag{5.18}$$

Given rewards $\mathbf{r} = \langle r_1, ..., r_n \rangle$, this leads to posterior

$$\mu \sim \mathcal{G}\text{amma}(\alpha_0 + \sum_{i=1}^{n} r_i, \beta_0 + n). \tag{5.19}$$

As $\beta_0 = 0$, this posterior needs to be initialized one time for it be proper.

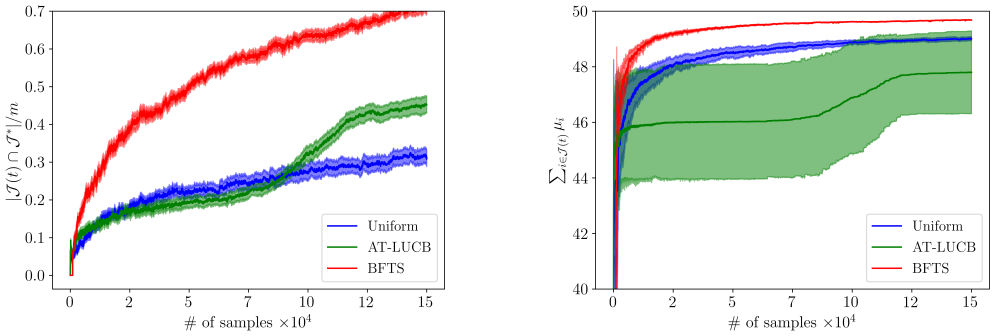


Figure 5.8: Results for the organic bandit benchmark

We present the results for the organic bandit for $m = 10$, in Figure 5.8. It is clear that AT-LUCB's performance grows very slowly and is similar to random sampling, while BFTS exhibits a much steeper learning curve. We further discuss these results in Section 11.

5.5 Overall performance of BFTS

In our experiments, BFTS consistently outperforms AT-LUCB, for all reported statistics. We do identify that BFTS needs an initialization period to meet AT-LUCB's performance, but we do not deem the lower performance during the first iterations of the algorithm problematic, as at these times both algorithms perform poorly and a fair amount of exploration is required to improve this.

Interestingly, BFTS also exhibits a significant performance improvement compared to AT-LUCB for the new settings we introduce. These additional experiments show that AT-LUCB struggles with the organic bandit, while BFTS performs much better. This demonstrates that BFTS has a great potential to be used with reward distributions that are not sub-Gaussian and non-symmetric. This is an important result, as we are unaware of any algorithms able to solve such problems efficiently.

BFTS outperforms AT-LUCB for both of the reported statistics. For the sum of means, a proxy for the simple regret, the difference is most evident during the earlier time steps of the experiments, as the difference in performance becomes less clear when the sum of means for both BFTS and AT-LUCB converge to a similar value. Supported by this observation, we argue that the number of correctly recommend arms (i.e., proportion of success) is a better proxy for the probability of success, i.e., the quantity that we actually attempt to optimize. Given this statistic, it is immediately clear how many mistakes an algorithm makes at a certain time, and BFTS' superior performance is even more evident.

6 Bayesian analysis of BFTS

In this section, we perform a Bayesian analysis of BFTS. We identify two heuristics that underlie BFTS's exploration strategy and relate these heuristics to the probability of error. We then experimentally demonstrate that these heuristics hold for all of the environments we consider in Section 5. While this analysis does not result in a bound on the probability of error, it provides a motivation for BFTS's exploration strategy.

In this Bayesian framework, we reason about the full distribution over bandits. Consequently, the actual means $\boldsymbol{\mu}$ are unknown, and we assert our belief over $\boldsymbol{\mu}$, given a posterior (Definition 5):

$$\pi(\cdot \mid \mathcal{H}^{(t-1)}), \tag{5.20}$$

i.e., the prior belief over the means $\pi(.)$ conditioned on the observed history

$$\mathcal{H}^{(t-1)} = \left\{a^{(i)}, r^{(i)}\right\}_{i=1}^{t-1} \tag{5.21}$$

at time t.

We remind the reader that we defined $\Psi_\rho(.)$ to be the ρ ordered arm (Definition 6). We introduce the random variables A_ρ^π as the ρ-ranked arms according to the prior belief, and A_ρ^{TS} as the ρ-ranked arm according to TS:

$$\begin{aligned} A_\rho^\pi &= \Psi_\rho(\boldsymbol{\mu}) \\ A_\rho^{TS} &= \Psi_\rho(\tilde{\boldsymbol{\mu}}^{(t)}) \end{aligned} \tag{5.22}$$

TS is a *probability matching* algorithm (Definition 5.23) [Agrawal and Goyal, 2012; Russo and Van Roy, 2016] and therefore it samples directly from the belief asserted in Equation 5.20.

Definition 20: Probability matching

Thompson sampling is a probability matching algorithm:

$$P(A_\rho^{TS} = \cdot \mid \mathcal{H}^{(t-1)}) = P(A_\rho^\pi = \cdot \mid \mathcal{H}^{(t-1)}) \tag{5.23}$$

We introduce $\rho^+ \in [1..m]$ and $\rho^- \in [m+1..K]$. Using this notation, we can express the true optimal arm set $\mathcal{J}^*$ and recommended arm set $\mathcal{J}^{TS}$ as:

$$\begin{aligned} \mathcal{J}^* &= \{A_{\rho^+}^\pi \mid \forall \rho^+\} \\ \mathcal{J}^{TS} &= \{A_{\rho^+}^{TS} \mid \forall \rho^+\}, \end{aligned} \tag{5.24}$$

we refer to $\overline{\mathcal{J}^*}$ as the complement of $\mathcal{J}^*$, i.e., the set of all arms excluding $\mathcal{J}^*$. Note that, both $\mathcal{J}^*$ and $\mathcal{J}^{TS}$ are random variables, as they are expressed as a union of random variables. Given this framework, we identify two heuristics that underlie BFTS' sampling strategy.

For the remainder of this Section, we use $P_t(.)$ to denote a probability that is conditioned on the observed history $\mathcal{H}^{(t-1)}$ at time t:

$$P_t(.) = P(. \mid \mathcal{H}^{(t-1)}) \tag{5.25}$$

Heuristic 1

The expectation that BFTS wrongly ranks an arm that is believed to be optimal, is bounded by the probability that BFTS wrongly ranks the arm on the sub-optimal side of the decision boundary:

$$\mathop{\mathbb{E}}_{\rho^-}[P_t(A^{TS}_{\rho^-} \in \mathcal{J}^*)] \leq P_t(A^{TS}_{m+1} \in \mathcal{J}^*) \tag{5.26}$$

Given this inequality, we expect that sampling the $m+1$-th arm will reduce

$$\mathop{\mathbb{E}}_{\rho^-}[P_t(A^{TS}_{\rho^-} \in \mathcal{J}^*)]. \tag{5.27}$$

Heuristic 2

The expectation that BFTS wrongly ranks an arm that is believed to be sub-optimal, is bounded by the probability that BFTS wrongly ranks the arm on the optimal side of the decision boundary.

$$\mathop{\mathbb{E}}_{\rho^+}[P_t(A^{TS}_{\rho^+} \in \overline{\mathcal{J}^*})] \leq P_t(A^{TS}_{m} \in \overline{\mathcal{J}^*}) \tag{5.28}$$

Given this inequality, we expect that sampling the m-th arm will reduce

$$\mathop{\mathbb{E}}_{\rho^+}[P_t(A^{TS}_{\rho^+} \in \overline{\mathcal{J}^*})]. \tag{5.29}$$

These heuristics stem from the fact that it is counter-intuitive for TS to often order an arm as optimal when it is *believed* to be sub-optimal. However, due to the stochastic nature of both the bandit and TS, it is possible to end up with a posterior for which the heuristics do not hold. Notwithstanding, we argue that, given the intuition behind probability matching, such events become unlikely when reasonable priors are chosen and

BFTS' exploration strategy is followed. We show this empirically in Section 6.1 for a diverse set of environments and their corresponding posteriors.

Given these heuristics, we expect that sampling A^{TS}_{m+1} decreases $P_t(A^{TS}_{m+1} \in \mathcal{J}^*)$ and, as a consequence, the expectation $\mathbb{E}_{\rho^-}[P_t(A^{TS}_{\rho^-} \in \mathcal{J}^*)]$. Equivalently, the second heuristic expresses that sampling A^{TS}_m decreases $P_t(A^{TS}_m \in \overline{\mathcal{J}^*})$ and, as a consequence, the expectation $E_{\rho^+}[P_t(A^{TS}_{\rho^+} \in \overline{\mathcal{J}^*})]$. We now show how the expectations in the heuristics relate to the probability of error. As such, given the heuristics, we can bound the probability of error with respect to both sides of the decision boundary (i.e., A^{TS}_{m+1} and A^{TS}_m), demonstrating that BFTS' exploration strategy is well-grounded.

First, we derive the bound in terms of A^{TS}_{m+1}:

$$
\begin{aligned}
P_t\left(\mathcal{J}^* \neq \mathcal{J}^{TS}\right) &= P_t\left(\bigvee_{\rho^-} A^{TS}_{\rho^-} \in \mathcal{J}^*\right) \\
&\leq \sum_{\rho^-} P_t\left(A^{TS}_{\rho^-} \in \mathcal{J}^*\right) \\
&= \frac{\sum_{\rho^-} P_t\left(A^{TS}_{\rho^-} \in \mathcal{J}^*\right) \cdot (K-m)}{(K-m)} \\
&= \mathop{\mathbb{E}}_{\rho^-}\left[P_t\left(A^{TS}_{\rho^-} \in \mathcal{J}^*\right)\right] \cdot (K-m) \\
&\overset{(H1)}{\leq} P_t\left(A^{TS}_{m+1} \in \mathcal{J}^*\right) \cdot (K-m)
\end{aligned}
\tag{5.30}
$$

In the first step, we express the probability of error in terms of the arms that are ranked as sub-optimal by TS. In the second step, we apply a union bound. In the third and fourth step, we transform the sum to an expected value. In the final step, we apply Heuristic 1 $(H1)$.

Following analogous arguments, we derive the bound in terms of A_m^{TS}:

$$\begin{aligned} P_t\left(\mathcal{J}^* \neq \mathcal{J}^{TS}\right) &= P_t\left(\bigvee_{\rho^+} A_{\rho^+}^{TS} \in \overline{\mathcal{J}^*}\right) \\ &\leq \sum_{\rho^+} P_t\left(A_{\rho^+}^{TS} \in \overline{\mathcal{J}^*}\right) \\ &= \frac{\sum_{\rho^+} P_t\left(A_{\rho^+}^{TS} \in \overline{\mathcal{J}^*}\right) \cdot m}{m} \\ &= \mathop{\mathbb{E}}_{\rho^+}\left[P_t\left(A_{\rho^+}^{TS} \in \overline{\mathcal{J}^*}\right)\right] \cdot m \\ &\overset{(H2)}{\leq} P_t\left(A_m^{TS} \in \overline{\mathcal{J}^*}\right) \cdot m \end{aligned} \tag{5.31}$$

In the first step, we express the probability of error in terms of the arms that are ranked as optimal by TS. In the second step, we apply a union bound. In the third and fourth step, we transform the sum to an expected value. In the final step, we apply Heuristic 2 $(H2)$.

These insights motivate a uniform selection of the two arms on both sides of the decision boundary, as is reflected in BFTS (see Algorithm 6).

The BFTS algorithm is constructed such that its sampling strategy is completely independent of its recommendation strategy. Likewise, in this analysis, we consider the belief that BFTS maintains over the problem, in terms of the random variable $\mathcal{J}^{TS}$ (Equation 5.24), rather than the statistic that is used to make recommendations (e.g., the mean of the posterior in our experiments). This observation shows that our analysis is independent from the statistic used to make recommendations with BFTS.

When inspecting other algorithms for the m-top setting, we observe that the decision boundary between the m^{th} and $m+1^{\text{th}}$ arms also plays an important role. For example, the frequentist algorithm AT-LUCB samples two arms each step; the one with the smallest lower-bound among the top m arms, and the one with the greatest upper-bound among the rest. This is analogous to choosing the optimal and sub-optimal arms that are closest to the decision boundary.

6.1 Empirical validation of the heuristics

To experimentally validate Heuristic 1 and Heuristic 2 we express the inequalities in terms of sums over all arms.

For Heuristic 1, we have on the left-hand side:

$$\begin{aligned} P_t(A^{TS}_{\rho^-} \in \mathcal{J}^*) &\overset{(I)}{=} \sum_a^K P_t(A^{TS}_{\rho^-} = a) P_t(a \in \mathcal{J}^*) \\ &\overset{(PM)}{=} \sum_a^K P_t(A^{\pi}_{\rho^-} = a) P_t(a \in \mathcal{J}^*), \end{aligned} \tag{5.32}$$

and on the right-hand side:

$$\begin{aligned} P_t(A^{TS}_{m+1} \in \mathcal{J}^*) &\overset{(I)}{=} \sum_a^K P_t(A^{TS}_{m+1} = a) P_t(a \in \mathcal{J}^*) \\ &\overset{(PM)}{=} \sum_a^K P_t(A^{\pi}_{m+1} = a) P_t(a \in \mathcal{J}^*), \end{aligned} \tag{5.33}$$

where we rely on the independence of BFTS rankings with respect to the optimal set (I), and probability matching (PM, Equation 5.23).

For Heuristic 2, following analogous arguments, we have:

$$\begin{aligned} P_t(A^{TS}_{\rho^+} \in \overline{\mathcal{J}^*}) &= \sum_a^K P_t(A^{\pi}_{\rho^+} = a) P_t(a \in \overline{\mathcal{J}^*}) \\ P_t(A^{TS}_{m} \in \overline{\mathcal{J}^*}) &= \sum_a^K P_t(A^{\pi}_{m} = a) P_t(a \in \overline{\mathcal{J}^*}) \end{aligned} \tag{5.34}$$

All ranking probabilities in Equations 5.32 to 5.34 are in terms of the belief over the means. Therefore, these ranking probabilities can be estimated at each step of BFTS by obtaining a set of samples from the posterior. Each element in this set represents a sample from our belief over the means, and by ordering the entries in this sample, we obtain a ranking. By doing this over all samples in the set, we obtain a frequency distribution over rankings.

Given that both heuristics consider an expected value in terms of K on the left-hand side and that the equations above denote a sum over all arms, estimating the heuristics experimentally has a computational complexity that is quadratic in K. Therefore, in our experiments, we use bandit environments with $K = 100$ and $m = 5$, instead of the 1000-armed bandit environments that were covered in Section 5. We evaluate the heuristics for the same environments as in Section 5. For each environment, we run 100 BFTS replicates for $3 \cdot 10^4$ time steps. As we show in the Appendix 8, these experiments also demonstrate a clear learning curve. While running BFTS, we compute the probabilities that make up

the heuristics at every 100-th time step. Consequently, for each heuristic a total of $3 \cdot 10^4$ samples were collected per environment.

Our experiments show that for the two Gaussian environments with fixed variance (i.e., linear and polynomial) both heuristics hold for all measurements. For the epidemic bandit, Heuristic 1 holds for all measurements. For Heuristic 2, we recorded 9 failures ($= 0.03\%$), of which most (8 out of 9) occurred during the initial time steps ($t \leq 2000$) of BFTS. For the organic bandit, Heuristic 1 holds for all measurements. For Heuristic 2, we recorded 46 failures ($< 0.154\%$), of which most (40 out of 46) occurred during the initial time steps ($t \leq 2000$) of BFTS. For the cartoon caption bandit, Heuristic 1 holds for all measurements. For Heuristic 2, we recorded 139 failures ($< 0.467\%$), of which most (111 out of 139) occurred during the initial time steps ($t \leq 2000$) of BFTS.

These results show that failures are rare. We also observe that failures mostly occur during the earlier time steps of BFTS' execution. This can be explained by the fact that the posteriors have not yet converged to a symmetric bell shape at this time. We would indeed expect to see less failures when the posteriors are bell-shaped, and this intuition is supported by the experiments with Gaussian posteriors.

Furthermore, as Heuristic 2 is to be interpreted as a zero-bounded difference in probabilities:

$$\mathop{\mathbb{E}}_{\rho^+}[P_t(A^{TS}_{\rho^+} \in \overline{\mathcal{J}^*})] - P_t(A^{TS}_m \in \overline{\mathcal{J}^*}) \leq 0, \tag{5.35}$$

we note that all failures are caused by differences that are close to zero. Moreover, we observe that the general trend (i.e., the mean over the trajectories) is well below the zero-bound, as shown in Figure 5.9. These results indicate that the heuristics hold in expectation.

7 Discussion

In this chapter we introduce BFTS, a new algorithm for the anytime explore-m problem. We show that BFTS' exploration strategy is well-grounded using a formal Bayesian analysis. We empirically show that BFTS consistently outperforms the current state-of-the-art algorithm AT-LUCB, in a variety of experimental settings, even when uninformative priors are used (i.e., Gaussian with fixed and unknown variance, Categorical and Poisson reward distributions).

BFTS is a Bayesian algorithm, which means that prior knowledge with respect to the problem can be easily incorporated. This is important, as for many real world problems such information is available, e.g., the cartoon caption contest [Jamieson et al., 2015], epidemic decision problems and settings with correlated arms [Hoffman et al., 2014].

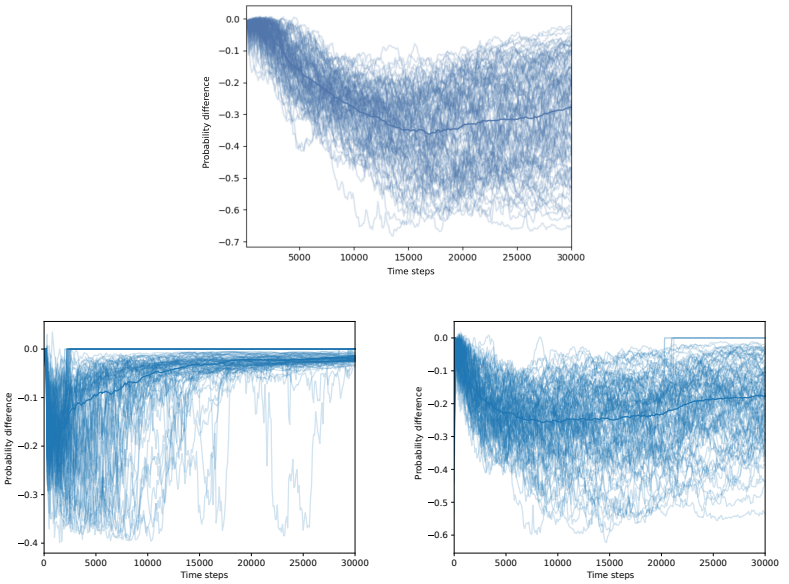


Figure 5.9: Trajectories of probability differences (Equation 5.35) over time for Heuristic 2 (traces: light blue lines, mean: thick blue line) for three environments: Categorical (top panel), scaled Gaussian (bottom left) and Poisson (bottom right).

As expected from its assumptions imposed on the reward distribution, AT-LUCB performs poorly in non sub-Gaussian settings, as we experimentally confirm in Section 5. This can be explained by the symmetric bound used by AT-LUCB (see Section 2), which will make bandit problems with a highly skewed reward distribution (e.g., Poisson) hard to solve. From our Bayesian analysis, it is clear that BFTS is not bound by such restrictions, and only relies on two heuristics, for which we argue that they are sensible in the context of probability matching when reasonable priors are chosen. This is an interesting observation that moves the assumption away from the reward distribution, which is inherently problem-specific. Instead, the assumption is placed on the posterior distribution, which represents the belief over the bandit's arms' means. While reward distributions are static, posteriors evolve when rewards are observed, and due to the central limit theorem, we expect that any specified prior over means will eventually tend towards a Gaussian [Billingsley, 2008]. This is important, as we expect both of the heuristics to hold well for bell-shaped posteriors, which was empirically supported through our experiments in Section 6.

While we performed a Bayesian analysis that provides important insights in BFTS' sampling strategy, a bound on the probability of error still needs to be established. We want to assert that, to our best knowledge, no such proofs have been established with respect to TS in the context of pure exploration. Even for vanilla TS, it took almost 80 years to come up with a tight bound on cumulative regret [Thompson, 1933; Agrawal and Goyal, 2012].

For future work, we acknowledge that additional efforts on theoretical guarantees are warranted, and we believe the heuristics proposed in Section 6 could provide a potential starting point, when additional constraints are imposed.

Moreover, we are in progress of evaluating a variant of BFTS, which we call δ-BFTS. δ-BFTS is a relaxed version of BFTS, where the m^{th} or $m+1^{\text{th}}$ ordered arm are still played with high probability, but with probability δ we uniformly choose a different arm. Our initial experimental analysis indicates that for small δ, the performance of δ-BFTS is similar to the performance of BFTS. Yet, by adding this additional exploration, we believe it to be possible to prove that the probability of error decays exponentially, and we are in the progress of pursuing this direction.

6 | Spatial model for pandemic influenza

> The purpose of models is not to fit the data but to sharpen the questions.
>
> Samuel Karlin

We introduce a new spatial epidemiological model in the context of pandemic influenza, that we aim to use in combination with reinforcement learning, to investigate preventive strategies. We construct a meta-population model, with a set of connected patches, where each patch corresponds to one of the administrative regions in Great Britain. To define the model, we describe the model internals of the patches and express the mobility model that connects the different patches. Finally, we conduct experiments to validate our new model, and discuss the model's computational complexity and performance.

The work presented in this chapter was conducted in collaboration with Prof. Dr. Niel Hens (UHasselt), Prof. Dr. Ann Nowé (Vrije Universiteit Brussel), Prof. Dr. Philippe Lemey (KU Leuven), Arno Moonens (Vrije Universiteit Brussel), Timothy Verstraeten (Vrije Universiteit Brussel). A manuscript to report the results presented in this and the next chapter was accepted at the 2020 Adaptive and Learning Agents Workshop at AAMAS.

1 Rationale and objectives

In this chapter, we set out the goal to build a realistic spatial epidemiological model with a focus on pandemic pathogens (e.g., pandemic influenza). We aim to use this model to learn adaptive preventive strategies using reinforcement learning, where our epidemiological model will act as the reinforcement learning environment. As the state-of-the-art of reinforcement learning techniques require many interactions with the environment in order to converge, an important objective is to construct a realistic model that still has a favourable computational performance [Yu, 2018]. To this end, we construct a meta-population model that consists out of a set of interconnected patches. Each patch corresponds to an administrative region in Great Britain and is internally represented by an age-dependent stochastic SEIR model. Great Britain consists out of three countries with the following administrative regions: 325 districts in England[1], 22 unitary authorities in Wales[2] and 32 council areas in Scotland[3] (see Figure 6.1).

Our choice for the age-dependent SEIR model is well grounded, as many preventive strategies are impacted by the age structure of the population, e.g., school closure and vaccine allocation [Eames et al., 2012; Medlock and Galvani, 2009]. The simulation is initialized by seeding a set of patches. In an uninfected patch, the infection process starts when the cumulative infectious potential of the surrounding infected patches reaches a stochastic threshold. Once a patch becomes infected, the infection process within this patch evolves independently of the rest of the system. In turn, this newly infected patch will contribute to the cumulative infectious potential of the other patches in the model.

The model's patches, that correspond to administrative regions in Great Britain, represent important urban and rural geographic areas and therefore present a realistic scale to model the mixing of children and adults within a compartment model. On the one hand, the size of these regions enables us to learn fine-grained preventive strategies. On the other hand, these regions are on an appropriate level that the outcome will still be interpretable to policy makers.

To validate that our model meets the predetermined goals, we conduct two experiments. Firstly, we compare the model to a SEIR compartment model that uses the same contact matrix and age structure. Secondly, we show that our model is able to reproduce the trends that were observed during the 2009 influenza pandemic.

[1] https://en.wikipedia.org/wiki/Districts_of_England
[2] https://en.wikipedia.org/wiki/Districts_of_Wales
[3] https://en.wikipedia.org/wiki/Subdivisions_of_Scotland

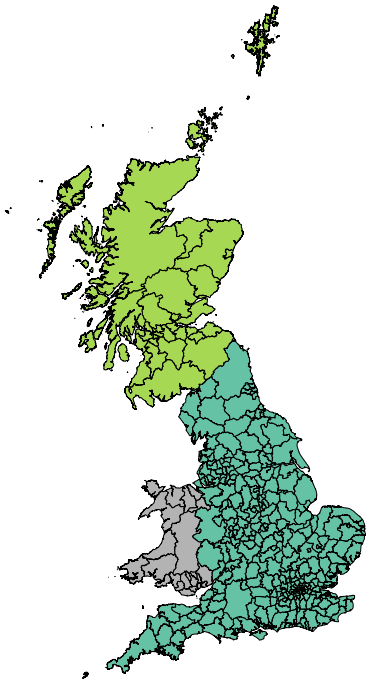


Figure 6.1: Great Britain and its three countries: England (mint-green), Wales (grey) and Scotland (lime-green). The thin black lines depict the border of the countries' administrative regions.

2 Related work

Meta-population models are used to investigate epidemics in a spatially structured host population. As detailed in Chapter 3, these models make it possible to balance the tradeoff of model resolution and the amount of computation that is required. Meta-population models are used to study many different pathogens, e.g., Ebola virus [D'Silva and Eisenberg, 2017], arboviruses such as Dengue virus [Marini et al., 2019], HIV [Coffee et al., 2007], and influenza [Klepac et al., 2018; De Luca et al., 2018].

For influenza, transmission models have been used to investigate the effect of school closures on an emerging pandemic [Merler et al., 2011; Apolloni et al., 2013; Fumanelli et al., 2016; Ciavarella et al., 2016; Eames et al., 2012].

Several model-assisted studies were performed to assess the quality of school closure policies [Fumanelli et al., 2016; Germann et al., 2019]. In this and the next chapter, it is our objective to study school closure policies using reinforcement learning, to investigate whether there are dynamic policies that can improve state-of-the-art policies with respect to their performance and flexibility. To our best knowledge, both this research question and the methodology that we use to address it is novel.

As recent work has shown that school closure decisions are best taken on a city or county level [Germann et al., 2019], we investigate school closure on the level of Great Britain's administrative regions. Therefore, in this chapter, we present a new model with two major objectives: the model should be able to realistically model school closures in a geospatial context and the model should be computationally efficient such that it can be used in combination with reinforcement learning algorithms. To allow for an efficient model we construct a patch-based meta-population model, which is inspired by the patch-based metapopulation model introduced by Klepac et al. [2018], where we use Poison processes and stochastic differential equations to introduce stochasticity. To model school closures realistically, for each patch, we use the age-dependent SEIR model Eames et al. [2012], that uses calibrated contact matrices based on a contact survey conducted in Great Britain.

3 Intra-patch age-dependent SEIR model

We consider a stochastic SEIR compartment model [Anderson and May, 1992], from which we sample trajectories. A SEIR model divides the population in a susceptible, exposed, infected and recovered compartment (see Section 1.2 in Chapter 3). More specifically, we consider an age-dependent SEIR model with a set of n disjoint age groups [Eames et al., 2012; Fumanelli et al., 2012]. This model is formally described by this system of ordinary

differential equations (ODEs), defined for each age group i:

$$\begin{aligned}\dot{S}_i(t) &= -\phi_i(t)S_i(t)\\ \dot{E}_i(t) &= \phi_i(t)S_i(t) - \zeta E_i(t)\\ \dot{I}_i(t) &= \zeta E_i(t) - \gamma I_i(t)\\ \dot{R}_i(t) &= \gamma I_i(t).\end{aligned} \tag{6.1}$$

Every susceptible individual in age group i is subject to an age-specific and time-dependent force of infection:

$$\phi_i(t) = \sum_{j=0}^{n} \beta M_{ij}(t) \frac{I_j(t)}{N_j(t)}, \tag{6.2}$$

which depends on:

- The probability of transmission β when a contact occurs.
- The average frequency $M_{ij}(t)$ of contacts that an individual in age group i has with an individual in age group j [Fung et al., 2015]. Note that the matrix $M_{ij}(t)$ is time-dependent, in order to enable the modelling of school closures. We discuss how this contact matrix M is established in Section 3.1. Following [Eames et al., 2012], we consider *conversational contacts*, i.e., contacts for which physical touch is not required.
- The frequency that contacts with infected individuals (in age group j) occur: $I_j(t)/N_j(t)$

Once exposed, individuals move to the infected state according to the latency rate ζ. Individuals recover from infection (i.e., get better or die) at a recovery rate γ.

We omit vital dynamics (i.e., births and deaths that are not caused by the epidemic) in this SEIR model, as the epidemic's time scale is short and we therefore assume that births and deaths will have a limited influence on the epidemic process [Towers and Feng, 2012]. Therefore, at any time:

$$N_i(t) = S_i(t) + E_i(t) + I_i(t) + R_i(t), \tag{6.3}$$

where N_i is initialized based on age-specific census data (see section 3.2).

3.1 Contact matrix

Our objective is to evaluate the effectiveness of social distancing interventions, more specifically school closures. For this reason, we use the United Kingdom contact matrices

presented in [Eames et al., 2012]. This study reports contact matrices for both conversational contacts and physical contacts. For each of these contact types a contact matrix, representative for school term and school holiday, is available. These contact matrices are the result of an internet-based social contact survey completed by a cohort of participants [Eames et al., 2012].

The contact matrices encode four age groups: children (0-4 years), adolescents (5-18 years), adults (19-64 years) and elderly (65 years and older).

The original contact matrices, as obtained from the survey report, i.e., C_{ij}, denote the average number of people in age group j met each day by a person in age group i. While in theory, we would expect this matrix to be reciprocal, in practice this will not necessarily be the case due to differences in reporting between the different age groups. We correct the contact matrices to compensate for this effect and explain this process in the next sub-section.

Contact matrix reciprocity

A contact matrix is reciprocal when for each tuple of age groups (i, j), the total number of contacts between individuals of age group i with individuals of age group j is equal to the total number of contacts between individuals of age group j with individuals of age group i [Wallinga et al., 2006]. Therefore, for any reciprocal contact matrix M we must have:

$$N_i M_{ij} = N_j M_{ji} \tag{6.4}$$

When we have a contact matrix C_{ij} that was obtained from a survey, it typically will not be fully reciprocal. In order to use the contact matrix in an epidemiological model, we therefore need to transform it into a reciprocal matrix [Eames et al., 2012; Fung et al., 2015].

Our aim is thus to find a reciprocal approximation M of the survey contact matrix C. For each entry M_{ij}, we have two error terms we want to minimize:

$$\begin{aligned} N_i M_{ij} - N_i C_{ij} \\ N_i M_{ij} - N_j C_{ji} \end{aligned} \tag{6.5}$$

When formulating this as a least squares problem, our objective is to minimize the following sum of squared residuals with respect to M_{ij}:

$$\Delta_{ij} = (N_i M_{ij} - N_i C_{ij})^2 + (N_i M_{ij} - N_j C_{ji})^2, \tag{6.6}$$

of which the derivative is:

$$\begin{aligned}\frac{d\Delta_{ij}}{dM_{ij}} &= 2N_i(N_iM_{ij} - N_iC_{ij}) + 2N_i(N_iM_{ij} - N_jC_{ji}) \\ &= 2N_i\left(2N_iM_{ij} - N_iC_{ij} - N_jC_{ji}\right),\end{aligned} \tag{6.7}$$

The optimum has a derivative that is equal to 0:

$$0 = 2N_iM_{ij} - N_iC_{ij} - N_jC_{ji}, \tag{6.8}$$

from which we obtain this transformation expression that minimizes the error terms specified in Equation 6.5:

$$M_{ij} = \frac{N_iC_{ij} + N_jC_{ji}}{2N_i}. \tag{6.9}$$

Note that M_{ij} is reciprocal:

$$N_iM_{ij} = N_jM_{ji} = \frac{N_iC_{ij} + N_jC_{ji}}{2} \tag{6.10}$$

The expression in Equation 6.10 is also used in [Eames et al., 2012; Fung et al., 2015], although in these works no source or derivation of the expression was given.

3.2 Census data

Each compartment model is representative of one of the districts defined in Section 1, and as such the compartment model is parametrised with the census data of the respective district, i.e., population counts stratified by age groups. We use the 2011 United Kingdom census data made available by NOMIS[4]. This dataset contains census data for all of the considered districts for the following age groups: 0-4, 5-7, 8-9, 10-14, 15, 16-17, 18-19, 20-24, 25-29, 30-44, 45-59,60-64,65-74,75-84,85-89, 90-90+.

To be compatible with our model, we need to map this census data to the age structure imposed by the Eames contact matrix (see Section 3.1): i.e., 0-4 years (children), 5-18 years (adolescents), 19-64 years (adults), 65-90+ years (elderly). To define this mapping, we will refer to the number of individuals with the symbol N, subscripted with the dataset type (i.e., NOMIS or Eames) and the age group.

For the age group 0-4 and 65-90+ we have a direct mapping:

$$\begin{aligned}N_{\text{Eames,Children}} &= N_{\text{NOMIS,0-4}} \\ N_{\text{Eames,Elderly}} &= N_{\text{NOMIS,65-90+}}\end{aligned} \tag{6.11}$$

[4] https://www.nomisweb.co.uk

However, as for the contact matrix the adolescents and adults are split between age 18 and 19, and for the census data these 2 age groups are aggregated, we need to make a custom mapping. We will aggregate all shared age groups and divide the common age group in two:

$$\begin{aligned} N_{\text{Eames,Adolescents}} &= N_{\text{NOMIS,5-17}} + \left\lceil \frac{N_{\text{NOMIS,18-19}}}{2} \right\rceil \\ N_{\text{Eames,Adults}} &= \left\lceil \frac{N_{\text{NOMIS,18-19}}}{2} \right\rceil + N_{\text{NOMIS,20-64}} \end{aligned} \tag{6.12}$$

When restructuring the census data according to the Eames age groups, we observe clear trends over the districts with respect to the proportion of children, adolescents, adults and elderly, as shown in Figure 6.2. However, the histograms in Figure 6.2 only show the marginalized distribution per age group. To reason about the distribution over all age groups, consider that we have a proportion of each of the age groups, and we thus can represent this data as a positive simplex [Aitchison, 1983], as defined in Equation 21.

Definition 21: Unit simplex

A unit simplex [Aitchison and Pawlowsky-Glahn, 1997], with D components, corresponds to the set:

$$\mathbb{S}^D = \{\langle x_1, ..., x_D \rangle \mid \forall x_i : x_i > 0, \sum_{i=1}^{D} x_i = 1\}. \tag{6.13}$$

This representation enables us to reason about this data in a statistical framework, as we will do in the next chapter, and to visualize the four-dimensional data in a three-dimensional space by using the Barycentric coordinate system, as shown in Figure 6.3. Figure 6.3 shows that the census distribution exhibits a dense region with a limited number of outliers.

Note that we use the 2011 census dataset, rather than the more recent 2018 census dataset, to be fully compatible with the mobility dataset used to inform our between-patch transition model (see Section 4).

For each district, the base contact matrix is corrected to make it reciprocal (see Section 3.1), using that district's census data.

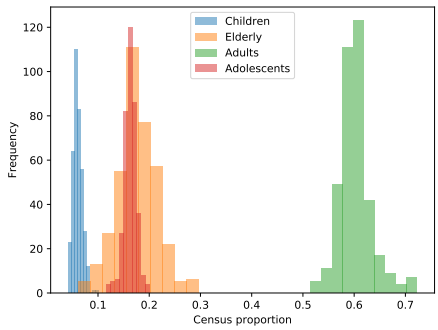


Figure 6.2: Histograms of the census proportions in the districts of Great Britain, according to Eames' age structure.

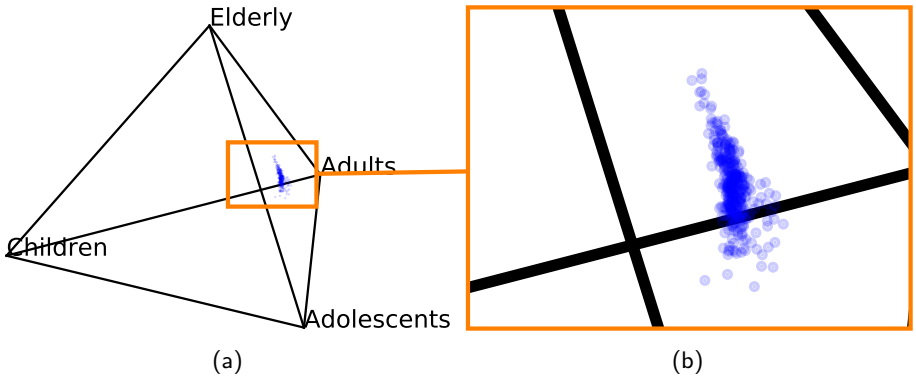


Figure 6.3: Barycentric projection of the census proportions in the districts of Great Britain, according to Eames' age structure. Each scatter point corresponds to one district, and each axis corresponds to the proportion of the age groups it connects. The left panel shows the original census pyramid, and the right panel zooms in on the point cloud.

3.3 Parametrising the model with R_0

In influenza modelling literature, it is common to parametrise the model in terms of a specific R_0 value. We will now introduce an equation that enables us to compute the transmission probability β for a given R_0 value.

To this end, we need to determine the next-generation matrix, which summarizes the number of secondary infections between age groups [Vynnycky and White, 2010], and determine the spectral radius of this matrix (see Definition 22).

Following Diekmann et al. [2009] and Eames et al. [2012], we construct the next-generation matrix for our SEIR model:

$$K = \frac{\beta M}{\gamma} \tag{6.14}$$

Definition 22: Spectral radius

The spectral radius of a matrix L, $\Upsilon(L)$, is the dominant eigenvalue of that matrix L.

Given K, we can compute R_0 as:

$$R_0 = \frac{\beta}{\gamma}\Upsilon(M). \tag{6.15}$$

As the contact matrix is a square matrix with positive real entries, according to the Perron-Frobenius theorem, the dominant eigenvalue exists and is unique. Note that, in a graph, the spectral radius is a measure of the graph's connectivity [Lewis, 2011]. As our contact matrix M can be seen as a graph that represents how strongly the different age groups are connected, this notion of connectivity applies here as well.

Using equation 6.15, we can now compute the transmission risk β for a given R_0, γ and contact matrix M.

Note that for each district, we have a contact matrix that is corrected for reciprocity by using that district's census data (see Section 3.1). Therefore, we have a distribution over $\Upsilon(M_d)$, where M_d is a contact matrix for district d. We would expect that this distribution is centred around $\Upsilon(M_{GB})$, where M_{GB} the contact matrix that is corrected for reciprocity using the census data representative for Great-Britain in its entirety (i.e., an aggregation of all the districts). This is confirmed in Figure 6.4, which shows that the median of the distribution over $\Upsilon(M_d)$ coincides with $\Upsilon(M_{GB})$. Furthermore, note that the contact matrix denotes the average frequency of contacts that an individual in age group i has with an individual in age group j. Figure 6.4 thus shows a limited variance ($\sigma^2 = 0.001$).

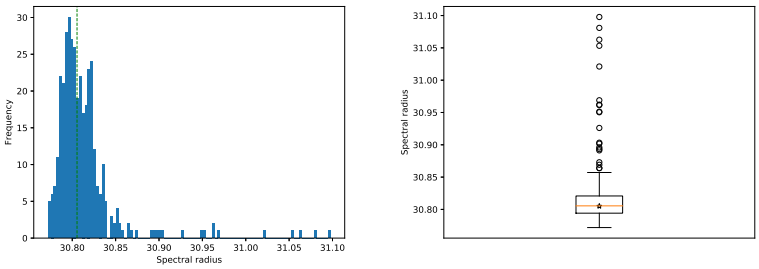


Figure 6.4: Both figures display the distribution of the contact matrix' spectral radius for the different districts. To demonstrate the shape of this distribution, the left panel shows a histogram, annotated with a dotted green line that represents the median. To demonstrate the distribution with respect to its quartiles and outliers, the right panel shows a box plot of the distribution, annotated with an orange line that represents the median and a yellow star that shows the spectral radius for Great Britain.

3.4 Stochastic trajectories

In order to obtain stochastic trajectories from the SEIR model, we transform the ODEs to stochastic differential equations (SDEs), using the transformation procedure presented by Allen et al. [2008]. We remind the reader that this procedure considers the compartments and transitions of the original ODE and adds white noise terms for each transition in the ODE, as specified in Section 1.4 in Chapter 3.

For our SEIR model, we have this set of SDEs for each age group i:

$$\begin{aligned}
\dot{S}_i &= -\phi_i S_i - \sqrt{\phi_i S_i}\dot{\mathcal{W}}_{(S_i \to E_i)} \\
\dot{E}_i &= \phi_i S_i - \zeta E_i + \sqrt{\phi_i S_i}\dot{\mathcal{W}}_{(S_i \to E_i)} - \sqrt{\zeta E_i}\dot{\mathcal{W}}_{(E_i \to I_i)} \\
\dot{I}_i &= \zeta E_i - \gamma I_i + \sqrt{\zeta E_i}\dot{\mathcal{W}}_{(E_i \to I_i)} - \sqrt{\gamma I_i}\dot{\mathcal{W}}_{(I_i \to R_i)} \\
\dot{R}_i &= \gamma I_i + \sqrt{\gamma I_i}\dot{\mathcal{W}}_{(I_i \to R_i)},
\end{aligned} \tag{6.16}$$

where $\mathcal{W}$ is a Wiener process.

From this system, we can sample trajectories using the Euler-Maruyama approximation method [Allen et al., 2008], using r random numbers per time step, where r has a linear complexity with respect to the number of model transitions (details on the Euler-Maruyama approximation in Section 1.4 in Chapter 3).

3.5 Compartment model parameters

To summarize, the compartment model is parametrised by:

- a contact matrix (see Section 3.1)
- the census data for the administrative regions (see Section 3.2)
- an R_0 value (see Section 3.3)
- a latency rate ζ
- a recovery rate γ

4 Between-patch model

Our model, that is comprised of a set of connected SEIR patches, is inspired by the recent BBC pandemic model [Klepac et al., 2018]. The BBC pandemic model was in its turn motivated by the model presented in [Gog et al., 2014].

At each time step, our model decides whether a patch p becomes infected. This is modulated by the patch's force of infection, which combines the potential of the infected patches in the system, weighted by a mobility model:

$$\mathring{\phi}_p(t) = \sum_{p' \in \mathcal{P}} \mathcal{M}_{p'p} \cdot \beta \cdot \left(S_p^{\mathsf{A}}(t)\right)^{\mu} \cdot \mathcal{I}_{p'}(t), \tag{6.17}$$

where $\mathcal{P}$ is the set of patches in the model, $\mathcal{M}_{p'p}$ is the mobility flux between patch p' and p, β is the probability of transmission on a contact (see Section 3.5), $S_p^{\mathsf{A}}(t)$ is the susceptible population of adults in patch p at time t and its contribution is modulated by parameter μ, and $\mathcal{I}_{p'}(t)$ is the infectious potential of patch p' at time t. We define this infectious potential as,

$$\mathcal{I}_{p'}(t) = I_{p'}^{\mathsf{A}}(t) \cdot M_{\mathsf{AA}}, \tag{6.18}$$

where $I_{p'}^{\mathsf{A}}(t)$ is the size at time t of the infectious adult population and M_{AA} is the average number of contacts between adults (see Section 3.1).

$\mathcal{M}$ is a matrix based on the mobility dataset provided by NOMIS[5]. This dataset describes the amount of commuting between the districts in Great-Britain.

In general, this between-patch model is constructed from first principles i.e., census data, a mobility model, the number of infected individuals and the transmission potential of the virus. However, for the parameter μ that modulates the contribution of the susceptibles in the receptive patch, while it is commonly used in literature [Gog et al., 2014; Eggo et al.,

[5]We use the NOMIS WU03UK dataset that was released in 2011.

2010; Kissler et al., 2019], no such intuition is readily available. Therefore, this parameter is fitted to match the properties of the epidemic that is under investigation [Gog et al., 2014; Eggo et al., 2010; Kissler et al., 2019].

To validate our model (see Section 5), we conduct two experiments. Firstly, we compare our model to the original compartment model and perform a sensitivity analysis with respect to parameter μ. Secondly, we show that our model fits the recent influenza pandemic of 2009, by choosing an appropriate value for all model parameters.

Given this time-dependent force of infection, we model the event that a patch becomes infected with a non-homogeneous Poisson process [Wang and Wu, 2018; Tomba and Wallinga, 2008; Barthélemy et al., 2010]. A Poisson process can be used to model the occurrence of events with a given intensity (see Definition 24), and non-homogeneous Poisson processes generalize this concept to time-dependent intensities (see Definition 25). As the process' intensity depends on how the model (i.e., the set of all patches) evolves, we cannot sample the time at which a patch becomes infected a priori. Therefore, we determine this time of infection using the time scale transformation algorithm [Cinlar, 2013]. Firstly, we explain the generic time scale transformation algorithm (Section 4.1). Secondly, we adjust the algorithm to our setting (Section 4.2).

4.1 Time scale transformation algorithm

The time scale transformation algorithm (TSTA) enables us to determine the time at which an event, modelled by a non-homogeneous Poisson process, will take place [Cinlar, 2013].

We will start by formally defining the homogeneous and non-homogeneous Poisson process. A Poisson process is a counting arrival process, defined on a sample space Ω with probability measure P.

Definition 23: Arrival process

An arrival process is a stochastic process [Cinlar, 2013],

$$\mathcal{N} = \{\mathcal{N}_t; t \geq 0\}, \tag{6.19}$$

such that for any $\omega \in \Omega$, the mapping $t \to \mathcal{N}_t(\omega)$, has $\mathcal{N}_0 = 0$, is non-decreasing, increases only by integer jumps and is right continuous.

Definition 24: Homogeneous Poisson process

A *homogeneous* Poisson process is an arrival process [Cinlar, 2013; Bertsekas and Tsitsiklis, 2002],

$$\mathcal{P} = \{\mathcal{P}_t; t \geq 0\}, \tag{6.20}$$

for which these axioms hold:

1. for almost all $\omega \in \Omega$, $t \to \mathcal{P}_t(\omega)$ jumps in steps of size 1
2. the number of arrivals within any interval $[t..t+s]$, is independent of the history of arrivals prior to t (i.e., $\forall t, s \geq 0 : \mathcal{P}_{t+s} - \mathcal{P}_t \perp\!\!\!\perp \{\mathcal{P}_u; u \leq t\}$)
3. the process is time-homogeneous (i.e., $\forall t, s \geq 0 : \mathcal{P}_{t+s} - \mathcal{P}_t \perp\!\!\!\perp t$)

From this definition, we can show that for each homogeneous Poisson process $\mathcal{P}$:

$$\forall t \geq 0 : P(\mathcal{P}_t = k) = \frac{e^{-\lambda t}(\lambda t)^k}{k!}, \tag{6.21}$$

for some constant $\lambda \geq 0$, where λ signifies the intensity (i.e., rate) of the process.

The concept of a Poisson process can be generalized to a *non-homogeneous* Poisson process by removing the time-homogeneity requirement:

Definition 25: Non-homogeneous Poisson process

A *non-homogeneous* Poisson process is an arrival process [Cinlar, 2013],

$$\mathcal{P}^{\lambda(t)} = \{\mathcal{P}_t^{\lambda(t)}; t \geq 0\}, \tag{6.22}$$

for which these axioms hold:

1. for almost all $\omega \in \Omega$, $t \to \mathcal{P}_t(\omega)$ jumps in steps of size 1
2. the number of arrivals within any interval $[t, t+s]$, is independent of the history of arrivals prior to t (i.e., $\forall t, s \geq 0 : \mathcal{P}_{t+s}^{\lambda(t)} - \mathcal{P}_t^{\lambda(t)} \perp\!\!\!\perp \{\mathcal{P}_u^{\lambda(t)}; u \leq t\}$)

$\mathcal{P}_t^{\lambda(t)}$ has a time-dependent rate that is specified by the intensity function $\lambda(t)$, where $\lambda(t) \geq 0$.

We define the process' cumulative intensity function:

Definition 26: Cumulative intensity function

A non-homogeneous Poisson process $\mathcal{P}^{\lambda(t)}$ with intensity function $\lambda(t)$ has a cumulative intensity function:

$$\Lambda(t) = \int_0^t \lambda(s)ds \tag{6.23}$$

Furthermore, we can show that [Osaki, 2012]:

$$\mathbb{E}[\mathcal{P}_{t+h}^{\lambda(t)} - \mathcal{P}_t^{\lambda(t)}] = \int_t^{t+h} \lambda(s)ds, \tag{6.24}$$

and thus we have that $\Lambda(t)$ is the expectation function of $\mathcal{P}_t^{\lambda(t)}$:

$$\Lambda(t) = \mathbb{E}[\mathcal{P}_t^{\lambda(t)}] \tag{6.25}$$

From Definition 26, it is clear that $\Lambda(t)$ will be a non-decreasing function and at least right-continuous.

The crucial theorem that underlies the time scale transformation algorithm denotes that the arrival times in a non-homogeneous Poisson process can be mapped to a homogeneous Poisson process with rate 1 (Theorem 1). We present an example that demonstrates this theorem in Figure 6.5.

Theorem 1: Mapping non-homogeneous Poisson processes

Let Λ be a continuous non-decreasing cumulative intensity function. Then,

$$T_1, T_2, ... \tag{6.26}$$

are the arrival times in a non-homogeneous Poisson process if and only if

$$\Lambda(T_1), \Lambda(T_2), ... \tag{6.27}$$

are the arrival times in a homogeneous Poisson process with rate 1 [Cinlar, 2013].

The time scale transformation algorithm uses the relation in Theorem 1 to transform a homogeneous Poisson process with $\lambda = 1$ into a non-homogeneous Poisson process with expectation function Λ. The homogeneous process is formed by sampling from an exponential probability distribution with $\lambda = 1$. To make this transformation possible, a time inverse function of Λ is required:

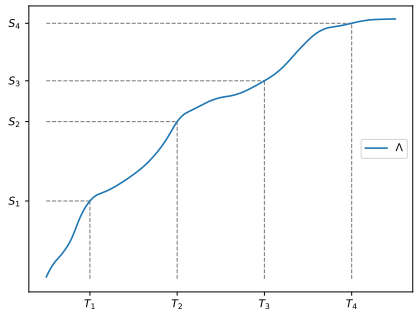


Figure 6.5: A visual example of Theorem 1: $T_1, T_2, ...$ form a non-homogeneous Poisson process with expectation function Λ if and only if $S_1, S_2, ...$ form a homogeneous Poisson process with rate 1.

Definition 27: Time inverse of $\Lambda(t)$

The time inverse τ of an expectation function $\Lambda(t)$ for a non-homogeneous Poisson process $\mathcal{P}_t^{\lambda(t)}$:

$$\tau(s) = \inf\{t : \Lambda(t) > s\} \tag{6.28}$$

In Algorithm 7, we formalize this procedure. At each step i, we obtain a sample X_i from an exponential probability distribution with rate $\lambda = 1$, which is added to the set of samples $\mathcal{X}_i$. The sum of the elements in $\mathcal{X}_i$ represents the i^{th} arrival in the homogeneous Poisson process, and is transformed into the i^{th} arrival in the non-homogeneous Poisson process using the inverse time function $\tau(.)$.

$\mathcal{X}_0 = \varnothing$
for $i = 1, \ldots$ **do**
  $X_i \sim \mathcal{E}\text{xp}(\lambda = 1)$
  $\mathcal{X}_i = \mathcal{X}_{i-1} \cup \{X_i\}$
  $t_i = \tau\left(\sum_{x \in \mathcal{X}_i} x\right)$
end

Algorithm 7: Time scale transformation algorithm

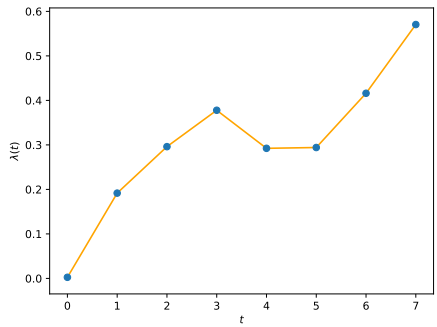


Figure 6.6: An example of a piecewise linear intensity function for a patch in our model (see Equation 6.29). The blue scatter points represent the evaluation of $\phi(t)$ (Equation 6.17) at discrete time steps (i.e., the end of each day). The orange connecting lines represent the linear interpolation between $\phi(i-1)$ and $\phi(i)$.

4.2 TSTA to model the infection of patches

In order to use the time scale transformation algorithm (Algorithm 7) in our epidemiological model, note that the patches' internal state is updated in a discrete number of time steps. We determine a patch's intensity $\phi_p(t)$ (Equation 6.17) at the end of each day. This results in a sequence of intensities between which we linearly interpolate to obtain a piecewise linear intensity function:

$$\lambda(t) = \text{line}(t, \lceil t \rceil - 1, \lceil t \rceil, \phi_p(t)), \tag{6.29}$$

where

$$\text{line}(x, x_1, x_2, f) = f(x_1) + \frac{f(x_2) - f(x_1)}{x_2 - x_1}(x - x_1) \tag{6.30}$$

interpolates linearly between $(x_1, f(x_1))$ and $(x_2, f(x_2))$.

This piecewise linear intensity function $\lambda(t)$ is continuous and thus its cumulative counterpart $\Lambda(t)$ is continuous as well. Furthermore, as $\phi_p(t) \geq 0$ for all $t \geq 0$, $\Lambda(t)$ is non-decreasing.

As $\phi_p(t)$ depends on the simulation state at time t, it is clear that we cannot evaluate this function beyond the current simulator time step. However, the definition of the time inverse τ (Definition 27) shows that we can use the arrival time in the homogeneous Poisson process as a threshold for the arrival time in the non-homogeneous Poisson process. We

formalize this threshold-based time scale transformation algorithm in Algorithm 8. Note that this algorithm approximates the original algorithm as we check whether the threshold is surpassed at discrete time steps.

$X \sim \mathcal{E}\text{xp}(\lambda = 1)$
for $t = 1, \ldots$ **do**
  if $\Lambda(t) \geq X$ **then**
    Trigger event
    $X^{(t)} \sim \mathcal{E}\text{xp}(\lambda = 1)$
    $X = X + X^{(t)}$
  end
end

Algorithm 8: Time scale transformation algorithm using discrete time steps

Following Klepac et al. [2018], we assume that a patch will become infected only once.

5 Model validation

Our objective is to construct a model that is representative for contemporary Great Britain with respect to population census and mobility trends. This model is to be used to study school closure intervention strategies for future influenza pandemics. While in many studies [Kissler et al., 2019; Gog et al., 2014; Eggo et al., 2010], a model is created specifically to fit one epidemic case, we aim for a model that is robust with respect to different epidemic parameters, most importantly R_0, the basic reproduction number (Definition 13).

To validate our model according to these goals, we conduct two experiments. In the first experiment, we compare our patch model to a SEIR compartment model that uses the same contact matrix and age structure. While we do not expect our model to behave exactly like the compartment model, as the patches and the mobility network that connects them induces a different dynamic, we do expect to see similar trends with respect to the epidemic curve and peak day. In the second experiment we show that our model is able to reproduce the trends that were observed during the 2009 influenza pandemic, commonly known as the swine-origin influenza pandemic, that originated in Mexico. The 2009 influenza pandemic in Great Britain is an interesting case to validate our model for three main reasons. Firstly, the pandemic occurred quite recently and thus our model's census and mobility scheme should be a good fit, as both the datasets on which we base our census and mobility model were released in 2011. Secondly, due to the time when the virus entered Great Britain, the summer holiday started 11 weeks after the emergence of the epidemic. The timing of the holidays had a severe impact on the progress of the epidemic and resulted in a epidemic

curve with two peaks. This characteristic epidemic curve enables us to demonstrate the predictive power of our age-dependent contact model with support for school closures. Thirdly, the number of symptomatic cases that occurred in Great Britain during the 2009 pandemic was recorded meticulously and is publicly available [Kubiak and McLean, 2012].

5.1 Comparison to the Eames SEIR compartment model

In this experiment, we compare our patch model to a simple SEIR model that encompasses the same age structure and contact matrix (details in Section 3) [Eames et al., 2012], to which we will refer as Eames-SEIR from this point onwards. We consider a stochastic implementation of the Eames-SEIR, using the same principles that we introduced in Section 3.4. This experimental setting will be central to the reinforcement learning experiments, related to finding optimal school closure policies, that we will present in the next chapter.

Following Eames et al. [2012] and Baguelin et al. [2010], we use a latent period of one day ($\zeta = \frac{1}{1}$) and an infectious period of 1.8 days ($\gamma = \frac{1}{1.8}$). As in Chapter 4, we perform our experiment for a set of R_0 values within the range of 1.4 to 2.4, in steps of 0.2. This range is considered representative for the epidemic potential of influenza pandemics [Basta et al., 2009; Medlock and Galvani, 2009].

Furthermore, we need to choose a value for the parameter in the between-patch model, i.e., μ, that modulates the contribution of susceptible adults in the receiving patch (see Section 4). This parameter is typically fitted towards data, however, in this experiment and in the reinforcement learning experiments in the next chapter, we consider a model to investigate future epidemics. Our goal is to calibrate our model such that it produces peak days that are similar to the peak days in Eames-SEIR [Eames et al., 2012], which is a prominent model for pandemic influenza that moreover generates peak days that are in agreement with earlier work [Ferguson et al., 2006]. Therefore, we investigate the effect of μ in this setting, through a sensitivity analysis. We consider μ in the interval $[0, 1]$, where the left end of the interval (i.e., $\mu = 0$) signifies that the contribution of susceptible adults is ignored and the right end of the interval (i.e., $\mu = 1$) signifies that the contribution of adults is not modulated. In Figure 6.7, we show the results for the sensitivity analysis for $\mu \in \{0, 0.1, 0.2, 0.3, 0.4, 0.5, 1\}$, together with the peak days for the Eames-SEIR model. From these results, it is clear that the different values for μ form a gradient within the interval $[0, 1]$.

However, no value of μ provides a good fit for all of the considered R_0's, when comparing the peak days to the Eames-SEIR model. Rather, we can discern a log-relationship between μ and the best fit for the different R_0's. Based on this observation, we propose to define:

$$\mu = \log(R_0) \cdot s, \tag{6.31}$$

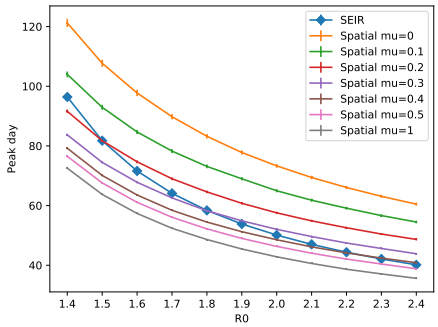


Figure 6.7: Time of peak day (y-axis) for $R_0 \in \{1.4, 1.6, 1.8, 2.0, 2.2, 2.4\}$ (x-axis). A curve is shown for different values of μ (plain curve) and the peak days as produced by the Eames-SEIR model (curve with diamond scatter points). For each R_0, 100 stochastic trajectories were sampled and the bound signifies the 95% confidence interval of the sample.

where s is a scaling factor. For this experimental setting, we find that $s = .6$ provides a good fit for all of the considered R_0's, which we show in Figure 6.8.

Provided this choice of μ, when we compare the epidemic trajectories of our spatial model with the Eames-SEIR model in Figure 6.9, we observe similar trends with respect to the shape of the trajectory distributions. The main difference is that the epidemic curves grow slower in our spatial model than in the Eames-SEIR model and also achieve a lower peak incidence. This is expected, as we constrain mixing in our spatial model within the districts, and thus increase the resolution of our model, which has been shown to more accurately predict peak incidence [Mills and Riley, 2014].

Furthermore, in Figure 6.10, we show the number of districts that get infected over time for different R_0 values. This shows that all districts get infected, and the time it takes for all districts to reach this point depends mainly on the transmission-ability of the virus strain.

We expect the attack rate to be similar for the Eames-SEIR and spatial model. When all districts get infected, the attack rate in the spatial model is the sum of the attack rates of a set of Eames-SEIR models (i.e., one Eames-SEIR model per district). We verified that the attack rates are indeed nearly identical, as shown in Figure 6.11, with little variance for either of the models.

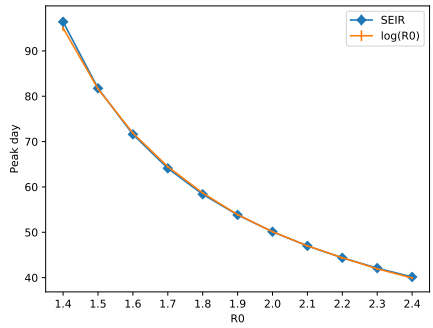


Figure 6.8: Number of peak days (y-axis) for $R_0 \in \{1.4, 1.6, 1.8, 2.0, 2.2, 2.4\}$ (x-axis). A curve is shown for $\mu = log(R_0) \cdot 0.6$ (orange curve) and the peak days as produced by the Eames-SEIR model (blue curve with diamond scatter points). For each R_0, 100 stochastic trajectories were sampled and the bound signifies the 95% confidence interval of the sample.

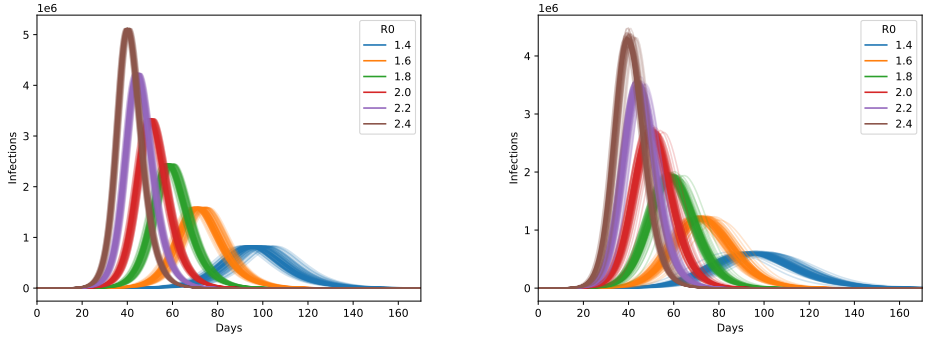


Figure 6.9: Epidemic trajectories for the Eames-SEIR model (left panel) and the spatial model (right panel). One epidemic trajectory encodes the number of infections per day. Trajectory distributions are shown for $R_0 \in \{1.4, 1.6, 1.8, 2.0, 2.2, 2.4\}$, with a different colour per reproductive number. For each R_0, the distribution consists out of 100 trajectory samples.

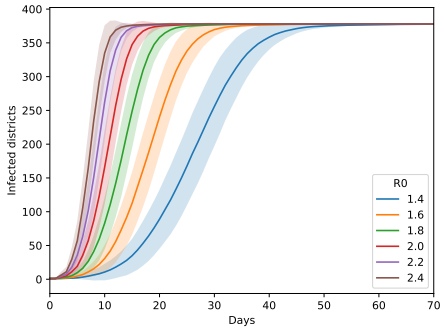


Figure 6.10: Number of infected districts (y-axis) per day (x-axis) for $R_0 \in \{1.4, 1.6, 1.8, 2.0, 2.2, 2.4\}$. For each R_0, 100 stochastic trajectories were were sampled, of which the curve represents the mean, and the bound represents the standard deviation of the samples.

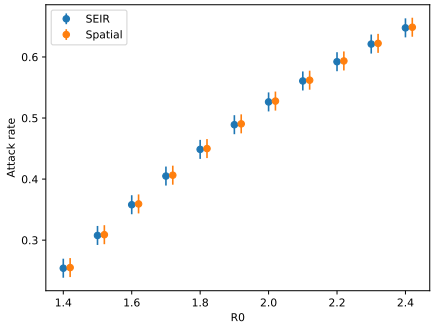


Figure 6.11: Attack rate (y-axis) for $R_0 \in \{1.4, 1.6, 1.8, 2.0, 2.2, 2.4\}$ (x-axis). Results are shown for the Eames-SEIR model (blue scatter) and the spatial model (orange scatter). For each model, we depict the standard deviation as bars on top of the scatter points. For each R_0, 100 stochastic trajectories were obtained.

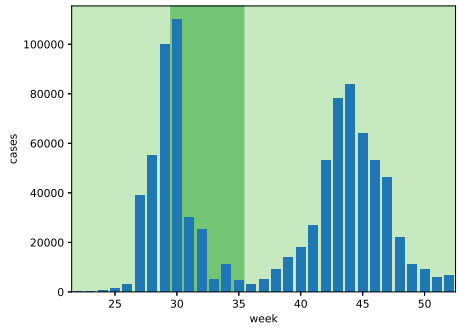


Figure 6.12: This figure shows the amount of cases that were recorded by the HPA on a weekly basis (blue bars). The background in this figure signifies the time of the summer holidays (dark green).

5.2 2009 influenza pandemic in Great Britain

The virus responsible for the 2009 influenza pandemic arrived in Great Britain during the first week of May 2009 (week 19). Following this introduction, the epidemic grew for 11 weeks until the summer school holidays started, after which the epidemic showed its first peak. After the school holidays, the epidemic was rekindled and grew to a second peak. In Figure 6.12 we show the weekly case count, as recorded by the British Health Protection Agency (HPA) and the time at which the school holidays take place.

To reproduce this distinctive epidemic curve, we use our original model as it was described throughout this chapter. We consider two free parameters: the basic reproduction number and the time of the infectious period. The general consensus is that the basic reproduction number was moderate during the 2009 influenza pandemic, with estimates ranging from 1.16 to 2 [Kubiak and McLean, 2012; Tuite et al., 2010; De Silva et al., 2009; Fraser et al., 2009; Yang et al., 2009; Nishiura et al., 2010; Balcan et al., 2009]. We present a detailed overview of the reported basic reproduction number estimates in Table 6.1. For the period of infectiousness we found estimates of 1.8, 2.5 and 3.38 days [Eames et al., 2012; Balcan et al., 2009; Tuite et al., 2010]. We present a detailed overview of the infectious period estimates in Table 6.2.

Given these prior estimates, we parametrize our model with a basic reproduction number that is in the range of 1.2 to 2.0 and consider a duration of infectiousness of respectively 1.8, 2.5 and 3.38 days.

R_0	**Source**
1.22-1.58	Fraser et al. [2009]
1.3-1.7	Yang et al. [2009]
1.21-1.35	Nishiura et al. [2010]
1.75	Balcan et al. [2009]
1.87-2.07	De Silva et al. [2009]
1.31	Tuite et al. [2010]
1.16-1.59	Kubiak and McLean [2012]

Table 6.1: Overview of basic reproduction numbers from literature.

Infectious period	**Source**
1.8	Eames et al. [2012]
2.5	Balcan et al. [2009]
3.38	Tuite et al. [2010]

Table 6.2: Overview of infectious periods from literature.

For this experiment, we found,

$$\mu = \log(R_0) \cdot 2.74, \tag{6.32}$$

to be a good fit for the overall comparison. In Figure 6.13 we show the epidemic curve for our model with respect to these parameters. In general, the epidemic curves that result from using an infectious period of 1.8 days are insufficient to reproduce the trends of the 2009 pandemic. For the other infectious periods (i.e., 2.5 and 3.38), we show that for all but the highest reproductive numbers we observe an epidemic curve with 2 peaks. Furthermore, we observe a deeper trough in the epidemic curve when an infectious period of 2.5 days is chosen.

In Figure 6.14, we show a set of model realisations in conjunction with the symptomatic case data, which shows that we were able to closely match the epidemic trends observed during the British pandemic in 2009. This model was configured with a basic reproductive number of 1.4 and infectious period of 2.6. The reproductive number is in good concordance with the general consensus that the virus responsible for the 2009 pandemic exhibited a moderate infectiousness. While the infectious period slightly differs from the value reported by Balcan et al. [2009] (i.e., 2.5 days), it lies well within the confidence bounds reported in this study (confidence interval: 1.1-4.0 days). Note that our model reports the number of infections, while the HPA only recorded symptomatic cases. Therefore we scale the epidemic curve with a factor of $\frac{1}{4}$. While atypical, this large number of asymptomatic

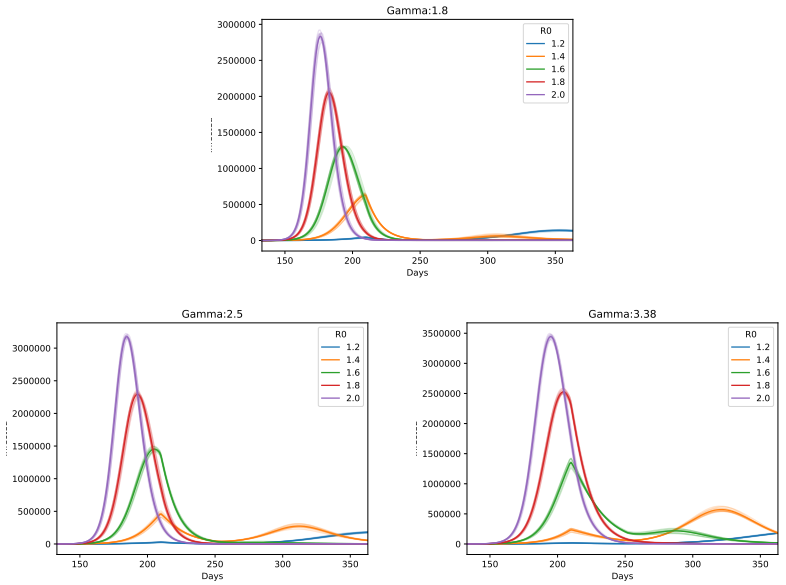


Figure 6.13: We demonstrate our model for $R_0 \in \{1.2, 1.4, 1.6, 1.8, 2.0\}$ (enumerated in the legend) and an infectious period of 1.8 days (top panel), 2.5 days (bottom left panel) and 3.38 days (bottom right panel). For each parameter combination, we show a set of stochastic trajectories (light coloured lines) and the mean of these trajectories (dark coloured line). For clarity, we only show 10 stochastic trajectories in this Figure. In Appendix 9, we show a Figure for the same model configuration, with 100 stochastic trajectories.

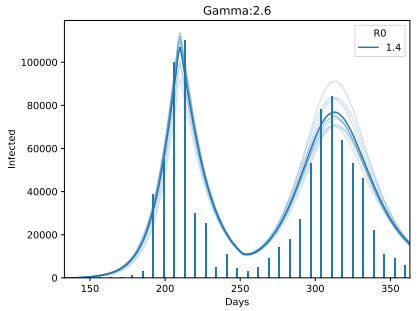


Figure 6.14: We show that our model, using a reproductive number of 1.4 and a an average duration of infectiousness of 2.6 days is able to match the trends observed in the British pandemic of 2009. For clarity, we only show 10 stochastic trajectories in this Figure. In Appendix 9, we show a Figure for the same model configuration, with 100 stochastic trajectories.

cases produced by our model is in line with earlier serological surveys [Miller et al., 2010] and with previous modelling studies [Kubiak and McLean, 2012].

6 Computational complexity and performance

An analysis of the computational complexity of our model needs to consider that the model incorporates two components. On the one hand, infection in the patches is triggered via the time scale transformation algorithm (see Section 4.2). On the other hand, once infected, each patch in the system evolves independently, and we use the Euler-Maruyama approximation method to obtain samples from the stochastic differential equation that is associated with the patch (see Section 3.4). The time scale transformation algorithm samples a random threshold for each patch, which is compared to the force of infection of the associated patch. This comparison occurs at each time step that the patch was not yet infected. Computing the force of infection in Equation 6.17 considers all model patches, and thus has a worst case complexity that is linear in the number of model patches, i.e.,

$$\mathcal{O}(|\mathcal{P}|). \tag{6.33}$$

In worst case, at each time step t, if only one of the model patches is infected, and we need to compute the force of infection for each patch, which has a quadratic complexity

in the number of model patches, i.e.,

$$\mathcal{O}(t \cdot |\mathcal{P}|^2). \tag{6.34}$$

However, when each patch only needs to be infected once, we observe that we have a complexity in terms of infected $\mathcal{P}_i$ and uninfected patches $\mathcal{P}_{\neg i}$. After all, at each time step, we only need to consider the force of infection of the uninfected patches, and this force of infection only takes into account the infected patches, i.e.,

$$\mathcal{O}(t \cdot |\mathcal{P}_i \cdot \mathcal{P}_{\neg i}|). \tag{6.35}$$

Since we have,

$$\mathcal{P} = \mathcal{P}_i + \mathcal{P}_{\neg i}, \tag{6.36}$$

we can see that while this expression has the same worst case complexity as in Equation 6.34, in practice less operations will be required.

For both the complexity in Equation 6.34 and Equation 6.35, it is clear that as long as the number of patches is limited, as is the case in our model, this procedure will be computationally efficient, as this can be implemented as a vector product, on vectors that all fit in memory (RAM).

When a patch is infected, at each time step it will be advanced by using a number of operations that is proportional to the number of compartments in the the age-dependent SEIR model.

This model was implemented in Python, and the performance critical sections were either implemented using NumPy when a vector representation was possible (e.g., to compute the force of infection) [Oliphant, 2006], or JIT-compiled using Numba (e.g., to evolve the age-dependent SEIR model in a patch) [Lam et al., 2015]. This implementation performs well, resulting in ≈ 2 simulation runs per second on a MacBook Pro.

7 Discussion

George Box said: "All models are wrong, but some are useful."

Indeed, in this chapter, we construct a model where we aim for realism, both with respect to the research question that we attempt to model, as in terms of computational efficiency. To this end, model decisions to balance this trade-off were made, which we motivated by references from literature. In this section, we will discuss some of these choices and some opportunities for future work.

Firstly, in this dissertation, we use conversational contact matrices (see Section 3), as these were shown to provide a better fit for the 2009 influenza pandemic by Eames et al.

[2012]. However, it would be interesting to quantify the difference between physical and conversational contact matrices, and investigate how they affect the epidemiological outcome of the model.

Secondly, we chose the contact matrices of Eames et al. [2012], to inform the social mixing of our model, as this study established contact matrices for both school terms and school holidays. These contact matrices were constructed based on a survey that was conducted in Great Britain, and we should be prudent when using these contact matrices outside of the British context. An interesting direction to use our models in different countries, is the work presented by Prem et al. [2017]. Prem et al. [2017] developed a methodology to project contact data surveys to different age-structured populations (i.e., home, work, and school), and can thus be used to model school closures. In [Prem et al., 2017], projections are performed on the well known contact survey of Mossong et al. [2008]. As Prem et al. [2017] also generated contact matrices for the United Kingdom, we believe that for future work, it would be interesting, to compare this to the work of Eames et al. [2012].

Thirdly, we chose to directly use the British mobility matrix provided by NOMIS (see Section 4). This mobility matrix is publicly available and has the same resolution as our patch structure. However, for many other countries, such a detailed and rich dataset may not be (publicly) available. As many epidemiological models use some kind of variation of the gravity model [Klepac et al., 2018; Gog et al., 2014], for future work, it would be interesting to use this mobility matrix to calibrate different gravity models, to improve our understanding on how different gravity models affect the epidemiological model. While a similar study have been conducted in the context of the mobility in the United Kingdom [Truscott and Ferguson, 2012], this study is limited to traditional gravity models, which might be hard to parametrize when no data is available, due to the large number of model parameters. To this end, it would be interesting to investigate how different intervening opportunity mobility models [Stouffer, 1940] relate to this mobility dataset. Such models rely on a stochastic process, and only require information on the population distribution as input [Simini et al., 2012]. The most well known intervening opportunity mobility model is the recent radiation model [Simini et al., 2012] that expresses the average flux between two geographic areas i and j as:

$$\mathcal{M}_{ij} = \mathcal{M}_i \frac{m_i n_j}{(m_i + s_{ij})(m_i + n_j + s_{ij})}, \tag{6.37}$$

where m_i and n_j are the population sizes of the respective areas i and j, s_{ij} is the circle of radius r_{ij} centred at area i, r_{ij} is the distance between area i and j and $\mathcal{M}_i$ is the total number of commuters. The radiation model was recently evaluated in the context of an Ebola model [Kraemer et al., 2019], and an evaluation in the context of the British mobility network would be complementary to this work. While the radiation model is an interesting model, and the model has been shown to work well when the scale of the mobility is

large enough, it has problems to explain intra-city mobility mobility dynamics [Masucci et al., 2013]. Therefore, it would be interesting to investigate new models that attempt to reconcile between intra- and inter-city mobility, such as the recent work presented by Liu and Yan [2019] and Mazzoli et al. [2019].

Finally, the techniques we used to construct this model could be applied in the context of other pathogens, such as for example Ebola virus [D'Silva and Eisenberg, 2017] and Dengue virus [Marini et al., 2019]. This could lead to interesting new research opportunities when these models are combined with reinforcement learning, to investigate dynamic mitigation policies.

7 | Studying school closure policies following a reinforcement learning approach

I believe in cooperating for the common good.

Erskine Bowles

We investigate the use of school closure policies to mitigate influenza pandemics in the model we constructed in the previous chapter, following a reinforcement learning approach. We formalize our learning environment and detail the epidemiological setting in which we conduct our experiments. We start our experiments by comparing Proximal Policy Optimization (PPO) and Deep Q-Networks (DQN). Next, we investigate the effect of population census compositions (i.e., the proportion of the different age groups in a population) on the outcome of school closure policies, using PPO. In this experimental setting, we establish the ground truth, that allows us to experimentally assess the performance of the policies learned by PPO. These experiments demonstrate the potential of deep reinforcement learning algorithms to learn policies in the context of more complex epidemiological

models (i.e., larger state and/or action space), for which no ground truth can be established. Finally, we investigate whether there is an advantage to consider multiple districts simultaneously when devising school closure policies. To this end, we associate each district with an agent that controls whether schools should be opened or closed, and we model this research question as a multi-agent reinforcement learning problem, as this allows us to investigate the joint behaviour of the agents. Through this analysis, we show that there is a collaborative advantage when designing school closures policies.

The work presented in this chapter was conducted in collaboration with Prof. Dr. Ann Nowé (Vrije Universiteit Brussel), Prof. Dr. Philippe Lemey (KU Leuven), Prof. Dr. Niel Hens (UHasselt), Arno Moonens (Vrije Universiteit Brussel), Fabian Ramiro Perez Sanjines (Vrije Universiteit Brussel), Timothy Verstraeten (Vrije Universiteit Brussel) and Dr. Jelena Grujic (Vrije Universiteit Brussel). A manuscript to report the results presented in this and the previous chapter was accepted at the 2020 Adaptive and Learning Agents Workshop at AAMAS.

1 Rationale and objectives

In the previous chapter, we constructed a spatial age-dependent meta-population model, to model school closure policies in the context of an influenza pandemic. Now, we will study the effect of such school closure policies following a reinforcement learning approach.

School closures policies can be used to serve two distinct objectives, on the one hand, to reduce the attack rate (Definition 19), and on the other hand, to delay the peak of the epidemic. Shifting the peak of an influenza pandemic can be useful for two main reasons. Firstly, to reduce the need for means of healthcare (e.g., hospital beds or antiviral drugs), which is important at the start of a pandemic, as healthcare providers will not be prepared for the sudden exponential rise in influenza cases [Cauchemez et al., 2009]. Secondly, to delay the epidemic until a vaccine, that is tailored towards the strain associated with the pandemic, is available [Markel et al., 2007]. In this chapter, we will conduct experiments to investigate policies that aim to reduce the attack rate, and discuss how these methods can be applied to investigate policies to delay the peak of the epidemic.

To investigate school closure policies, we study the use of deep reinforcement learning algorithms to learn directly on an epidemic model. Firstly, we consider two prominent deep reinforcement learning algorithms to investigate the learning of optimal school closure policies on the level of a single district, i.e., Deep Q-Networks (DQN), a Q-learning algorithm, and Proximal Policy Optimization (PPO), a policy gradient algorithm. Secondly, we investigate the effect of population census compositions (i.e., the proportion of the different age groups in a population) on the outcome of school closure policies, using PPO. In this experimental setting, we establish a ground truth, that allows us to experimentally

assess the performance of the policies learned by PPO. Through these experiments, we demonstrate the potential of deep reinforcement learning algorithms to learn policies in the context of complex epidemiological models, opening the prospect to learn in even more complex stochastic models with large action spaces. Furthermore, we show insights on the effect of population census compositions on school closures. Thirdly, we investigate a multi-agent reinforcement learning approach, to examine whether there is an advantage in collaborating between districts, when implementing school closure policies. On the one hand, the current state-of-the-art of deep multi-agent reinforcement learning algorithms is only able to deal with a limited number of agents (≈ 10 agents) [Hernandez-Leal et al., 2019]. On the other hand, in our model, we have 379 agents, one for each district, as agents represent a district for which they can control school closure. Therefore, we need to partition these agents into smaller groups, such that the use of multi-agent reinforcement learning algorithms becomes feasible. We do this by analysing the mobility graph that connects the different districts. Through this analysis, we identify components of tightly coupled districts. Finally, we discuss how these methods can be used to investigate policies to delay the peak of the epidemic, by defining and evaluating a reward function targeted towards this objective.

To conduct our experiments, we establish a learning environment, based on the epidemiological model that we introduced in the previous chapter. Therefore, we construct a Markov Decision Process (MDP), with a state space that directly corresponds to our epidemiological model, an action space that allows us to open and close schools on a weekly basis, a transition function that follows the epidemiological model's dynamics, and a reward function that is targeted to the objective of reducing the attack rate.

In this work, we consider an epidemiological model for which the parameters are assumed to be known (i.e., reproductive number, latency period and infectious period). We keep the latency and infectious period fixed throughout our experiments and consider two distinct values of the reproductive number: $R_0 = 1.8$ (i.e., moderate transmission potential) and $R_0 = 2.4$ (i.e., high transmission potential). We conduct our experiments for three distinct school closure budgets: 2, 4 and 6 weeks.

2 Related work

The closing of schools has been found an effective way to limit the spread of an influenza pandemic [Earn et al., 2012; Copeland et al., 2013; Chowell et al., 2011; Heymann et al., 2004; Markel et al., 2007; Germann et al., 2019]. For this reason, the use of school closures as a mitigation strategy has been explored in variety of modelling studies [Halder et al., 2010; Brown et al., 2011; Milne et al., 2013; De Luca et al., 2018; Ciavarella et al., 2016;

Germann et al., 2019; Halloran et al., 2008; Haber et al., 2007; Eames et al., 2012], of which the work by Germann et al. [2019] is the most recent and comprehensive study.

The concept to learn dynamic policies by formulating the decision problem as a Markov decision process (MDP) (see Chapter 2) was first introduced in [Yaesoubi and Cohen, 2011]. This technique was used to investigate dynamic tuberculosis case-finding policies in HIV/tuberculosis co-epidemics [Yaesoubi and Cohen, 2013]. Later, the technique was extended towards a methodology to include cost-effectiveness in the analysis [Yaesoubi and Cohen, 2016], and applied to investigate mitigation policies (i.e., school closures and vaccines) in the context of pandemic influenza in a simplified epidemiological model. On the one hand, the work presented in Yaesoubi and Cohen [2011, 2016] uses a policy iteration algorithm to solve the MDP. On the other hand, the use of on-line reinforcement learning techniques (e.g., TD-learning, policy gradient) has only been explored to a limited extent[1], and motivated us to do the work presented in this chapter. Note that Deep Q-networks (see Chapter 2) was recently used to investigate culling and vaccination in farms in a simple individual-based model to delay the spread of viruses in a cattle population [Probert et al., 2019]. However, to our best knowledge, the work presented in this chapter is the first attempt to use deep reinforcement learning algorithms directly on a complex meta-population model. Furthermore, we experimentally validate the performance of these algorithms using a ground truth, in a variety of model settings (i.e., different census compositions and different R_0's). This is the first validation of this kind and it demonstrates the potential of on-line deep reinforcement learning techniques in the context of epidemic decision making. Finally, we present a novel approach to investigate how intervention policies can be improved by enabling collaboration between different geographic districts, by formulating the setting as a multi-agent problem, and by solving it using deep multi-agent reinforcement learning algorithms.

3 Learning environment

In order to apply reinforcement learning, we construct a Markov Decision Process (MDP, Definition 7) based on the epidemiological model that we introduced in the previous chapter. This epidemiological model consists out of patches that correspond to administrative regions.

We have an agent for each patch that we attempt to control, and for each agent we have an action space $\mathcal{A} = \{\text{open}, \text{close}\}$ that allows us to open and close schools for one week. Each agent has a predefined budget b of school closure actions it can execute

[1] The recent perspective report by Yanez et al. reached the same conclusion as we did.

(Definition 28). Once this budget is depleted, executing a close action will default to executing an open action.

Definition 28: School closure budget

Each agent has a budget b of school closure actions it can execute. This budget thus corresponds to the number of weeks the agent can close schools during an influenza pandemic. We refer to the remaining budget at time t as $b^{(t)}$.

For each patch, we consider a state space that combines the state of the SEIR model and the remaining budget of school closures $b_p^{(t)}$. For the SEIR model, we have 16 state variables (i.e., $\mathbb{R}^{16}$), as we have a SEIR model (4 state variables) for each of the four age groups. The remaining school closure budget is encoded as an integer, resulting in a combined state space of 17 variables. We refer to the state space of one patch p, that thus combines the epidemiological states and the budget, as $\mathcal{S}_p$. The state space of the MDP, $\mathcal{S}$, corresponds to the aggregation of the state space of each patch that we attempt to control:

$$\bigtimes_{p \in \mathcal{P}^{\mathsf{c}}} \mathcal{S}_p, \tag{7.1}$$

where $\mathcal{P}^{\mathsf{c}}$ is the set of patches that we control.

The transition probability function $T(\mathbf{s}' \mid \mathbf{s}, \mathbf{a})$ evolves the epidemic state to the next week in the epidemic, taking into account the school closure actions that were chosen, using the epidemiological dynamics as defined in the previous chapter.

As stated in Section 1, school closures can serve two objectives, i.e., to reduce attack rate and to shift the peak day of the epidemic. To reduce the attack rate (Definition 19), we consider an immediate reward function that quantifies the negative loss in susceptibles over one simulated week, as formalized in Definition 29. In Section 10, we will define and evaluate a reward function targeted at shifting the peak day.

Definition 29: Reward function to reduce attack rate

The reward function to reduce the attack rate is:

$$R_{\mathsf{AR}}(\mathbf{s}, \mathbf{a}, \mathbf{s}') = -(S(\mathbf{s}) - S(\mathbf{s}')), \tag{7.2}$$

where $S(.)$ is the function that determines the total number of susceptible individuals given the state of the epidemiological model.

4 Epidemiological setting

Throughout this chapter, we consider the model introduced in Chapter 6. We conduct two kinds of experiments: in the context of a single district and in the context of the Great Britain model that combines all 379 districts. Following Eames et al. [2012]; Baguelin et al. [2010], we use a latent period of one day ($\zeta = \frac{1}{1}$) and an infectious period of 1.8 days ($\gamma = \frac{1}{1.8}$). We consider two distinct values for the reproductive number, i.e., $R_0 = \{1.8, 2.4\}$, where $R_0 = 1.8$ represents an epidemic with moderate transmission potential [Ferguson et al., 2006] and $R_0 = 2.4$ represents an epidemic with high transmission potential [Longini et al., 2005]. We investigate the effect of different school closure budgets (Definition 28), i.e., $\mathcal{b} = \{2, 4, 6\}$ weeks. The epidemic is simulated for a fixed number of weeks, chosen beforehand, to ensure that the epidemic finishes.

5 Proximal Policy Optimization and Deep Q-Networks

Throughout this chapter, we will mostly focus on the Proximal Policy Optimization algorithm (PPO), as PPO is a policy gradient variant (see Section 7 in Chapter 2), and such algorithms tend to be more suitable to deal with large action spaces, that are associated with the multi-agent context that we will investigate later in Section 9. Nonetheless, in this section we compare PPO to the Deep Q-Networks algorithm (DQN), which is a Q-learning variant (see Section 4 in Chapter 2), to show that both algorithms exhibit similar performance, once the algorithms' hyper-parameters are tuned. This preliminary analysis demonstrates that there is potential for both PPO and DQN to support the epidemiological decision making process.

For DQN, we use a neural network that accepts the state of the epidemiological model as input[2] and outputs a value for each action. Every hidden layer in the DQN network uses the rectified linear activation function. For PPO, the policy network also accepts the state of the epidemiological model, and the output of the network contains 1 unit, which is passed through a sigmoid activation function. This output thus represents the probability of keeping the schools open in the district. Every hidden layer in the PPO network uses the hyperbolic tangent activation function. The value network has the same architecture as the policy network, with the exception that the output is not passed through an activation function.

For this analysis, we operate in a single district (the Greenwich district in London, England), that is seeded on the first day. We obtain five trials of both PPO and DQN on the

[2]As detailed in Section 3, the state of the epidemiological model is comprised of the SEIR values for each age group, along with the remaining budget for closing schools in the district.

epidemiological setting introduced in Section 4, for which we show the reward curves for $R_0 = 2.4$ in Figure 7.1. These results show that both DQN and PPO quickly converge to a similar policy. For $R_0 = 1.8$, we have similar results, that we show in Appendix 10.

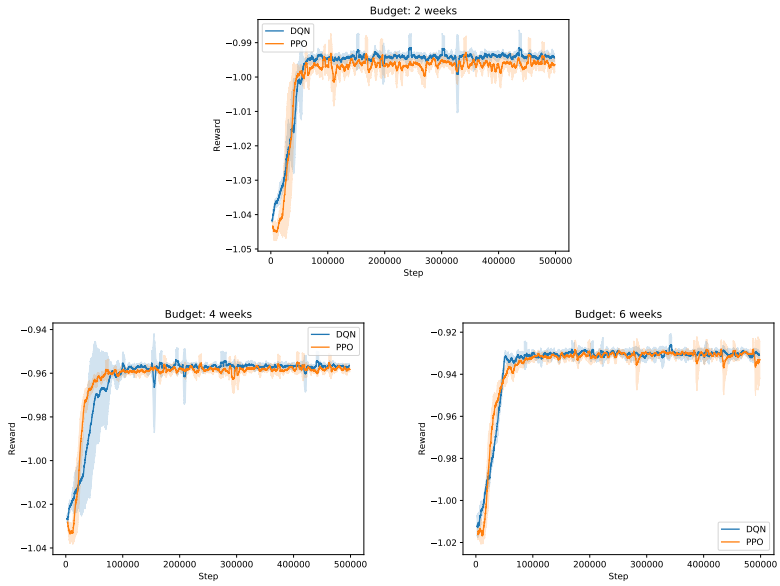


Figure 7.1: Learning results (i.e., learning curve that depicts the total reward per episode) for the Greenwich environment with $R_0 = 2.4$, for school closure budgets $b = \{2, 4, 6\}$. Reward curves for PPO (orange) and DQN (blue), using a rolling window of 100 steps. The shaded area shows the standard deviation of the reward signal.

Clearly, these results only demonstrate that both algorithms converge to a similar policy, which is no guarantee that an appropriate policy was learned. To verify this, we will establish a ground truth in the next section, and empirically validate that PPO converges to this ground truth.

6 Studying the effect of the population composition

We now establish a ground truth that investigates the effect of population compositions (i.e., the proportion of the different age groups in a population) on the outcome of school

closure policies. We will use this ground truth to empirically validate that PPO converges to the appropriate policy.

6.1 Establishing a ground truth

To establish a ground truth[3], first consider that when we deal with a single district, we can approach the average behaviour of the model by removing the stochastic terms from the differential equations (see Section 6 in Chapter 6). Hence, for a particular parameter configuration (i.e., district, R_0, γ, ζ), the model will always produce the same epidemic curve. This means that the state space of this deterministic epidemic model directly corresponds to the time of the epidemic. For an epidemic that spans w weeks, we can formulate a school closure policy as a binary number with w digits, where the digit at position i signifies whether schools should be open (1) or closed (0) during the i-th week. For short-lived epidemics, such as influenza pandemics[4], we can enumerate these policies and evaluate them in our model (i.e., using exhaustive policy search).

6.2 Selecting districts

To investigate the effect of population census compositions, we select 10 districts that are representative of the population heterogeneity in Great Britain. To this end, we remind the reader that in Section 3.2 of Chapter 6, we analysed the population heterogeneity by representing the population structure as a positive simplex. We select 10 districts: one district that is representative for the average of this distribution and a set of nine districts that is representative for the diversity in this distribution. To determine the average district, we consider the population heterogeneity distribution over all districts, and determine the Aitchison's mean (Definition 30) of this distribution [Aitchison, 1994]. We then select the district that is closest to the Aitchison's mean (Definition 30) according to the Aitchison distance (Definition 31), as shown in Figure 7.2.

[3]Note that this is a proxy to the ground truth, as we use a deterministic version of the model.
[4]In the settings we consider (see Section 4), the epidemic spans $w \leq 25$.

Definition 30: Aitchison's mean

Given a set of points from a unit simplex (Definition 21),

$$P = \{p^{(i)} \mid \ p^{(i)} \in \mathbb{S}^D\}_{i=1}^{N}, \tag{7.3}$$

the Aitchison's mean [Aitchison and Pawlowsky-Glahn, 1997] is:

$$C_A(P) = \frac{\langle h_1, ..., h_D \rangle}{\sum_{d=1}^{D} h_d}, \tag{7.4}$$

where,

$$h_d = \left(\prod_{p^{(i)} \in P} p_d^{(i)} \right)^{(1/N)}, \tag{7.5}$$

is the geometric mean of the d-th component over all simplex points in P.

Definition 31: Aitchison distance

Given two points from a unit simplex $p, q \in \mathbb{S}^D$ (Definition 21), we define the Aitchison distance function [Aitchison, 1992]:

$$d_A(p, q) = \left[\sum_{d=1}^{D} \left(\log \frac{p_d}{h(p)} - \log \frac{q_d}{h(q)} \right) \right]^{1/2}, \tag{7.6}$$

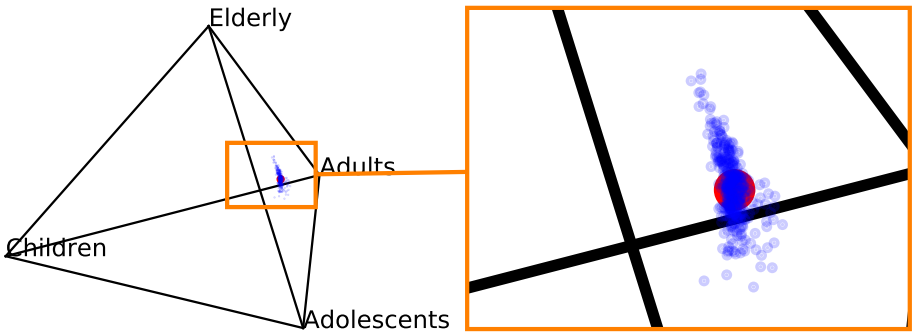
where,

$$h(p) = \left(\prod_{d=1}^{D} p_d \right)^{(1/D)}, \tag{7.7}$$

denotes the geometric mean of p. This distance defines a metric on the simplex sample space.

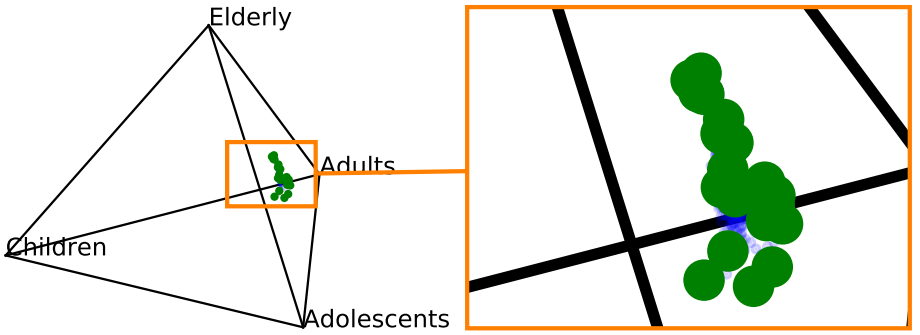
Next, we determine the outer extreme points, as these represent the most diverse census points. To do this, we compute the convex hull of the point cloud (i.e., the smallest convex set of points that contains the point cloud), as shown in Figure 7.3.

We proceed by taking the points that belong to the surface of the convex hull, of which we make a sub-selection of 9 census points. As the convex hull consists out of 21 points, we consider all k-combinations (with $k = 9$) and select the set of points that maximizes the minimum Aitchison distance between the selected points, as shown in Figure 7.4.



(a) (b)

Figure 7.2: Barycentric projection of the census proportions in the districts of Great Britain (blue scatter points), according to Eames' age structure. The geometric mean of this distribution is shown as a red scatter point. The left panel shows the original census pyramid, and the right panel zooms in on the point cloud.



(a) (b)

Figure 7.3: Barycentric projection of the census proportions in the districts of Great Britain (blue scatter points), according to Eames' age structure. The census points that are part of the convex hull are shown in green. The left panel shows the original census pyramid, and the right panel zooms in on the point cloud.

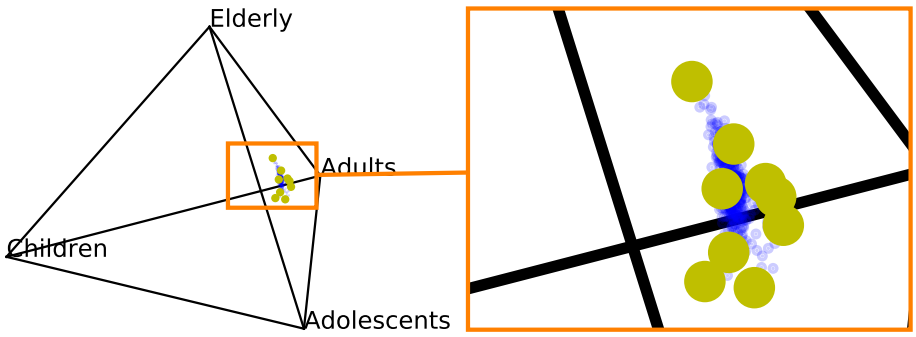


Figure 7.4: Barycentric projection of the census proportions in the districts of Great Britain (blue scatter points), according to Eames' age structure. The census point that were selected out of the convex hull are shown in yellow. The left panel shows the original census pyramid, and the right panel zooms in on the point cloud.

6.3 Exhaustive policy search

We consider the epidemiological setting that was introduced in Section 4 on the 10 districts that were selected in the previous section. For each district, we enumerate and evaluate all possible policies as described in Section 6.1, based on the improvement that the policy induces with respect to the attack rate as compared to the baseline, i.e., the model executed without school closures.

We show the range of the attack rates of the top 10 policies in Figure 7.5. This figure shows that there is some variation over the districts, which is expected, as these districts have different census configurations, which impacts the amount of influence school closures will have.

We show an overview of the top 10 policies as histograms, for the district that is closest to the geometric mean (i.e., the Barnsley district), with $R_0 \in \{2.4\}$ and school closure budgets $\acute{b} \in \{2, 4, 6\}$, in Figure 7.6. This figure indicates that the optimal policy is to focus the school closure budget around the peak day of the epidemic. This is interesting, as in previous work [Germann et al., 2019] that investigated the efficacy of school closure policies, only school closure events were considered at the start of the epidemic.

When inspecting these histograms for the different districts (Figures shown in Appendix 11), we observe that the optimal policy is the same for the different districts. Furthermore, for the other top policies there is some variability, yet all districts exhibit similar trends.

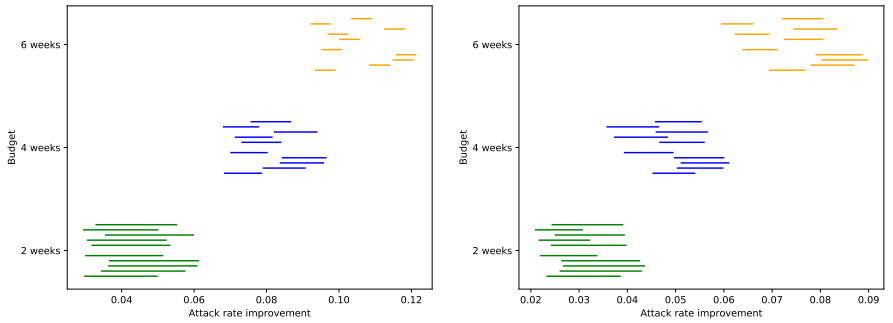


Figure 7.5: Top 10 policies for the experiment with $R_0 = 1.8$ (left panel) and $R_0 = 2.4$ (right panel), using a school closure budget of 2 (green), 4 (blue) and 6 (yellow) weeks. Each bar represents the range of attack rate improvements (over the top 10 policies) for one district.

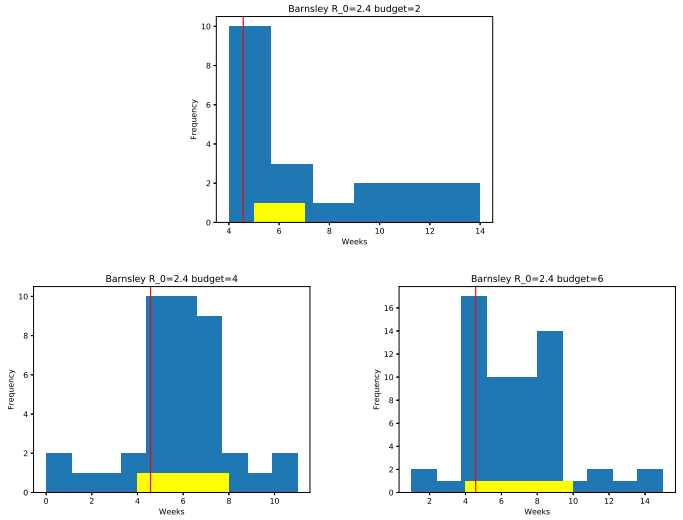


Figure 7.6: We visualize the top 10 policies as a histogram (blue) for $R_0 = 2.4$, where each bar in the histogram signifies the frequency of school closure events at a particular week. On top of this histogram, we show the top policy in yellow. The red vertical line indicates the peak day of the baseline epidemic. We show results for three budgets: 2 weeks (top), 4 weeks (bottom left) and 6 weeks (bottom right).

We show the epidemic trajectories, with optimal school closure strategy, for Barnsley with $R_0 = 2.4$ in Figure 7.7.

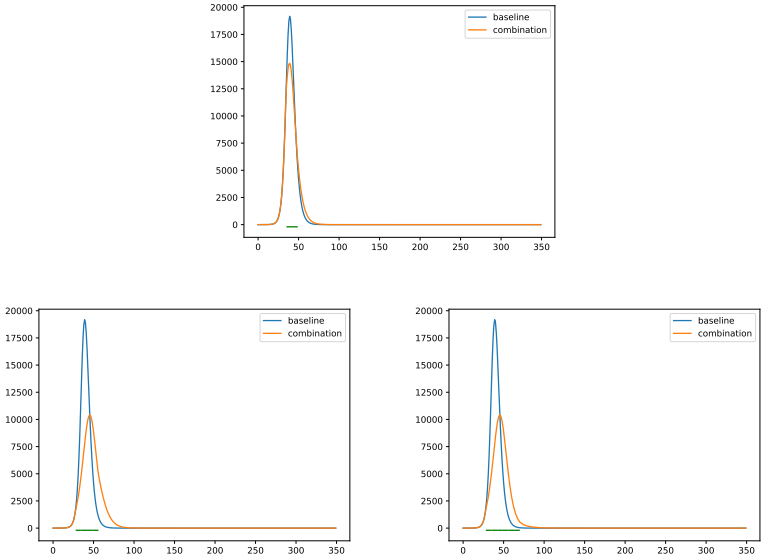


Figure 7.7: Epidemic trajectories for the Barnsley district with R_0=2.4. We show a baseline epidemic curve (blue) and an epidemic curve that depicts the impact of the school closure policy (orange). The duration of the school closure period is annotated with a green horizontal line. We show results for three budgets: 2 weeks (top), 4 weeks (bottom left) and 6 weeks (bottom right).

In the next section, we will empirically evaluate PPO, with respect to the established ground truth.

7 Evaluate PPO with respect to the ground truth

We repeat the experiment for which we established a ground truth (i.e., $R_0 \in \{1.8, 2.4\}$, 10 districts and $\acute{b} \in \{2, 4, 6\}$) and learn a policy using PPO, in the stochastic model. For each experimental setting (i.e., the combination of a district, an R_0 value, and a school closure budget $\acute{b}$), we run PPO 5 times (5 trials), to assess the variance of the learning curve (i.e., total reward per episode). Each PPO trial is run for 500000 time steps. These

experiments demonstrate learning curves that are similar to the ones presented in Section 5, as shown in Appendix 12 ($R_0 = 1.8$) and Appendix 13 ($R_0 = 2.4$).

To compare each of the learned policies to its ground truth, we take the learned policy and apply it 1000 times in the stochastic model, which results in a distribution over model outcomes (i.e., attack rate improvements). We then compare this distribution to the attack rate improvement that was recorded for the ground truth. We show these results, for the setting with a school closure budget of 6 weeks, in Figure 7.8, and for the other settings in Appendix 14. These results show that PPO learns a policy that matches the ground truth for all districts and combinations of R_0 and $\mathcal{b}$.

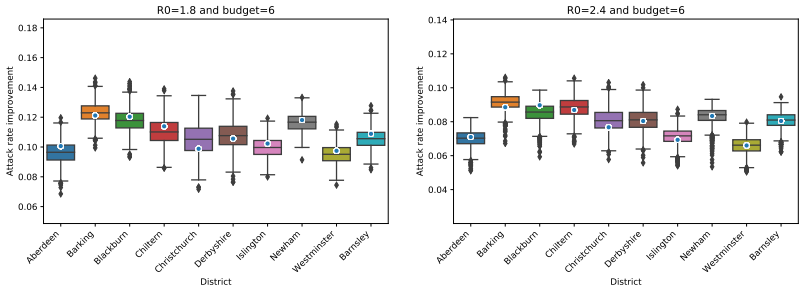


Figure 7.8: We compare the PPO results to the ground truth for $R_0 \in \{1.8, 2.4\}$ and $\mathcal{b} = 6$. Per district, we show a box plot that denotes the outcome distribution that was obtained by simulating the policy learned by PPO 1000 times. On top of this box plot, we show the ground truth as obtained in Section 6.1 (blue scatter).

Note that for these experiments, we use the same hyper-parameters for PPO that were introduced in Section 5. This demonstrates that, for different values of R_0 and for different census compositions, which induce a significant change in dynamics in the epidemic model, these hyper-parameters work well. This indicates that these hyper-parameters are adequate to be used for different variations of the model.

In this section, we compare a proxy to the ground truth (that has been found through an exhaustive policy search) to a policy learned by PPO, a deep reinforcement learning algorithm. This allows us to empirically validate that PPO converges to the optimal policy. This experimental validation is important, as it demonstrates the potential of deep reinforcement learning algorithms to learn policies in the context of complex epidemiological models. This indicates that it is possible to learn in even more complex stochastic models with large action spaces, for which it is impossible to compute a proxy to the ground truth. In Section 9, we investigate such as setting, where we aim to learn a joint policy for a set of agents, using deep multi-agent reinforcement learning.

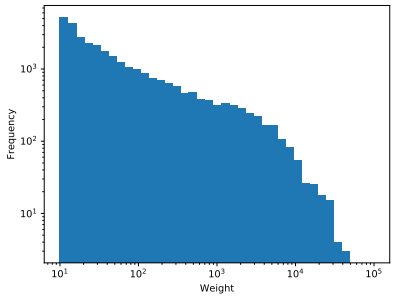


Figure 7.9: We show a histogram that visualizes the weight distribution of the commute flux matrix, where the x-axis represents the weight (log scale) and the y-axis (log scale) denotes the frequency.

8 Analysing the mobility network to partition

To investigate the collaborative nature of school closure policies, we apply deep multi-agent reinforcement learning algorithms. In our model, we have 379 agents, one for each district, as agents represent the district for which they can control school closure. As the current state-of-the-art of deep multi-agent reinforcement learning algorithms is limited to deal with ≈ 10 agents [Hernandez-Leal et al., 2019], we thus need to partition our model into smaller groups of agents, such that deep multi-agent reinforcement learning algorithms become feasible. To this end, we analyse the mobility graph that connects the different districts, to detect communities in this network of districts [Fortunato, 2010].

We remind the reader that we connect the different model districts via a mobility dataset that describes the daily commuting flow between the districts in Great-Britain (see Section 4 of Chapter 6).

This dataset consists out of a matrix $\mathcal{M}$, with elements $\mathcal{M}_{ij}$ that encode the amount of commute from district i to district j, for all of the districts in our model. In this matrix, $\mathcal{M}_{ij} = 0$ when $i = j$, as we do not consider commute within the same district. We show a histogram of the weight distribution of $\mathcal{M}$ in Figure 7.9, which shows that this distribution exhibits a long-tail trend.

Given $\mathcal{M}_{ij} \geq 0$, we can define a directed commute graph (Definition 32).

Definition 32: Commute graph

For a commuting matrix $\mathcal{M}$ that describes the mobility flux between a set of districts $\mathcal{D}$, we define a commute graph,

$$G_c = \langle V_c, A_c \rangle, \tag{7.8}$$

where V_c is the set of vertices, with a vertex for each of the districts in $\mathcal{D}$, and A_c is the adjacency matrix that specifies the vertices that are connected:

$$(A_c)_{ij} = \begin{cases} 1, & \mathcal{M}_{ij} > 0 \\ 0, & \mathcal{M}_{ij} = 0 \end{cases} \tag{7.9}$$

Each pair of connected vertices i and j has a weight $\mathcal{M}_{ij}$.

For our mobility matrix, this results in a graph where 0.725% of all possible edges are connected.

To start our analysis, we consider the vertex strength distribution, which is the analogue to a degree distribution in a weighted graph. The strength of a vertex $i \in V_c$, when we consider both incoming and outgoing edges, is defined as [Barrat et al., 2004]:

$$s_i = \sum_{j \in V_c} (A_c)_{ij} \cdot \mathcal{M}_{ij} + (A_c)_{ji} \cdot \mathcal{M}_{ji}. \tag{7.10}$$

We show the strength distribution for our commute graph in Figure 7.10. This figure shows the strength distribution for different weight thresholds, by using the threshold function $t(\mathcal{M}_{ij}, t_c)$ that censors any commute flux below the threshold t_c:

$$t(\mathcal{M}_{ij}, t_c) = \begin{cases} \mathcal{M}_{ij} & \text{if } \mathcal{M}_{ij} \geq t_c \\ 0 & \text{if } \mathcal{M}_{ij} < t_c \end{cases} \tag{7.11}$$

This figure confirms that while there is a large amount of connections with low commute weights (as is shown in Figure 7.9), these connections contribute little to the shape of the strength distribution. This indicates that the commute graph consists of components that are weakly connected and that there is the potential to partition the graph.

To detect communities in the commute graph, we used the Leiden algorithm [Traag et al., 2019]. This algorithm searches for communities that maximize the modularity (Definition 33). The Leiden algorithm extends the Louvain algorithm[5] [Blondel et al., 2008], as it adds guarantees that the identified communities are well-connected.

[5]The Louvain algorithm is another well known algorithm for community detection.

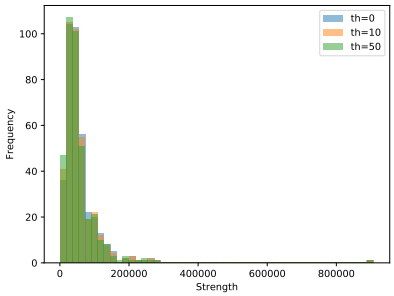


Figure 7.10: Histogram that visualizes the strength distribution of the commute graph, for different thresholds: $\{0, 10, 50\}$. The x-axis represents the strength and the y-axis denotes the frequency.

Definition 33: Modularity

For a set of communities in a weighted directed graph $G = \langle V, A \rangle$, with the set of vertices V and adjacency matrix A, we define modularity [Leicht and Newman, 2008; Reichardt and Bornholdt, 2006]:

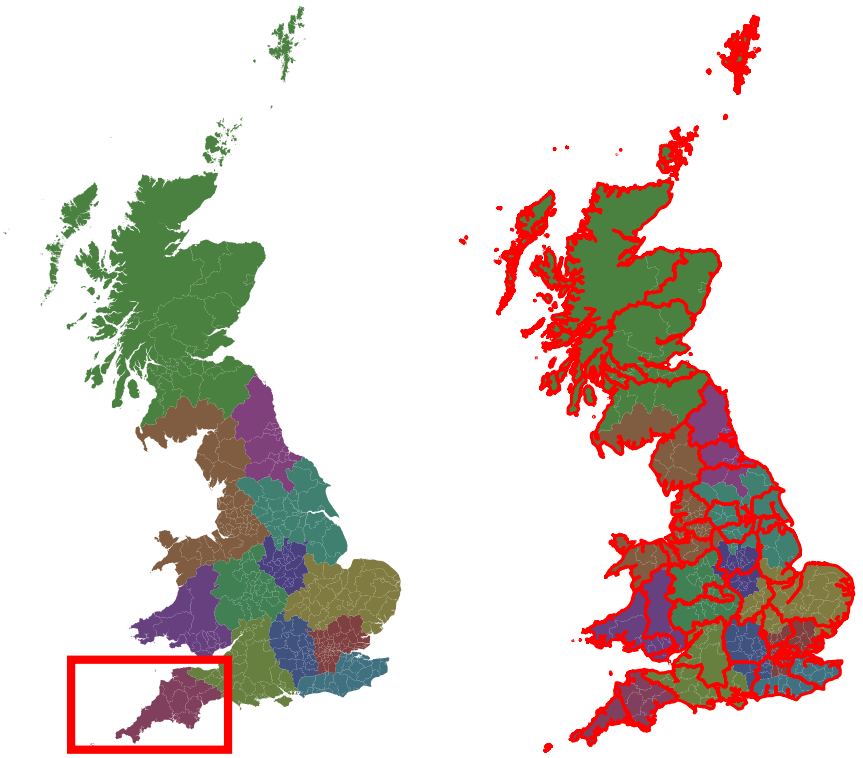
$$\sum_{i,j \in V} \left(A_{ij} - r \frac{k_i^{\text{out}} k_j^{\text{in}}}{m} c(i,j) \right), \tag{7.12}$$

where m is the sum of edge weights, k_i^{out} and k_i^{in} are respectively the weighted out-degree and in-degree of node i, the boolean function $c(i,j)$ specifies whether node i and j are in the same community, and r is a resolution parameter.

We used the Leiden algorithm, with the default[6] resolution parameter $r = 1$, for 100 iterations and we observed that the algorithm converged after 3 iterations. We found a partition of which we demonstrated the robustness (p-value ≤ 0.001) using a bootstrapping approach presented in [Radivojević and Grujić, 2017]. Furthermore, by rendering this partition on top of the map of Great Britain, as is shown in Figure 7.11 (left panel), we show that the districts belonging to the same community are close to each other geographically,

[6] In Section 11, we discuss the effect of different resolution parameters on the resulting number of communities in our mobility graph.

as we would expect. Moreover, when we overlay the NUTS-2 administrative regions[7] on the partitioning (Figure 7.11, right panel), we observe that our partitioning scheme mostly overlaps with the NUTS-2 regions, which indicates that Leiden algorithm produces a sensible partitioning.



(a) (b)

Figure 7.11: We show the communities, that resulted from applying the Leiden algorithm, on the map of Great Britain. The left panel shows all administrative districts colour-coded by the community they belong to, and the right panel adds the borders of the NUTS-2 administrative regions. We annotate the Cornwall-Devon community on the left panel with a red rectangle.

[7]NUTS (Nomenclature of Territorial Units for Statistics) is a geocode standard constructed by Eurostat to reference the subdivisions of European countries (https://ec.europa.eu/eurostat/web/nuts). NUTS-2 is the second level and corresponds to basic regions for the application of regional policies.

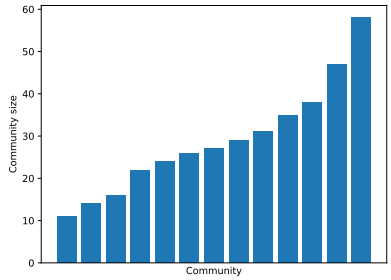


Figure 7.12: Bar chart that shows the distribution of community sizes, as retrieved by the Leiden algorithm.

However when looking at the bar chart in Figure 7.12, it is clear that the size of the majority of the communities is still too large to make multi-agent reinforcement learning feasible. Therefore, in the next section, we conduct our multi-agent reinforcement learning experiments in the community with 11 districts, to which we will refer as the Cornwall-Devon community (see Figure 7.11), as it is comprised of the Cornwall and Devon NUTS-2 regions. In Section 11 we will discuss possible ways to deal with larger communities.

9 Multi-agent reinforcement learning

We now examine whether there is an advantage to consider the collaboration between districts when designing school closure policies. We conduct an experiment in our epidemiological model with 379 districts, and attempt to learn a joint policy to control the districts in the Cornwall-Devon community (details Section 8). To this end, we assign an agent to each of the 11 districts of the Cornwall-Devon community, and use a multi-agent reinforcement learning approach to learn a joint policy.

We remind the reader, that we refer to the state space of one patch p as $\mathcal{S}_p$, as detailed in Section 3. The state space of the MDP, $\mathcal{S}$, corresponds to the aggregation of the state space of each patch that we attempt to control:

$$\mathcal{S} = \bigtimes_{p \in \mathcal{P}^c} \mathcal{S}_p, \tag{7.13}$$

where $\mathcal{P}^c$ is the set of patches we attempt to control. In this experiment, $\mathcal{P}^c$ corresponds to the 11 districts in the Cornwall-Devon community.

In order to learn a joint policy, we need to consider an action space that combines the actions for each district $p \in \mathcal{P}^{\mathsf{c}}$ that we attempt to control. This results in a joint action space with a size that is exponential with respect to the number of agents. To approach this problem, we use a Proximal Policy Optimization agent that controls multiple districts simultaneously. To this end, we use a custom policy network that gets as input the combined model state of each district $p \in \mathcal{P}^{\mathsf{c}}$ (Equation 7.13), and as a result, the input layer has $17 \cdot |\mathcal{P}^{\mathsf{c}}|$ input units[8]. In contrast to the single-district PPO, that was introduced in Section 5, the output layer of the policy network of this agent has a unit for each district that we attempt to control. Again, each output unit is passed through a sigmoid activation function, and hence corresponds to the probability of closing the schools in the associated district. Similar to the single-district PPO, each hidden layer uses the hyperbolic tangent activation function. The value network has the same architecture for the input layers and hidden layers, but only has a single output unit that represents the value for the given state. We will refer to this agent as *multi-district-PPO*.

We conduct experiments for $R_0 = 1.8$ (i.e., moderate transmission potential) and $R_0 = 2.4$ (i.e., high transmission potential), and we consider a school closure budget of 6 weeks, i.e., $\mathcal{b} = 6$. We run multi-district-PPO 5 times, to assess the variance of the learning signal, for $5 \cdot 10^6$ time steps, and we show the learning curves in Figure 7.13. These learning curves demonstrate a stable and steady learning process, for $R_0 = 1.8$ the reward curve is still increasing, while for $R_0 = 2.4$ the reward curve indicates that the learning process has converged.

To investigate whether these *joint policies* provide a collaborative advantage, we compare it to the aggregation of single district policies, to which we will refer as the *aggregated policy*. To construct this aggregated policy, we learn a distinct school closure policy for each of the 11 districts in the Cornwall-Devon community, using PPO, following the same procedure as in Section 7. To evaluate this aggregated policy, we execute the distinct policies simultaneously. For the districts that are not controlled (both for the joint and aggregated policy) we keep the schools open for all time steps. For both $R_0 = 1.8$ and $R_0 = 2.4$, respectively, we simulate the joint and the aggregated policy 1000 times, and we show the attack rate improvement distribution in Figure 7.14. These results corroborate that there is a collaborative advantage when devising school closures policies, for both $R_0 = 1.8$ and $R_0 = 2.4$. However, the improvement is most significant for $R_0 = 1.8$. We conjecture that this difference is due to the fact that there is less flexibility when the transmission potential of the epidemic is higher, since there is less time to act. This conjecture is supported by our earlier analyses in Section 6, that show that the top 10

[8] As detailed in Section 3, the state of the epidemiological model is comprised of the SEIR values for each age group (16 state variables), along with the remaining budget for closing schools in the district. The combined state space thus is comprised of 17 state variables.

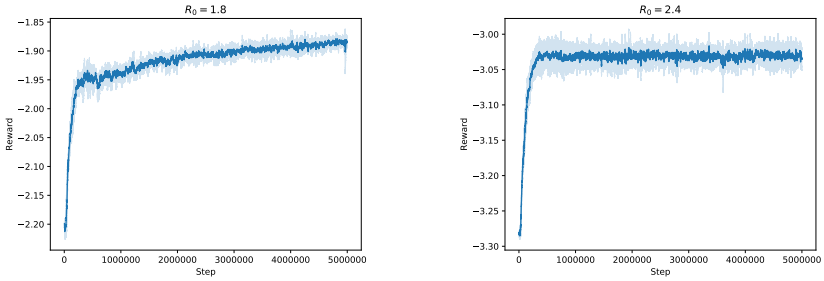


Figure 7.13: We show the reward curves (i.e., total reward per episode) for multi-district-PPO for $R_0 = 1.8$ (left panel) and $R_0 = 2.4$ (right panel). The reward curves are visualized using a rolling window of 100 steps. The shaded area shows the standard deviation of the reward signal, over 5 multi-district-PPO runs.

policies are more diverse for $R_0 = 1.8$, and more concentrated towards the peak day for $R_0 = 2.4$. Although, we observe an improvement when a joint policy is learned, it remains challenging to interpret deep multi-agent policies, and we discuss in Section 11 possible direction for future work with respect to explainable multi-agent reinforcement learning.

In this analysis, where we have a limited number of actions per agent, the use of multi-district-PPO proved to be successful. However, the use of more advanced multi-agent reinforcement learning methods is warranted when a more complex action space is considered. For this reason, we also investigated the recently introduced QMIX [Rashid et al., 2018] algorithm. We searched for hyper-parameters to optimize QMIX's performance, but the resulting learning curve proved to be quite unstable. Next to QMIX, there are other algorithms (e.g., Counterfactual multi-agent policy gradients [Foerster et al., 2018], Actor-Attention-Critic for Multi-Agent Reinforcement Learning [Iqbal and Sha, 2019] and Deep coordination graphs [Böhmer et al., 2019]) of interest to epidemiological decision making. In particular, we discuss the attention-based multi-agent reinforcement learning algorithms (e.g., Actor-Attention-Critic for Multi-Agent Reinforcement Learning [Iqbal and Sha, 2019]) as a direction for future work in Section 11.

We conducted our experiments in the setting of school closures, and our findings are of direct relevance with respect to the mitigation of pandemic influenza. Furthermore, our novel approach to investigate the collaborative nature of prevention strategies as a multi-agent reinforcement learning problem, can be applied to other epidemiological settings, as we discuss in Section 11.

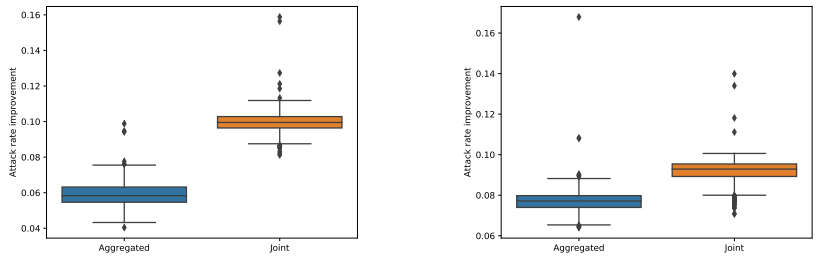


Figure 7.14: We compare the simulation results of the aggregated policy (blue) and the joint policy (orange) for $R_0 = 1.8$ (left panel) and $R_0 = 2.4$ (right panel). For both distributions (i.e., aggregated versus joint) , we show a box plot that denotes the outcome distribution that was obtained by simulating the respective policy 1000 times.

10 Reward function to investigate policies to shift the epidemic's peak day

As stated earlier (see Section 1), school closures policies can be used to serve two distinct objectives, on the one hand, to reduce the attack rate (Definition 19), and on the other hand, to delay the peak of the epidemic (i.e., to shift the peak day). We show an example of a peak day shift, through means of school closures, in Figure 7.15.

To shift the peak day, we require a delayed reward function, that returns the peak day at the end of the episode. The concept of an epidemic's peak day is straightforward in the case of an epidemic curve with a single peak. However, by closing schools, we can change the epidemic such that it is split in two distinct peaks, as shown in Section 5.2 of Chapter 6. To take this into account in our reward function, we first detect all local maxima in the epidemic curve, and select the first peak of which the normalized peak count exceeds the threshold t. We determine these local maxima by determining the first and second derivative of the epidemic curve. As we have a stochastic epidemic model, this epidemic curve will not be smooth and we therefore determine these derivatives using a Savitzky-Golay filter [Savitzky and Golay, 1964]. We show an example of this procedure in Figure 7.16 and formalize the reward function in Definition 34.

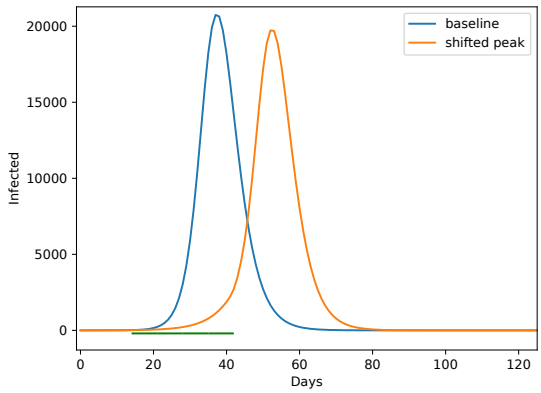


Figure 7.15: We show a baseline epidemic curve (blue), and an epidemic curve where schools were closed for 4 weeks (orange), to demonstrate a shift of the epidemic's peak day. The period that schools were closed is annotated with a green bar on top of the days axis.

Definition 34: Reward function to shift the peak day

The reward function to shift the peak day is:

$$R_{\text{PD}}(s, a, s') = \begin{cases} 0, & s' \text{ is a non-terminal state} \\ 0, & s' \text{ is a terminal state, but no peak occurred} \\ \text{peak\_days}(s')[0], & s' \text{ is a terminal state, and a peak occured,} \end{cases} \tag{7.14}$$

where peak_days(s') returns the local maxima using the first and second derivative of the epidemic curve as obtained through a Savitzky-Golay filter.

11 Discussion

In this chapter, we demonstrate the potential of deep reinforcement learning in the context of epidemiological decision making by conducting experiments that show that the Proximal Policy Optimization (PPO) algorithm converges to the optimal policy. Furthermore, we show that the impact of the census composition on school closure policies is limited.

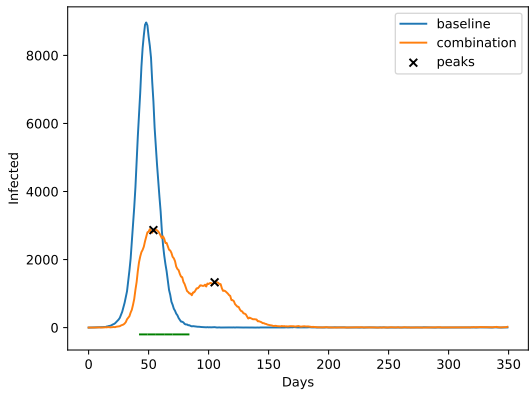


Figure 7.16: We show a baseline epidemic curve (blue), and an epidemic curve where schools were closed for 6 weeks (orange). The period that schools were closed is annotated with a green bar on top of the days axis. For the orange epidemic curve, we show its two epidemic peaks.

Finally, we investigate and show that there is a collaborative advantage when devising school closures policies, by formulating the hypothesis as a multi-agent problem.

The work conducted throughout this chapter indicates that there is the potential to use reinforcement learning in the context of complex stochastic epidemiological models. For future work, it would be interesting to investigate how well these algorithms scale to even larger state and/or action spaces. To increase the scalability, a possible research direction is the use of convolutional neural networks, instead of the multi-layer perceptron networks we use in this chapter. However, convolutional neural networks assume that the input data is structured in a grid-wise fashion [Collobert and Weston, 2008], which is not typically the case for geographic districts. To circumvent this assumption, the use of graph convolutional neural networks could be explored [Defferrard et al., 2016].

Another important concern is to scale these reinforcement learning methods to epidemiological models with a greater computational burden. In this dissertation we constructed a custom model where we attempt to balance between model complexity and computational efficiency. However, constructing such models is cumbersome and time-consuming, and the resulting model is specifically tailored to address one particular research question (in our case the evaluation of school closure policies). An alternative to such custom models is the use of individual-based models, as such models can be easily configured to

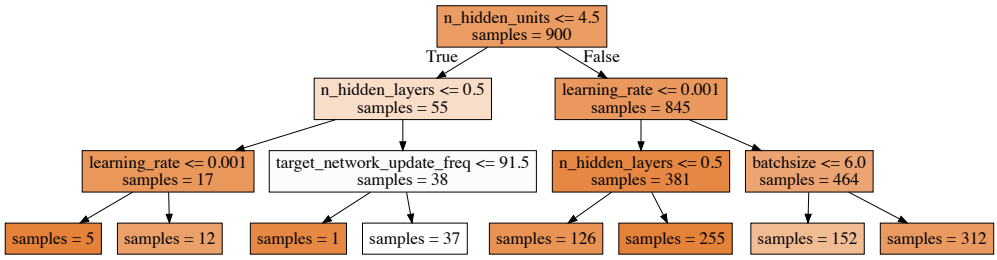
approach a variety of research scenarios. However, the computational burden that is associated with individual-based models (see Chapter 4) complicates the use of reinforcement learning methods, such as DQN and PPO. To this end, it would be interesting to devise methods to automatically learn a surrogate model from the individual-based model, such that the reinforcement learning agent can learn in this computationally leaner surrogate model [Willem et al., 2014].

We performed a hyper-parameter sweep using Latin hypercube sampling ($n = 1000$) for both DQN and PPO [Stein, 1987]. On the one hand, this Latin hypercube sampling sweep found a good fit for the hyper-parameters, on the other hand, it provides insights in the effect of the different hyper-parameters on the learning performance, as shown in Figure 7.17. As we demonstrated in Section 5, both DQN and PPO exhibit similar performance for their top hyper-parameters. However, considering the complete space of hyper-parameters, DQN seems to be less sensitive to the choice of hyper-parameters than PPO. Nonetheless, our extensive evaluation of PPO shows that once a good set of hyper-parameters has been identified, this set works well for different variations of the epidemiological model, e.g., with respect to census composition and reproductive number.

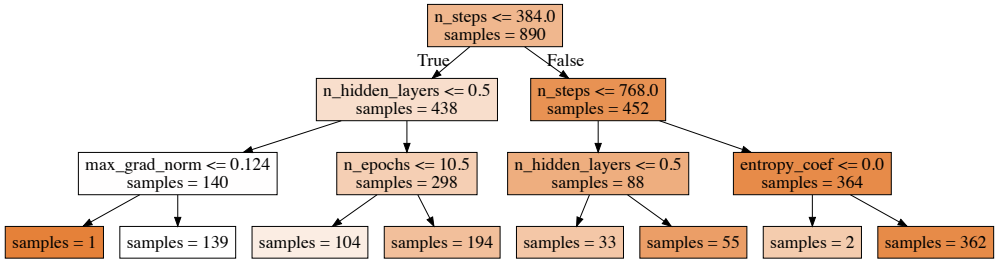
While we show in this chapter that deep reinforcement learning algorithms can be used to learn optimal mitigation strategies, the interpretation of such policies remains challenging. This is especially the case for the multi-agent setting we considered. To this end, further research into explainable reinforcement learning, both in a single-agent and multi-agent setting, is warranted.

In Section 8, we partitioned the districts based on an analysis of the mobility network. Through this analysis, we were able to obtain a community of 11 districts, that made the use of multi-agent reinforcement learning algorithms feasible. However, the other communities proved too large, rendering the methods we propose unachievable. We foresee two different research directions to address this problem. Firstly, we could attempt to partition the communities to a more fine-grained resolution[9] or to group the most tightly coupled districts into super-districts. Secondly, we could consider an algorithm that incorporates the network structure that connects the agents in the learning process. In this regard, we acknowledge the recent Cooperative Prioritized Sweeping multi-agent reinforcement learning algorithm [Bargiacchi et al., 2020], that takes into account the connections between agents, and scales to many agents. While this algorithm is currently limited to tabular settings, we expressed interest to the authors of this work to extend this algorithm to continuous state spaces.

[9] There are different ways to approach this deeper partitioning. Firstly, by finding a resolution parameter r for the Leiden algorithm that provides a more fine-grained partitioning. Note that this approach was not successful on our mobility network, but it could prove beneficial for other settings. Secondly, by dividing the community graph using hierarchical clustering techniques.



(a) DQN



(b) PPO

Figure 7.17: We show the CART decision trees [Steinberg, 2009] of the Latin hypercube sampling parameter sweep for DQN (a) and PPO (b). The tree nodes' hue indicates the performance of that branch of hyper-parameters, i.e. darker is better.

Furthermore, in order to address problems with a larger state/action space and to scale to a larger number of agents, the use of attention-based multi-agent reinforcement learning algorithms could be explored [Jiang and Lu, 2018; Iqbal and Sha, 2019; Vaswani et al.]. Based on this mechanism, algorithms can be applied on a graph of agents, which is either assumed [Khan et al., 2019; Jiang et al., 2018] or learned [Liu et al., 2019].

8 | General discussion

> Le but de la discussion ne doit pas être la victoire, mais l'amélioration.
>
> Joseph Jouber

In this chapter, we summarize the contributions that we presented throughout this dissertation. Next, we discuss the valorisation potential of our research, both with respect to the control of epidemics, and the potential of our algorithms and methods beyond epidemiological decision making. Finally, we discuss different opportunities for future work, and we express the aspiration to use reinforcement learning to react to emerging infectious diseases in real-time.

1 Contributions

Our first contribution concerns the decision making problem where an optimal prevention strategy needs to be identified from a set of alternatives, where we assume that these prevention strategies can be evaluated in a stochastic individual-based epidemiological model. Due to the computationally intensive nature of such models, it is important to minimize the number of model evaluations required to make a decision. In this regard, in Chapter 4, **we formulate this decision making problem as a multi-armed bandit problem, where the different prevention strategies are modelled as bandit arms.** We demonstrate that it is possible to efficiently identify the optimal prevention strategy in this multi-armed bandit setting, by using fixed-budget best-arm identification algorithms.

Moreover, we show that by using epidemiological modelling theory, we can derive knowledge on the outcome distribution of the epidemiological model, and exploit this knowledge using Bayesian best-arm identification algorithms. Following this Bayesian approach, we experimentally show that it is possible to identify the optimal prevention strategy 2-to-3 times faster compared to the uniform sampling method, the predominant technique used for epidemiological decision making in the literature. Furthermore, we show that the uncertainty distribution constructed by Bayesian best-arm identification algorithms can be used to inform decision makers about the confidence of an arm recommendation. From a technical perspective, we enable Bayesian best-arm identification algorithms (i.e., Thompson sampling and BayesGap) to be used on a Gaussian reward distribution with unknown variance, by using a t-distributed posterior. For BayesGap, we formally prove that using this posterior yields a probability of simple regret that asymptotically reaches the exponential lower bound of [Hoffman et al., 2014].

Our second contribution (Chapter 5) **generalizes this decision making framework into an anytime m-top exploration setting**, which has two significant advantages. Firstly, while the best-arm identification approach only returns the best prevention strategy, m-top exploration algorithms can return the m best prevention strategies, providing public health scientists with more flexibility. Secondly, the use of an anytime algorithm removes the need for decision makers to choose a computational budget upfront. This is important, as choosing a computational budget upfront can be challenging, especially for computationally intensive models. We demonstrate this generalized decision making framework with the state-of-the-art anytime m-top exploration algorithm AT-LUCB.

As AT-LUCB is an Upper Confidence Bound (UCB) variant, it is challenging to inform it with prior knowledge. Therefore, **our third contribution is a new Bayesian anytime m-top algorithm: Boundary Focused Thompson Sampling** (Chapter 5). We demonstrate that Boundary Focused Thompson Sampling (BFTS) outperforms AT-LUCB for a set of benchmarks, and show that BFTS significantly outperforms AT-LUCB in the context of decision making in epidemics, by introducing a new and challenging benchmark problem. We further establish BFTS's potential in a bandit setting with Poisson reward distributions, to show that BFTS is able to handle skewed and high-variance (i.e., challenging) reward distributions. Finally, we perform a formal Bayesian analysis of BFTS, to provide additional insights in BFTS' exploration strategy, which confirms that this strategy is well-grounded.

While our first three contributions are in the realm of multi-armed bandits, the next two contributions concern the use of reinforcement learning techniques to learn adaptive mitigation policies. In this regard, we used reinforcement learning to study and optimize school closure policies in the context of pandemic influenza.

Our fourth contribution regards **a new epidemiological meta-population model to study school closure policies with reinforcement learning.** This model consists out

of a set of interconnected patches, where each patch corresponds to an administrative region in Great Britain and is internally represented by a SEIR compartment model with four age groups. On the one hand, we designed this model such that it is sufficiently fine-grained to evaluate school closure policies. On the other hand, the model was engineered to be computationally efficient such that it can be used in combination with the state-of-the-art of reinforcement learning algorithms, that to date suffer from sample inefficiency. We conducted experiments to assess the model's performance, of which one experiment demonstrated the model's capability to reproduce the 2009 pandemic in Great Britain. While this new meta-population model addresses a specific research questions, it can be extended to investigate other preventive measures in the context of pandemic influenza (i.a., vaccines, antiviral drugs), by changing the structure of the compartment model. Furthermore, while this particular model was constructed to model influenza pandemics in Great Britain, the modelling methodology can be used to construct models for other geographic regions and/or scales, as long as there is census data available for each of the patches and a model (or dataset) that expresses the mobility network between the different patches. Moreover, the methods we introduced to achieve the computational efficiency in this model can be used to construct multi-patch models for other pathogens (e.g., Ebola, arboviruses). Such computationally efficient models could render the use of reinforcement learning techniques attainable to study adaptive mitigation policies for a wider range of pathogens.

Our fifth contribution concerns the introduction of a reinforcement learning approach to learn adaptive mitigation policies in a complex epidemiological model. Firstly, we experimentally evaluate the use of Proximal Policy Optimization (PPO), a deep reinforcement learning algorithm, in a single district setting. We establish a ground truth and show that PPO converges to this optimal policy. Furthermore, through this evaluation, we show that the impact of the census composition on school closure policies is limited. Secondly, we present a new approach to investigate whether there is a collaborative advantage when devising mitigation policies. We do this by formulating this research question as a multi-agent decision making problem, and we solve this decision problem by using deep multi-agent reinforcement learning algorithms. In this new framework, we show that there is a collaborative advantage when devising school closures policies. These findings are of direct relevance with respect to the mitigation of pandemic influenza. Furthermore, this novel deep multi-agent reinforcement learning approach, has the potential to study other epidemiological settings.

2 Dissemination and valorisation

The research in this dissertation was funded by an FWO[1] grant for strategic basic research. The FWO allocates this grant to challenging and innovative research, which may in the longer term lead to innovative applications with economic and/or societal added value. Therefore, in this section, we discuss the potential for the dissemination and valorisation of our work.

In this dissertation, we contribute different reinforcement learning techniques to support the decision making process to mitigate epidemics of infectious diseases. Improving the response to epidemics is important, as outbreaks of infectious diseases cause a great societal and economic burden. Our contributions add value on three main fronts. Firstly, by introducing the multi-armed bandit decision framework, we reduce the number of model evaluations that are necessary to come to a solution. Minimizing the number of required model evaluations reduces the total time required to evaluate a given set of preventive strategies. This renders the use of individual-based models attainable in studies where it would otherwise not be computationally feasible, opening the prospect for novel directions to investigate the control of epidemics. Additionally, reducing the number of model evaluations will free up computational resources in studies that already use individual-based models, capacitating researchers to explore a wider set of model scenarios. This is important, as considering a wider range of scenarios increases the confidence about the overall utility of preventive strategies. Secondly, we show that deep reinforcement learning techniques can be used to learn adaptive mitigation strategies in complex epidemiological models. As we showed, learning such adaptive strategies can improve the utility of a prevention strategy, thereby reducing the societal and economic burden that an epidemic induces. Moreover, these techniques can be used to further optimize prevention strategies, for example to reduce the cost of a prevention strategy. Thirdly, next to the mitigation of epidemics, there is a great potential for the anytime m-top exploration to support decision makers with complex societal challenges in general. These decision making processes are often guided by intricate simulation models, to evaluate a set of alternative policies that can be modelled as bandit arms. Our new Bayesian algorithm, Boundary Focused Thompson sampling (BFTS), allow decision makers to introduce prior knowledge into the decision making process. Such prior information is readily available in many practical applications, and enables faster learning and the ability to report a confidence measure when a decision is made. In this regard, we believe that BFTS is a promising algorithm to support decision makers across different research fields.

Finally, our experiments show that reinforcement learning techniques can be used to investigate mitigation strategies off-line to evaluate and come up with new mitigation

[1] Fonds voor Wetenschappelijk Onderzoek - Vlaanderen, Research Foundation – Flanders

protocols. However, we believe that there is a great potential to use reinforcement learning to support decision makers in a real-time fashion, which is especially important in the context of emerging infectious diseases. We discuss this prospect in Section 3.

3 Future work

We believe that there is a great potential to use reinforcement learning in the context of emerging infectious diseases to support policy makers in a real-time fashion.

The control of emerging epidemics is notoriously challenging, as there is typically a great deal of uncertainty associated with the pathogen's characteristics and the course that the epidemic will take [Metcalf and Lessler, 2017]. As we argue throughout this dissertation, potential mitigation strategies thus need to be evaluated using reliable epidemiological models. However, modelling epidemics when they emerge is difficult, as the number of infected individuals is typically limited at this phase of the epidemic, which convolutes the fitting of such models [Britton and Scalia Tomba, 2019].

Epidemiological models are routinely fit on case statistics (e.g., incidence data), as these are readily available when an epidemic emerges[Britton and Scalia Tomba, 2019]. However, due to the rapid evolution of many viral pathogens, and as the acquisition of genetic virus sequences is becoming increasingly abundant and accessible due to recent advances in sequencing technologies [Grubaugh et al., 2019], the potential to spatial epidemiological models from virus genomes rises.

We argue that to support policy makers in real time, two main methodological advances are required. On the one hand, a method is needed to continuously update a geospatial epidemiological model whilst an epidemic unravels, by using virus genomes as an additional epidemiological marker, taking into account the uncertainty of the ongoing epidemic. On the other hand, given this model that captures the uncertainty of the epidemic, a reinforcement learning agent is necessary that can suggest decisions under uncertainty to policy makers and can use the feedback of the policy makers to further guide the learning process.

Another important direction for future work is to extend the proposed techniques to a multi-objective setting. As mentioned in Chapter 4, we did some preliminary research to use the Interactive Thompson sampling algorithm [Roijers et al., 2017] in a pure-exploration multi-armed bandit setting, such that we can learn about the environment (i.e., the decision problem) and the user's preferences simultaneously. We believe that further research in this direction would be useful. In the context of stateful multi-objective reinforcement algorithms there are two categories of algorithms [Vamplew et al., 2011]: single-policy and multi-policy algorithm. On the one hand, single-policy algorithms try to learn a single optimal policy for a known utility function. On the other hand, multi-policy algorithms attempt to approximate the Pareto front. Both approaches are useful in the context of

epidemiological decision making, and an experimental validation of the state-of-the-art of multi-objective reinforcement algorithms is warranted.

Finally, in this dissertation, we show that multi-agent reinforcement learning techniques can be used to investigate intricate hypotheses. However, the state-of-the-art in deep multi-agent reinforcement learning is only able to consider a limited number of agents. As discussed in Chapter 7 we believe that there is great potential to construct algorithms that incorporate the network structure between the different agents in the learning process. An interesting research direction in this regard is the recent Cooperative Prioritized Sweeping multi-agent reinforcement learning algorithm [Bargiacchi et al., 2020], that takes into account the connections between agents and scales to many agents. While this algorithm is currently limited to tabular settings, it would be interesting to consider extending it towards continuous settings, and evaluate this new algorithm in the context of epidemic control.

A | Appendices

1 BayesGap simple regret bound for t-distributed posteriors

In this section, we provide a proof for BayesGap's simple regret bound for Gaussians with unknown means and variances (see Chapter 4). To this end, we first introduce and prove three lemmas. Note that this proof is a novel contribution.

In Section 4.4 of Chapter 4, we specified problem-specific bounds. To remind the reader, given our posteriors (Equation 4.12), we have:

$$\begin{aligned} U_k^{(t)} &= \mathbb{E}\left[\pi_{\mathcal{T}}^{(t-1)}\right]_k + \kappa\sqrt{\mathbb{V}\left[\pi_{\mathcal{T}}^{(t-1)}\right]_k} \\ L_k^{(t)} &= \mathbb{E}\left[\pi_{\mathcal{T}}^{(t-1)}\right]_k - \kappa\sqrt{\mathbb{V}\left[\pi_{\mathcal{T}}^{(t-1)}\right]_k} \end{aligned} \tag{A.1}$$

where $\mathbb{E}\left[\pi_{\mathcal{T}}^{(t-1)}\right]$ is the mean, and $\mathbb{V}\left[\pi_{\mathcal{T}}^{(t-1)}\right]$ is the variance of the posterior of arm a_k at time step $t-1$, and κ is an exploration coefficient. Furthermore, we defined the arm-dependent ϵ-hardness (Definition 18):

$$H_{k,\epsilon} = \max\left(0.5(\delta_k^{(t))} + \epsilon), \epsilon\right). \tag{A.2}$$

Lemma 1

Consider a Jeffrey's prior $(\mu_k, \sigma_k^2) \sim \sigma_k^{-3}$ over the parameters of the Gaussian reward distributions. Then the posterior mean of arm k has the following non-standardized t-distribution at pull $n_k^{(t)}$:

$$\mu_k \mid \hat{\mu}_k^{(t)}, S_k^{(t)} \sim \mathcal{T}_{n_k^{(t)}}(\hat{\mu}_k^{(t)}, (n_k^{(t)})^{-1}\sqrt{S_k^{(t)}}),$$

where $n_k^{(t)}$ is the number of pulls for arm k, $\hat{\mu}_k^{(t)}$ is the sample mean and $S_k^{(t)}$ is the sum of squares.

Proof. This lemma was presented and proved by Honda and Takemura [2014]. □

Lemma 2

Consider a non-standardized t-distributed random variable $X \sim \mathcal{T}_\nu(\mu, \lambda)$ with variance $\sigma^2 = \frac{\nu}{\nu-2}\lambda^2$, $\nu > 2$ and $\kappa > 0$. The probability that X is within a radius $\kappa\sigma$ from its mean can be bounded as:

$$P(|X - \mu| < \kappa\sigma) \geq 1 - 2\frac{\sqrt{\nu(\nu-2)}}{\nu-1}\frac{C(\nu)}{\kappa}\left(1 + \frac{\kappa^2}{\nu}\right)^{-0.5(\nu-1)},$$

where:

$$C(\nu) = \frac{\Gamma(0.5\nu + 0.5)}{\Gamma(0.5\nu)\sqrt{\pi\nu}},$$

is the normalizing constant of a standard t-distribution.

Proof. Consider a random variable $Z \sim \mathcal{T}_\nu(0,1)$, $\nu > 2$ and $\kappa > 0$. Then the probability of Z being greater than $\kappa\sqrt{\frac{\nu}{\nu-2}}$ is:

$$
\begin{aligned}
P\left(Z > \kappa\sqrt{\frac{\nu}{\nu-2}}\right) &\overset{(1)}{=} \int_{\kappa\sqrt{\frac{\nu}{\nu-2}}}^{+\infty} \mathcal{T}_\nu(z \mid 0,1)dz \\
&= C(\nu)\int_{\kappa\sqrt{\frac{\nu}{\nu-2}}}^{+\infty} \left(1+\frac{z^2}{\nu}\right)^{-0.5(\nu+1)} dz \\
&\overset{(2)}{\leq} C(\nu)\int_{\kappa\sqrt{\frac{\nu}{\nu-2}}}^{+\infty} \frac{z}{\kappa\sqrt{\frac{\nu}{\nu-2}}} \left(1+\frac{z^2}{\nu}\right)^{-0.5(\nu+1)} dz \\
&= \frac{\sqrt{\nu-2}}{\kappa\sqrt{\nu}} C(\nu)\int_{\kappa\sqrt{\frac{\nu}{\nu-2}}}^{+\infty} z\left(1+\frac{z^2}{\nu}\right)^{-0.5(\nu+1)} dz \\
&= -\frac{\nu}{\nu-1}\frac{\sqrt{\nu-2}}{\kappa\sqrt{\nu}} C(\nu)\int_{\kappa\sqrt{\frac{\nu}{\nu-2}}}^{+\infty} -\frac{\nu-1}{\nu} z\left(1+\frac{z^2}{\nu}\right)^{-0.5(\nu+1)} dz \\
&\overset{(3)}{=} -\frac{\sqrt{\nu(\nu-2)}}{\nu-1}\frac{C(\nu)}{\kappa}\left(1+\frac{z^2}{\nu}\right)^{-0.5(\nu-1)}\Bigg|_{\kappa\sqrt{\frac{\nu}{\nu-2}}}^{+\infty} \\
&\overset{(4)}{=} \frac{\sqrt{\nu(\nu-2)}}{\nu-1}\frac{C(\nu)}{\kappa}\left(1+\frac{\kappa^2}{\nu-2}\right)^{-0.5(\nu-1)}
\end{aligned}
$$

The probability of Z being greater or equal than the lower bound $\kappa\sqrt{\frac{\nu}{\nu-2}}$ is the integral over its probability density function, starting from that lower bound (1). In the integral, we introduce a factor $\frac{z}{\kappa\sqrt{\frac{\nu}{\nu-2}}}$, which is greater or equal than 1 for the considered values of z (2). We then take note of the following derivative, and use this result to analytically solve the integral (3):

$$
\frac{d}{dx}\left(1+\frac{x^2}{\nu}\right)^{-0.5(\nu-1)} = -\frac{\nu-1}{\nu}x\left(1+\frac{x^2}{\nu}\right)^{-0.5(\nu+1)}
$$

Finally, we solve the primitive from $\frac{z}{\kappa\sqrt{\frac{\nu}{\nu-2}}}$ to infinity (4).

Next, we apply a union bound to obtain a lower bound on the probability that the magnitude of Z is smaller than $\kappa\sqrt{\frac{\nu}{\nu-2}}$:

$$\begin{aligned}
P\left(|Z| < \kappa\sqrt{\frac{\nu}{\nu-2}}\right) &= 1 - P\left(|Z| > \kappa\sqrt{\frac{\nu}{\nu-2}}\right) \\
&= 1 - \left[P\left(Z > \kappa\sqrt{\frac{\nu}{\nu-2}}\right) + P\left(Z < -\kappa\sqrt{\frac{\nu}{\nu-2}}\right)\right] \\
&= 1 - 2P\left(Z > \kappa\sqrt{\frac{\nu}{\nu-2}}\right) \\
&\geq 1 - 2\frac{\sqrt{\nu(\nu-2)}}{\nu-1}\frac{C(\nu)}{\kappa}\left(1 + \frac{\kappa^2}{\nu-2}\right)^{-0.5(\nu-1)}
\end{aligned}$$

In the third step, we use the fact that $P\left(Z > \kappa\sqrt{\frac{\nu}{\nu-2}}\right)$ is equal to $P\left(Z < -\kappa\sqrt{\frac{\nu}{\nu-2}}\right)$, as Z is a standardized t-distributed random variable.

Finally, consider $Z = \frac{(X-\mu)}{\lambda}$:

$$P\left(|X-\mu| < \kappa\sqrt{\frac{\nu}{\nu-2}}\lambda\right) \geq 1 - 2\frac{\sqrt{\nu(\nu-2)}}{\nu-1}\frac{C(\nu)}{\kappa}\left(1 + \frac{\kappa^2}{\nu-2}\right)^{-0.5(\nu-1)}$$

□

Lemma 3

Consider a K-armed bandit problem with budget T and K arms. Let $U_k^{(t)}$ and $L_k^{(t)}$ be upper and lower bounds that hold for all times $t \leq T$ and all arms $k \leq K$ with probability $1 - \delta_k^{(t)}$, and $n_k^{(t-1)}$ is the number of times arm k has been pulled at time $t-1$. Finally, let g_k be a monotonically decreasing function such that

$$U_k^{(t)} - L_k^{(t)} \leq g_k(n_k^{(t-1)}), \tag{A.3}$$

and,

$$\sum_{k=1}^{K} g_k^{-1}(H_{k,\epsilon}) \leq T - K. \tag{A.4}$$

We can then bound the simple regret $R^{(T)}$ (Definition 4) as:

$$P(R^{(T)} < \epsilon) \geq 1 - \sum_{k=1}^{K} \sum_{t=1}^{T} \delta_k(t)$$

Proof. First, we define $\mathcal{E}$ as the event in which every mean μ_k is bounded by its associated bounds (i.e., $U_k^{(t)}$ and $L_k^{(t)}$) for each time step.

$$\mathcal{E} := \forall k \leq K, \forall t \leq T : L_k^{(t)} \leq \mu_k \leq U_k^{(t)}$$

The probability of regret is equal to the probability of the event $\mathcal{E}$ occurring, as shown by Hoffman et al. [2014]. The probability of μ_k deviating from a single bound at time t is by definition $\delta_k(t)$. When applying the union bound, we obtain

$$P(\mathcal{E}) \geq 1 - \sum_{k=1}^{K} \sum_{t=1}^{T} \delta_k(t). \tag{A.5}$$

□

Theorem 2

Consider a K-armed Gaussian bandit problem with budget T and unknown variance, with a Jeffrey's prior $(\mu_k, \sigma_k^2) \sim \sigma_k^{-3}$ over the parameters of the Gaussian reward distributions. Let σ_G^2 be a generalization of that variance over all arms, and $U_k^{(t)}$ and $L_k^{(t)}$ respectively be the upper and lower bounds for each arm k at time t. The simple regret is then bounded as:

$$
\begin{aligned}
&P(R^{(T)} \leq \epsilon) \\
&\geq 1 - 2\sum_{k=1}^{K}\sum_{t=1}^{T} \frac{\sqrt{n_k(t)(n_k(t)-2)}}{n_k(t)-1} \frac{C(n_k(t))}{\kappa} \left(1 + \frac{\kappa^2}{n_k(t)-2}\right)^{-0.5(n_k(t)-1)} \\
&\geq 1 - O\left(KT\left(1 + \frac{\kappa^2}{\min\limits_{k,t} n_k(t)}\right)^{-0.5 \min\limits_{k,t} n_k(t)}\right),
\end{aligned}
$$

where,

$$\kappa = \sqrt{\frac{T-3K}{4\sigma_G^2 \sum_{k=1}^{K} H_{k,\epsilon}^{-2}}}.$$

Proof. According to Lemma 1, the posterior over the average reward is a t-distribution with scaling factor $\lambda_k(t) = \left(n_k^{(t)}\right)^{-1}\sqrt{S_k^{(t)}}$. Therefore, the difference between the lower and upper bounds is equal to:

$$
\begin{aligned}
U_k^{(t)} - L_k^{(t)} &= 2\kappa \mathbb{V}\left[\pi_{\mathcal{T}}^{(t-1)}\right]_k \\
&\overset{(1)}{=} 2\kappa \sqrt{\frac{n_k(t-1)}{n_k(t-1)-2}} \frac{\sqrt{S_k^{(t-1)}}}{n_k^{(t-1)}} \\
&= 2\kappa \sqrt{\frac{(s_k^{(t)})^2}{n_k(t-1)-2}},
\end{aligned}
$$

where $(s_k^{(t)})^2$ is the variance over n rewards for arm k and κ is a free parameter that will be chosen later. The standard deviation of a t-distribution is equal to $\sqrt{\frac{\nu}{\nu-2}}\lambda_k(t)$ for arm k at time t, where ν is the degrees of freedom and $\lambda_k(t)$ is the scaling factor of the t-distribution described in Lemma 1 (1). We generalize this standard deviation $s_k^{(t)}$ to σ_G, which is assumed to be representative for all arms. Note that, in Section 4.4 of

Chapter 4, we choose $\sigma_G^2 = \bar{s}_G^2$ to be the mean over all arm-specific variances obtained after the initialization phase.

We define the monotonically decreasing function

$$g_k(n) = 2\kappa\sqrt{\frac{\sigma_G^2}{n-2}}, \tag{A.6}$$

such that the condition,

$$U_k^{(t+1)} - L_k^{(t+1)} \leq g_k(n_k^{(t-1)}), \tag{A.7}$$

in Lemma 3 is satisfied.

Next, we compute the inverse of $g_k(x)$:

$$\begin{aligned} & 2\kappa\sqrt{\frac{\sigma_G^2}{g_k^{-1}(x)-2}} = x \\ \Leftrightarrow\ & \frac{1}{g_k^{-1}(x)-2} = \frac{x^2}{4\kappa^2\sigma_G^2} \\ \Leftrightarrow\ & g_k^{-1}(x) = \frac{4\kappa^2\sigma_G^2}{x^2} + 2 \end{aligned}$$

We restrict the sum of this function applied to the hardness over all arms k to be equal to $T-K$, which satisfies the last condition on g_k in Lemma 3:

$$\begin{aligned} & \sum_{k=1}^{K} g_k^{-1}(H_{k,\epsilon}) \\ & = \sum_{k=1}^{K} \frac{4\kappa^2\sigma_G^2}{H_{k,\epsilon}^2} + 2 \\ & = 4\kappa^2\sigma_G^2 \sum_{k=1}^{K} H_{k,\epsilon}^{-2} + 2K \\ & = T - K \end{aligned}$$

From the restrictions put on $g_k(x)$, we can derive κ as follows:

$$\begin{aligned} & 4\kappa^2\sigma_G^2 \sum_{k=1}^{K} H_{k,\epsilon}^{-2} + 2K = T - K \\ \Leftrightarrow\ & \kappa = \sqrt{\frac{T-3K}{4\sigma_G^2 \sum_{k=1}^{K} H_{k,\epsilon}^{-2}}} \end{aligned}$$

Finally, as the conditions on g_k are now satisfied, the simple regret bound can be obtained using Lemma 3 (L3) and the probability that the true mean is out of the arm-specific bounds $U_k^{(t)}$ and $L_k^{(t)}$, given in Lemma 2 (L2).

$$
\begin{aligned}
& P(R^{(T)} < \epsilon) \\
& \overset{(L\ 3)}{\geq} 1 - \sum_{k=1}^{K} \sum_{t=1}^{T} \delta_k(t) \\
& \overset{(L\ 2)}{=} 1 - 2 \sum_{k=1}^{K} \sum_{t=1}^{T} \frac{\sqrt{n_k^{(t)}(n_k^{(t)} - 2)}}{n_k^{(t)} - 1} \frac{C(n_k^{(t)})}{\kappa} \left(1 + \frac{\kappa^2}{n_k^{(t)} - 2}\right)^{-0.5(n_k^{(t)} - 1)}
\end{aligned}
$$

□

Corollary 1

If the number of pulls of all arms tend to infinity, the probability of regret decreases exponentially in T.

$$
\begin{aligned}
P(R^{(T)} \leq \epsilon) &\geq 1 - \sqrt{\frac{2\sigma_G^2 \sum_{k=1}^{K} H_{k,\epsilon}^{-2}}{\pi\,(T - 3K)}} KT \exp\left(-\frac{T - 3K}{2\sigma_G^2 \sum_{k=1}^{K} H_{k,\epsilon}^{-2}}\right) \\
&\geq \Omega\left(1 - \sqrt{T}\exp(-T)\right)
\end{aligned}
$$

Note that this bound is similar to the bandit setting with Gaussians with known variances presented in Hoffman et al. [2014]. Intuitively, this result makes sense, as for known variances, a Gaussian can be used to describe the posterior means, and indeed, as the number of pulls approaches infinity, our t-distributions converge to Gaussians.

Proof. Take the limit of the regret bound, established in Theorem 2, with the number of arm pulls going towards infinity:

$$
\begin{aligned}
& \lim_{n_k^{(t)} \to +\infty} P(R^T < \epsilon) \\
& \geq \lim_{n \to +\infty} 1 - 2 \sum_{k=1}^{K} \sum_{t=1}^{T} \frac{\sqrt{n(n-2)}}{n-1} \frac{C(n)}{\kappa} \left(1 + \frac{\kappa^2}{n-2}\right)^{-0.5(n-1)}
\end{aligned}
$$

The limit of the sum/product of factors is the sum/product of the limits of the factors, as long as the limits of the factors exist.

We derive now the limits of each factor in the regret bound (and thereby prove they exist):

1)

$$\begin{aligned}
&\lim_{n\to+\infty} \frac{\sqrt{n(n-2)}}{n-1} \\
&= \lim_{n\to+\infty} \sqrt{\frac{n^2-2n}{n^2-2n+1}} \\
&= 1
\end{aligned}$$

2)

$$\begin{aligned}
&\lim_{n\to+\infty} C(n) \\
&= \lim_{n\to+\infty} \frac{\Gamma(0.5n+0.5)}{\Gamma(0.5n)\sqrt{\pi n}} \\
&\overset{(S)}{\geq} \lim_{n\to+\infty} \frac{\sqrt{2\pi}\,(0.5n+0.5)^{0.5n+0.5-0.5}\exp\left(-0.5n-0.5\right)}{\sqrt{2\pi}\,(0.5n)^{0.5n-0.5}\exp\left(-0.5n+\frac{1}{12(0.5n)}\right)\sqrt{\pi n}} \\
&= \lim_{n\to+\infty} \frac{(0.5n+0.5)^{0.5n}}{(0.5n)^{0.5n-0.5}} \frac{\exp\left(-0.5n-0.5\right)}{\exp\left(-0.5n+\frac{1}{12(0.5n)}\right)} \frac{1}{\sqrt{\pi n}} \\
&= \lim_{n\to+\infty} \left(\frac{0.5(n+1)}{0.5n}\right)^{0.5n} \sqrt{0.5n}\exp\left(-0.5-\frac{1}{6n}\right)\frac{1}{\sqrt{\pi n}} \\
&= \lim_{n\to+\infty} \left(1+\frac{1}{n}\right)^{0.5n} \exp\left(-0.5-\frac{1}{6n}\right)\frac{1}{\sqrt{2\pi}} \\
&\overset{(E)}{=} \lim_{n\to+\infty} \exp\left(0.5\right)\exp\left(-0.5-\frac{1}{6n}\right)\frac{1}{\sqrt{2\pi}} \\
&= \frac{1}{\sqrt{2\pi}} \lim_{n\to+\infty} \exp\left(-\frac{1}{6n}\right) \\
&= \frac{1}{\sqrt{2\pi}},
\end{aligned}$$

where, at (S), we use Stirling's inequalities for the Gamma function [Andrews et al., 1999]:

$$\sqrt(2\pi)S(x) \leq \Gamma(x) \leq \sqrt(2\pi)S(x)\exp\left(\frac{1}{12x}\right), \tag{A.8}$$

with,

$$S(x) = x^{x-\frac{1}{2}}\exp(-x), \tag{A.9}$$

and, at (E), the exponential function limit:

$$\lim_{x\to+\infty}\left(1+\frac{k}{x}\right)^{mx}=\exp(mk). \tag{A.10}$$

3)

$$\begin{aligned}
&\lim_{n\to+\infty}\left(1+\frac{\kappa^2}{n-2}\right)^{-0.5(n-1)}\\
&=\lim_{m\to+\infty}\left(1+\frac{\kappa^2}{m}\right)^{-0.5(m+1)} \quad \text{, with } m=n-2\\
&=\left(\lim_{m\to+\infty}\left(1+\frac{\kappa^2}{m}\right)^{-0.5m}\right)\cancel{\left(\lim_{m\to+\infty}\left(1+\frac{\kappa^2}{m}\right)^{-0.5}\right)}\\
&=\exp\left(-0.5\kappa^2\right)
\end{aligned}$$

Combining all previous limits, we obtain the following exponential regret bound in the limit:

$$\begin{aligned}
&\lim_{n_k^{(t)}\to+\infty} P(R^{(T)}<\epsilon)\\
&\geq 1-2\sum_{k=1}^{K}\sum_{t=1}^{T}\frac{1}{\sqrt{2\pi}}\frac{1}{\kappa}\exp\left(-0.5\kappa^2\right)\\
&=1-\frac{1}{\sqrt{0.5\pi}\kappa}KT\exp\left(-0.5\kappa^2\right)
\end{aligned}$$

Finally, we set $\kappa=\sqrt{\frac{T-3K}{4\sigma_G^2\sum_{k=1}^{K}H_{k,\epsilon}^{-2}}}$, as described in Theorem 2.

□

2 Epidemic bandit experiments: computational details

2.1 FluTE source

FluTE is a stochastic individual-based model, that is implemented in C++. The original source code, as release by FluTE's author (i.e., D. Chao), is available from https://github.com/dlchao/FluTE. This GitHub repository contains FluTE's C++ source code, GNU/Linux-specific make files and a set of population density descriptions that can be used to simulate particular geographical settings (i.e., 2000-individual population, Seattle, Los Angelos and the entire United States).

Some changes were made to the source code to make our research easier: we organized the source code in a directory structure and added a CMake meta-make file. This CMake build file allows us to build the source code on GNU/Linux and MacOS [1]. These changes are publicly available on the https://github.com/vub-ai-lab/FluTE-bandits GitHub repository.

2.2 FluTE configurations

To run our experiments, we defined a model environment to evaluate pre-vaccination with little vaccine available, as described in detail in section 5. The pre-vaccination configuration script can be found in the 'configs/bandits' directory of the https://github.com/vub-ai-lab/FluTE-bandits GitHub repository. Note that this configuration script is a python Mako template (http://makotemplates.org/), to enable easy parameterization of the configuration script.

2.3 Bandit implementation

We implemented a flexible bandit framework in Scala, the code is publicly available on GitHub: https://github.com/vub-ai-lab/scala-bandits. This framework is specifically designed to enable us to easily experiment with new algorithms and environments (i.e., both Scala environments and external environments, such as e.g., the FluTE simulator environment).

2.4 High performance computing

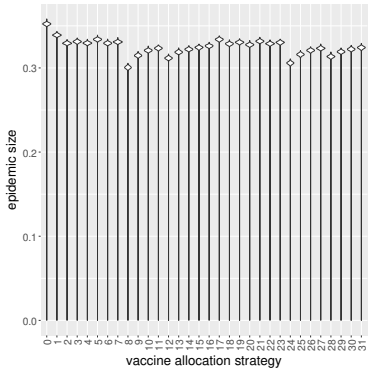
Simulating epidemics using individual-based models is a computationally intensive process. Therefore, our experiments were run on a powerful high performance computing cluster: the Flemish Supercomputer Center. We report that, to make this possible, all software

[1] Microsoft Windows should also work with little changes, but this was not tested yet.

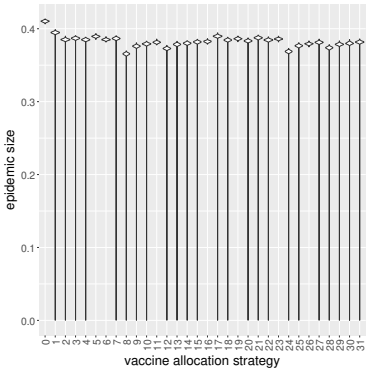
had to be installed (or build) for the high performance computing cluster. We report that our FluTE CMake file allows the generation of efficient code (i.e., using SSE instructions) for all plaforms used in our analyses (i.e., MacOS, XUbuntu desktop GNU/Linux and GNU/Linux on the high performance computing cluster).

On this HPC, we used ‘Ivy Bridge‘ nodes, more specifically nodes with two 10-core "Ivy Bridge" Xeon E5-2680v2 CPUs (2.8 GHz, 25 MB level 3 cache) and 64 GB of RAM. This infrastructure allowed us to run 20 FluTE simulations per node.

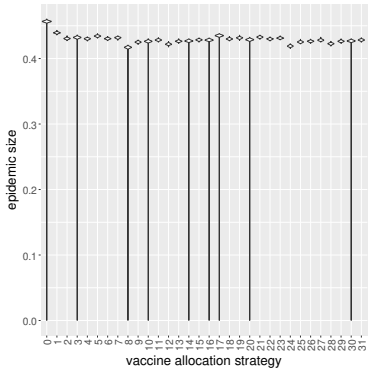
3 Outcome (i.e., epidemic size) distributions



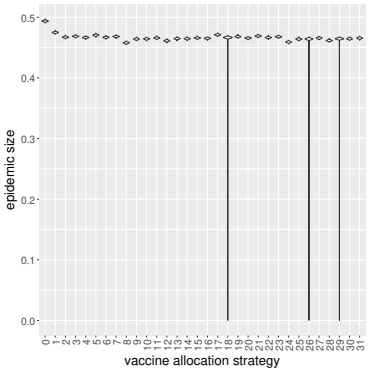
(a) Outcome distributions for $R_0 = 1.6$.



(b) Outcome distributions for $R_0 = 1.8$.

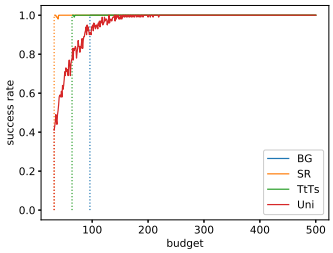


(c) Outcome distributions for $R_0 = 2.0$.

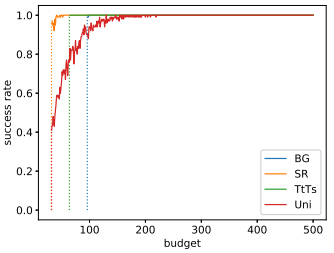


(d) Outcome distributions for $R_0 = 2.2$.

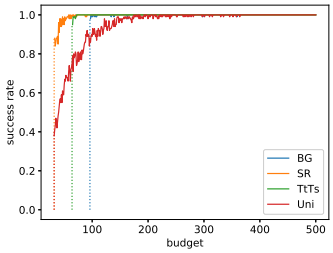
4 Bandit run success rates



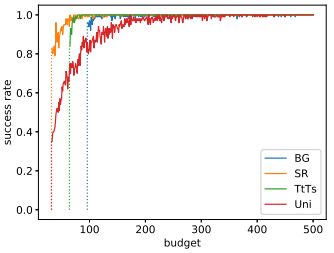
(a) Bandit run results for $R_0 = 1.6$.



(b) Bandit run results for $R_0 = 1.8$.

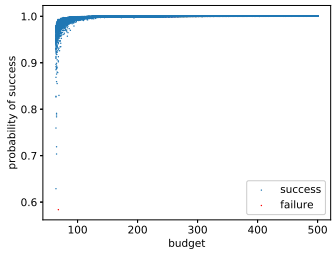


(c) Bandit run results for $R_0 = 2.0$.

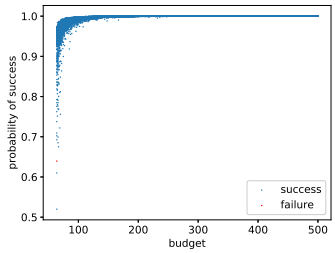


(d) Bandit run results for $R_0 = 2.2$.

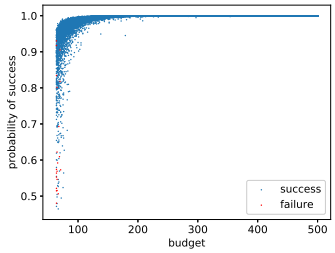
5 P_s values for Top-two Thompson sampling



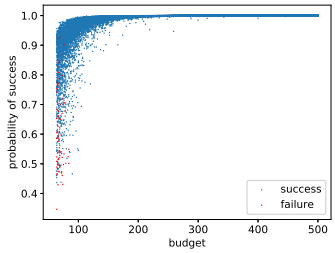
(a) P_s values for $R_0 = 1.6$.



(b) P_s values for $R_0 = 1.8$.

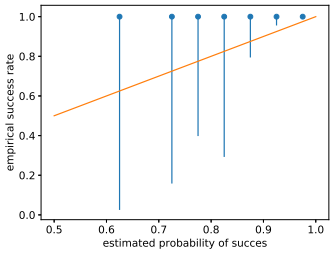


(c) P_s values for $R_0 = 2.0$.

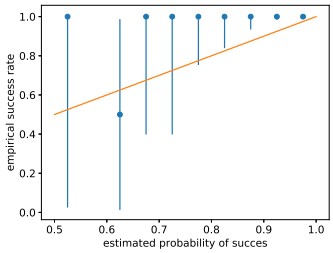


(d) P_s values for $R_0 = 2.2$.

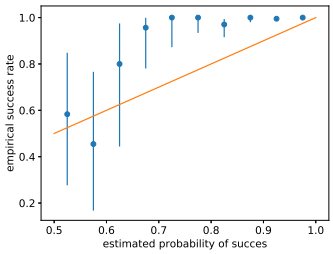
6 Binned distribution of P_s values for Top-two Thompson sampling



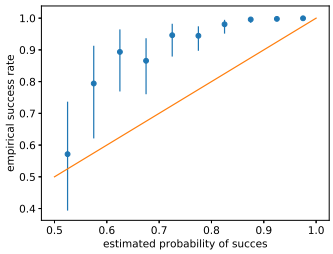
(a) Binned distribution for $R_0 = 1.6$.



(b) Binned distribution for $R_0 = 1.8$.



(c) Binned distribution for $R_0 = 2.0$.



(d) Binned distribution for $R_0 = 2.2$.

7 Expectation of the truncated t-distribution posterior

We consider a Gaussian reward distribution with unknown variance and assume an uninformative Jeffreys prior $(\sigma)^{-3}$ on (μ, σ^2).

Given rewards $\mathbf{r} = \{r_1, ..., r_n\}$, this prior leads to the non-standardized t-distributed posterior, that we truncate given that we know that the arm's means are in $[0, 1]$:

$$\mu \sim \mathcal{T}_{n,[0,1]}\left(\mu_0 = \frac{\sum_{i=1}^n r_i}{n}, \sigma_0^2 = \frac{\sum_{i=1}^n (r_i - \mu_0)^2}{n^2}\right). \tag{A.11}$$

Given the probability density function (pdf) $f(\cdot)$ of a non-standardized t-distribution $\mathcal{T}_\upsilon(\mu, \sigma^2)$

$$f(x; \mu, \sigma^2) = \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1 + \frac{(x-\mu)^2}{\upsilon\sigma^2}\right)^{-\frac{\upsilon+1}{2}}, \tag{A.12}$$

and cumulative density function (cdf) $F(\cdot; \mu, \sigma^2)$, we can compute the mean of the truncated non-standardized t-distribution using this normalized definite integral:

$$\begin{aligned} &\frac{\int_0^1 x f(x; \mu, \sigma^2) dx}{F(1, \mu, \sigma) - F(0, \mu, \sigma)} \\ &= \frac{\int_0^1 x f(x; \mu, \sigma^2) dx}{F(\frac{1-\mu}{\sigma}; 0, 1) - F(\frac{0-\mu}{\sigma}; 0, 1)} \end{aligned} \tag{A.13}$$

From this, we can derive an analytic expression by first considering the nominator:

$$\begin{aligned} &\int_0^1 x f(x; \mu, \sigma^2) dx \\ &\overset{(1)}{=} \int_0^1 x \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sigma\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1 + \frac{(x-\mu)^2}{\upsilon\sigma^2}\right)^{-\frac{\upsilon+1}{2}} dx \\ &= \int_{x=0}^{x=1} \sigma\frac{x-\mu+\mu}{\sigma} \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1 + \frac{(x-\mu)^2}{\upsilon\sigma^2}\right)^{-\frac{\upsilon+1}{2}} \frac{1}{\sigma} dx \\ &\overset{(2)}{=} \int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} (\sigma u + \mu) \frac{\Gamma\left(\frac{\upsilon+1}{2}\right)}{\sqrt{\upsilon\pi}\Gamma\left(\frac{\upsilon}{2}\right)}\left(1 + \frac{u^2}{\upsilon}\right)^{-\frac{\upsilon+1}{2}} du \\ &\overset{(3)}{=} \int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} \sigma u f(u; 0, 1) du + \int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} \mu f(u; 0, 1) du \end{aligned} \tag{A.14}$$

First (1), we fill in the pdf of the non-standardized. Next (2), we use integration by substitution, with,

$$u = \frac{x-\mu}{\sigma}, du = \frac{1}{\sigma}dx. \tag{A.15}$$

Finally (3), we have an expression that is the pdf of a standardized t-distribution.

Substituting the result of Equation A.14 in Equation A.13, we have:

$$\begin{aligned}
&\frac{\int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} \sigma u f(u;0,1)du + \int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} \mu f(u;0,1)du}{F(\frac{1-\mu}{\sigma};0,1) - F(\frac{0-\mu}{\sigma};0,1)} \\
&= \sigma \frac{\int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} u f(u;0,1)du}{F(\frac{1-\mu}{\sigma};0,1) - F(\frac{0-\mu}{\sigma};0,1)} + \mu \frac{\int_{u=\frac{0-\mu}{\sigma}}^{u=\frac{1-\mu}{\sigma}} f(u)du}{F(\frac{1-\mu}{\sigma};0,1) - F(\frac{0-\mu}{\sigma};0,1)} \\
&= \sigma \mathbb{E}[u] + \mu,
\end{aligned} \tag{A.16}$$

where, u follows a standardized t-distribution that is truncated on the interval $[\frac{-\mu}{\sigma}, \frac{1-\mu}{\sigma}]$.

8 Empirical validation of BFTS' heuristics

In Chapter 5, we conduct an experiment to empirically validate the heuristics that we derived in the Bayesian analysis of the Boundary Focused Thompson sampling algorithm in Section 6. We evaluate the heuristics for the same environments as in the experiments section, but with a limited number of arms ($K = 100$) and time steps (i.e., $3 \cdot 10^4$). In this section, we show the results for these additional experiments.

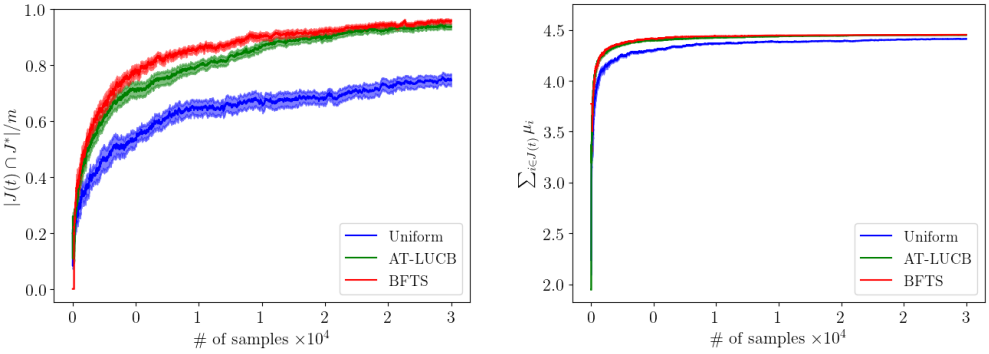


Figure A.5: Results for the linear Gaussian benchmark with fixed variance ($K = 100, m = 5$).

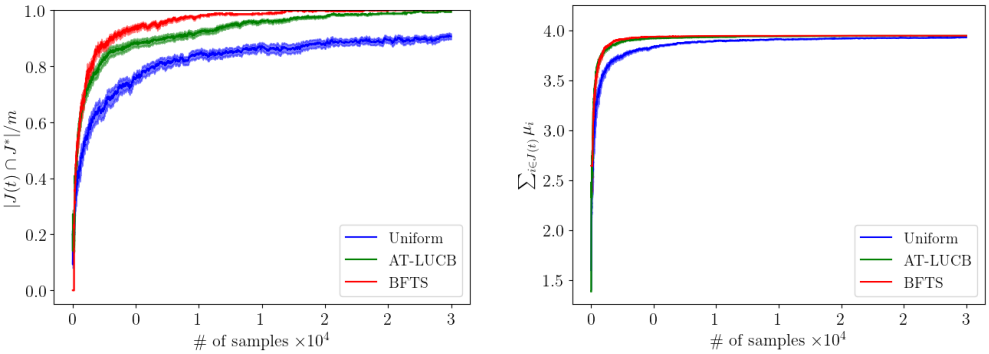


Figure A.6: Results for the polynomial Gaussian benchmark with fixed variance ($K = 100, m = 5$).

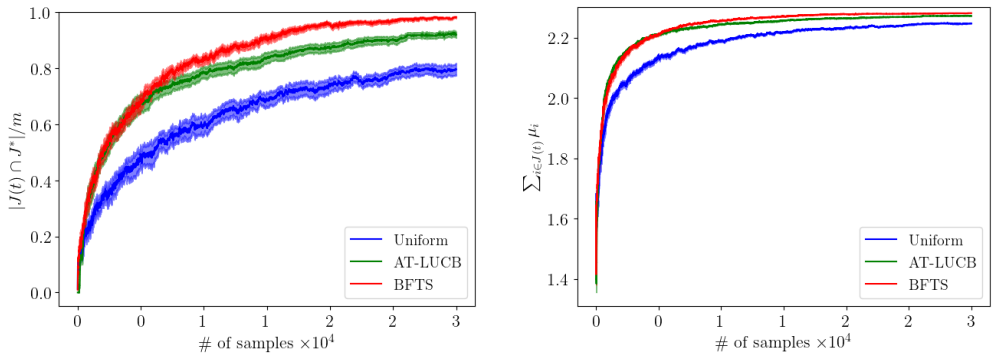


Figure A.7: Results for the caption benchmark ($K = 100, m = 5$).

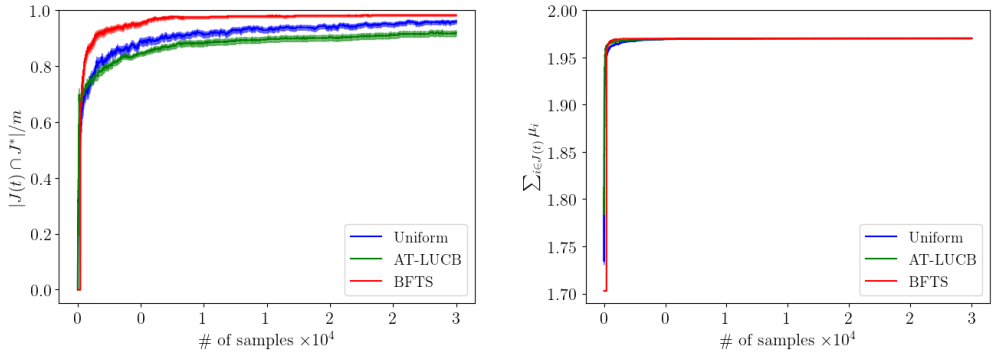


Figure A.8: Results for the scaled Gaussian benchmark ($K = 100, m = 5$).

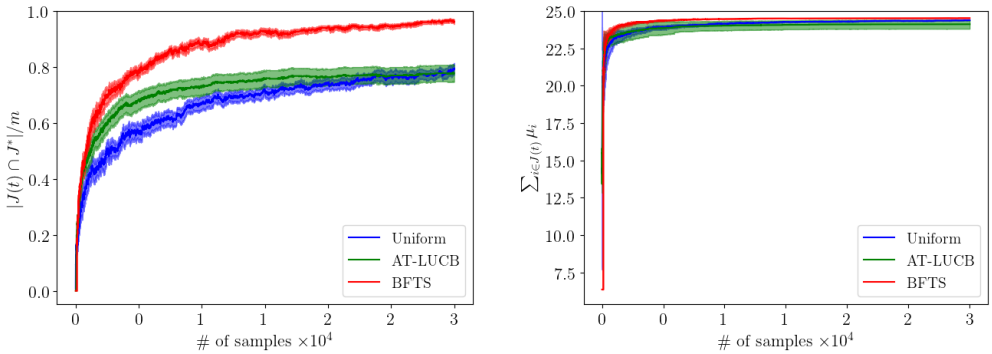


Figure A.9: Results for the Poisson benchmark ($K = 100, m = 5$).

9 Influenza model validation

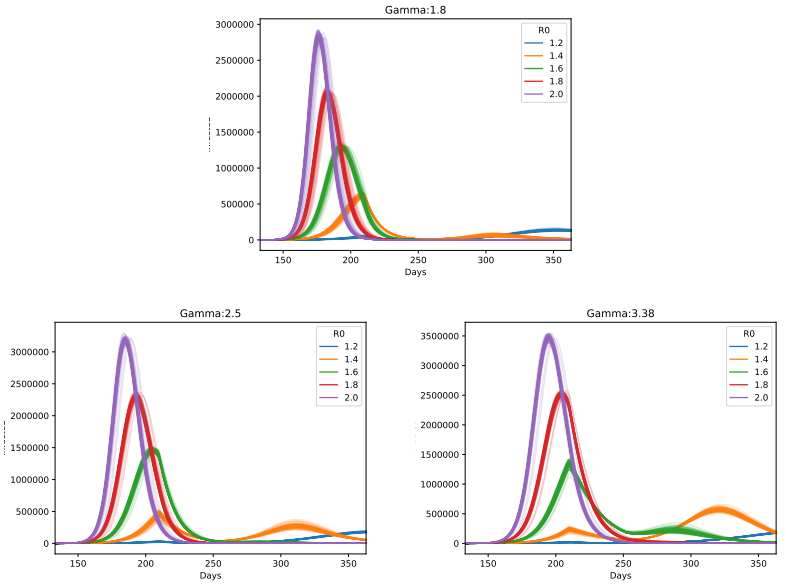


Figure A.10: We demonstrate our model for $R_0 \in \{1.2, 1.4, 1.6, 1.8, 2.0\}$ (enumerated in the legend) and an infectious period of 1.8 days (top panel), 2.5 days (bottom left panel) and 3.38 days (bottom right panel). For each parameter combination, we show 100 stochastic trajectories (light coloured lines) and the mean of these trajectories (dark coloured line).

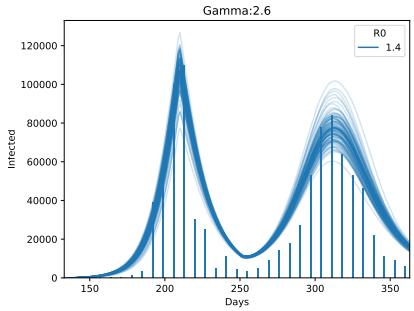


Figure A.11: We show that our model, using a reproductive number of 1.4 and a an average duration of infectiousness of 2.4 days is able to match the trends observed in the British pandemic of 2009. We show 100 stochastic trajectories in this Figure.

10 Comparing PPO and DQN

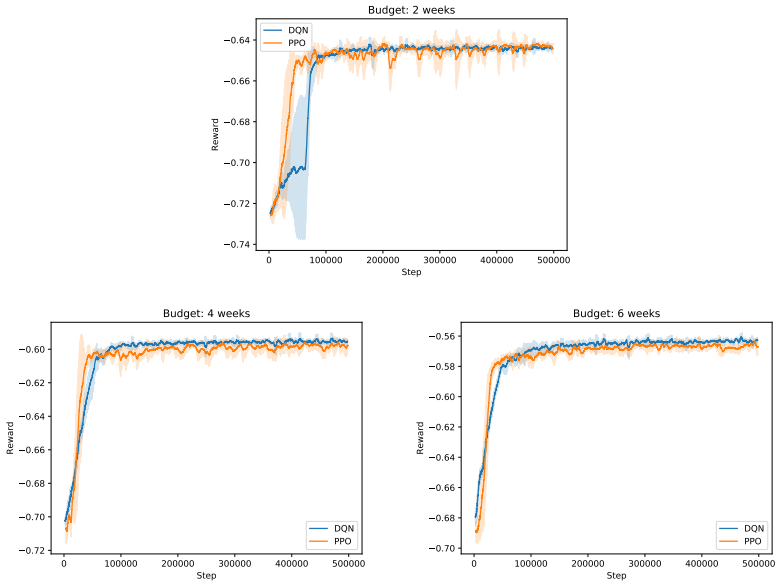


Figure A.12: Learning results for the Greenwich environment with $R_0 = 1.8$, for school closure budgets $b = \{2, 4, 6\}$. Reward curves for PPO (orange) and DQN (blue), using a rolling window of 100 steps. The shaded are shows the standard deviation of the reward signal.

11 Histograms for the top policies (attack rate)

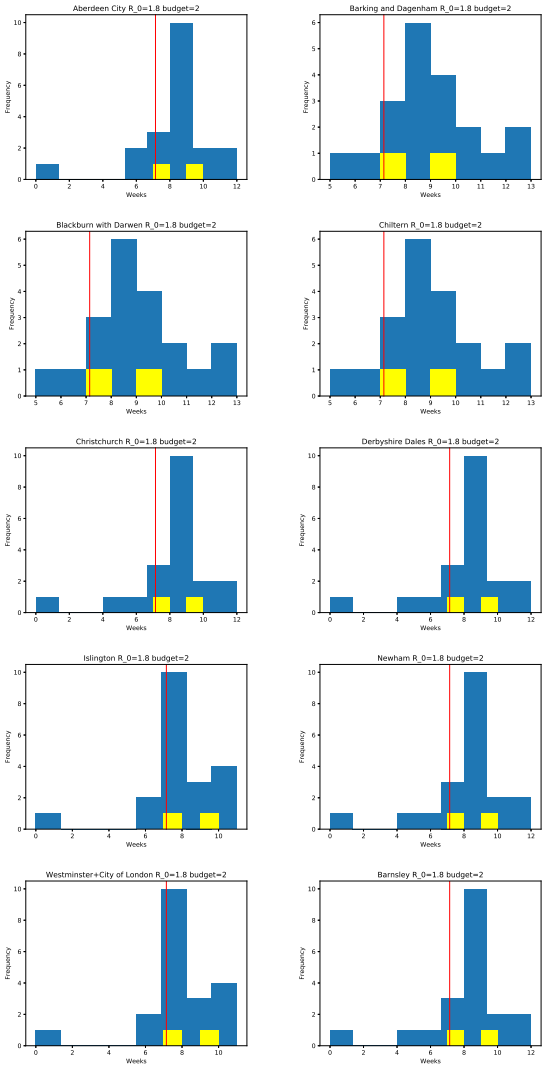


Figure A.13: Histogram of the top policies for $R_0 = 1.8$ and budget=2.

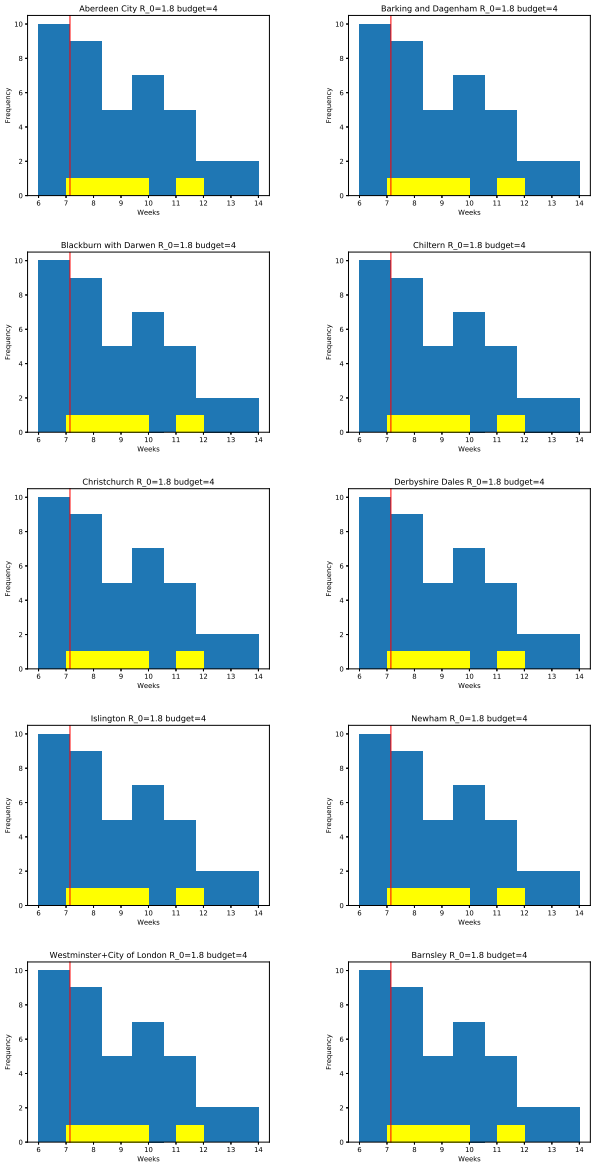


Figure A.14: Histogram of the top policies for $R_0 = 1.8$ and budget=4.

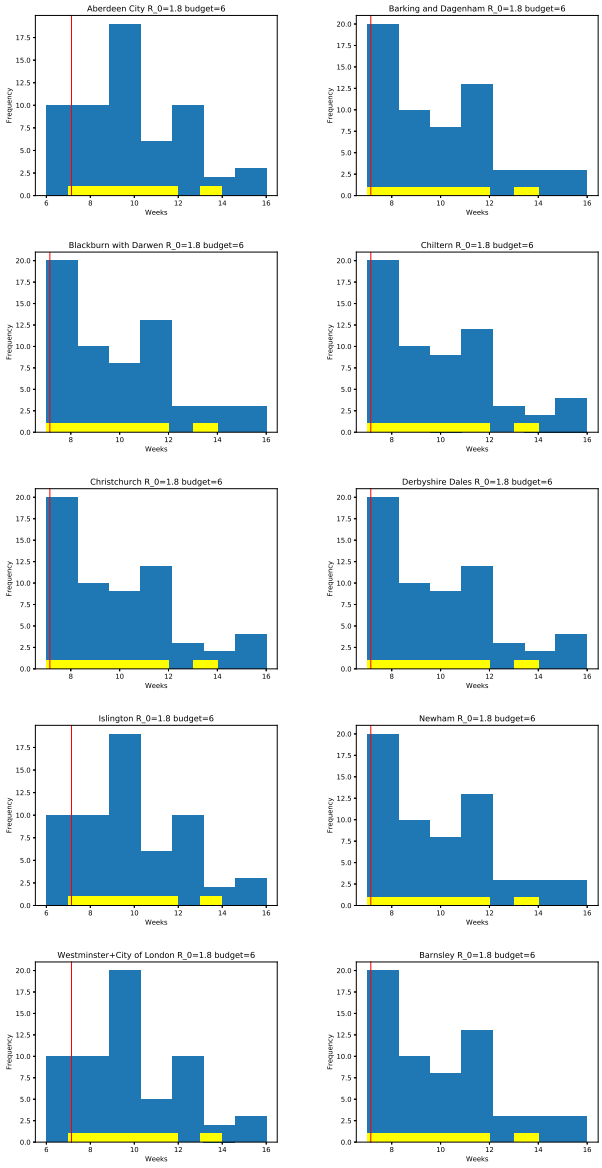


Figure A.15: Histogram of the top policies for $R_0 = 1.8$ and budget=6.

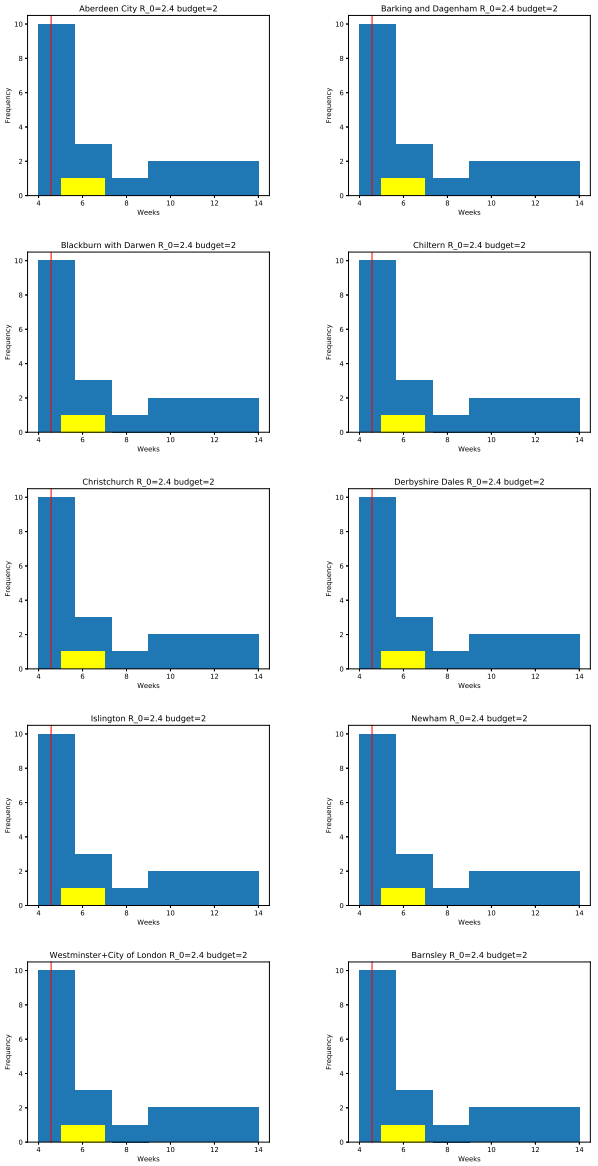


Figure A.16: Histogram of the top policies for $R_0 = 2.4$ and budget=2.

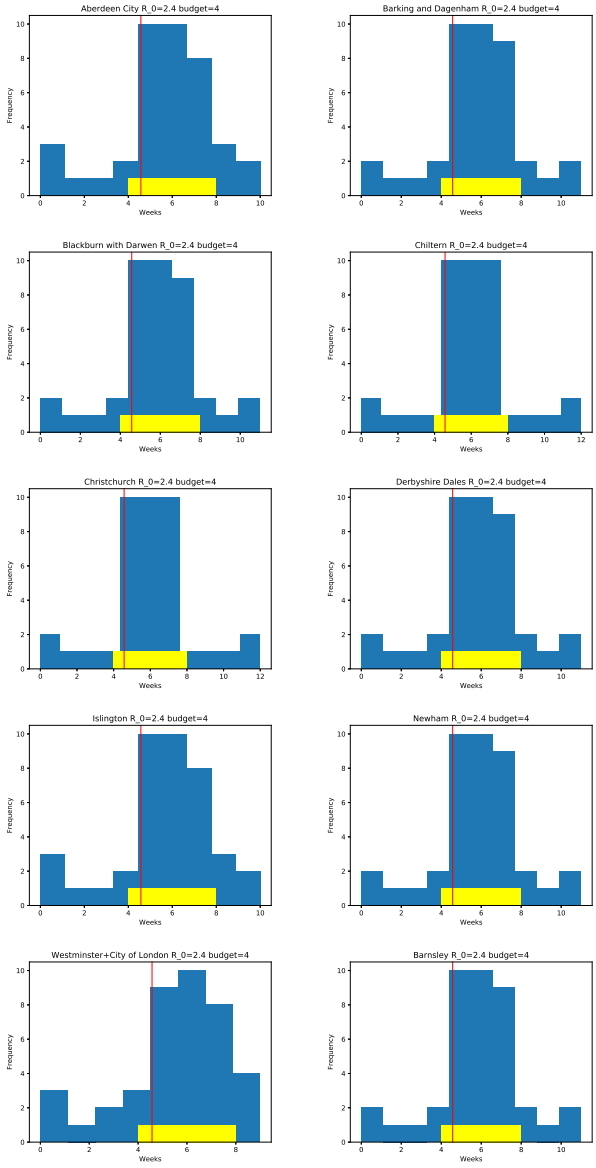


Figure A.17: Histogram of the top policies for $R_0 = 2.4$ and budget=4.

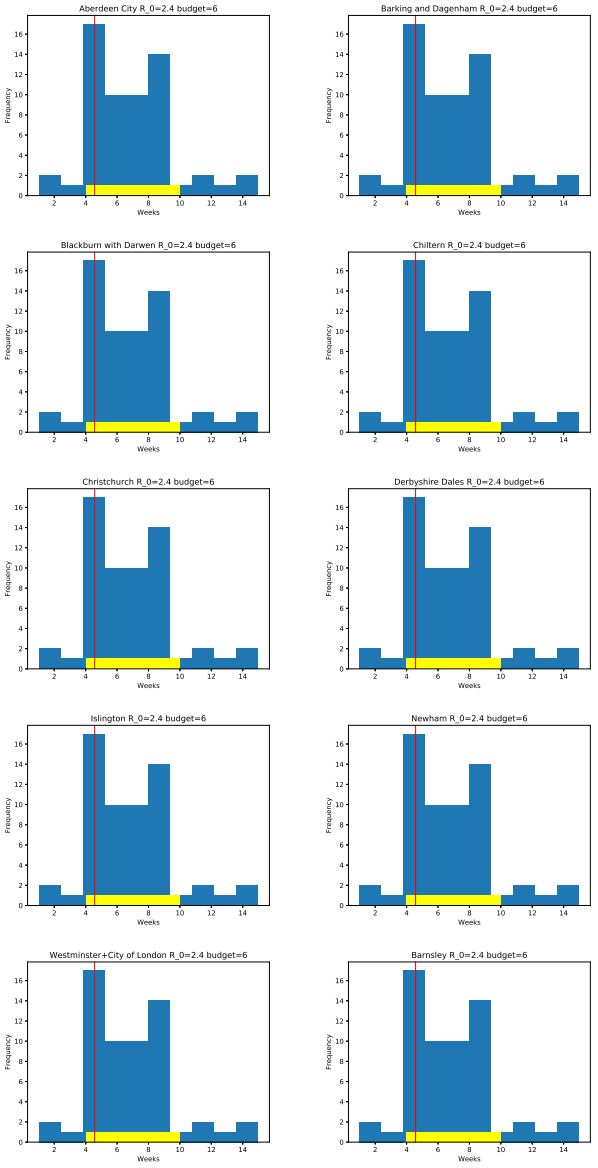


Figure A.18: Histogram of the top policies for $R_0 = 2.4$ and budget=6.

12 PPO learning curves ($R_0 = 1.8$)

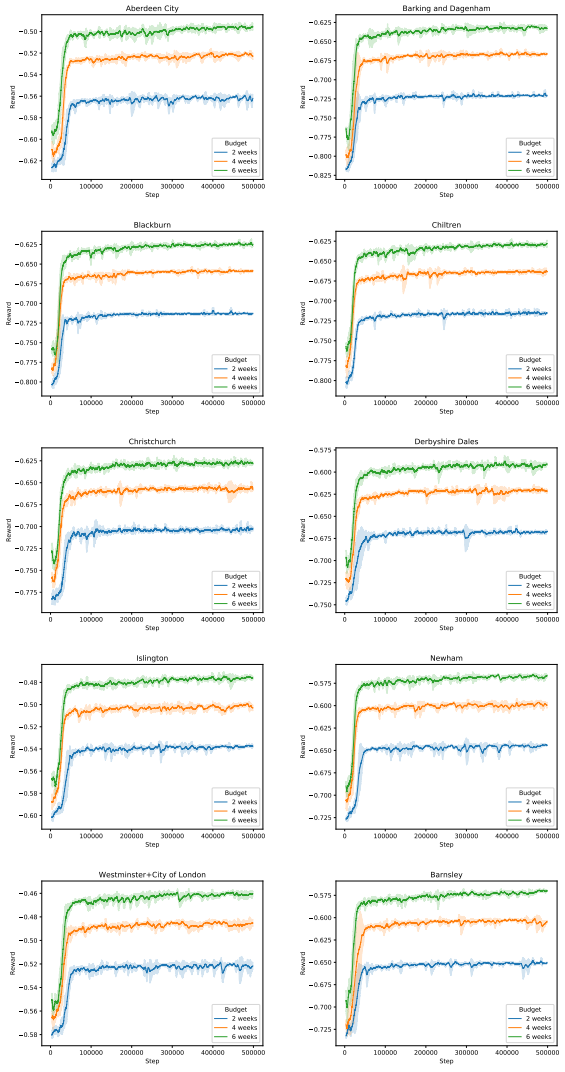


Figure A.19: PPO learning curves for $R_0 = 1.8$.

13 PPO learning curves ($R_0 = 2.4$)

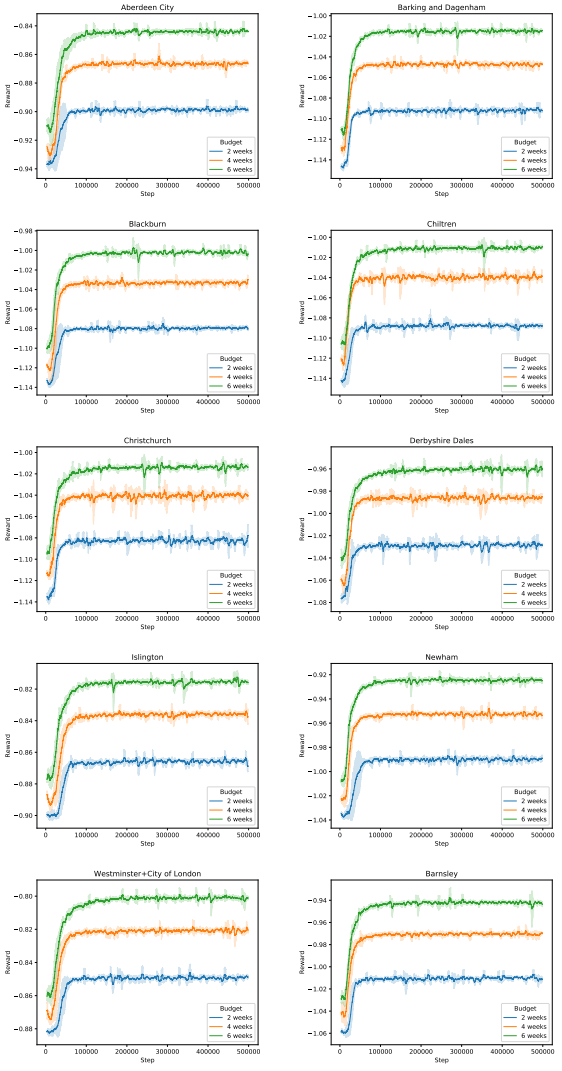


Figure A.20: PPO learning curves for $R_0 = 2.4$.

14 Comparing PPO to the ground truth (attack rate)

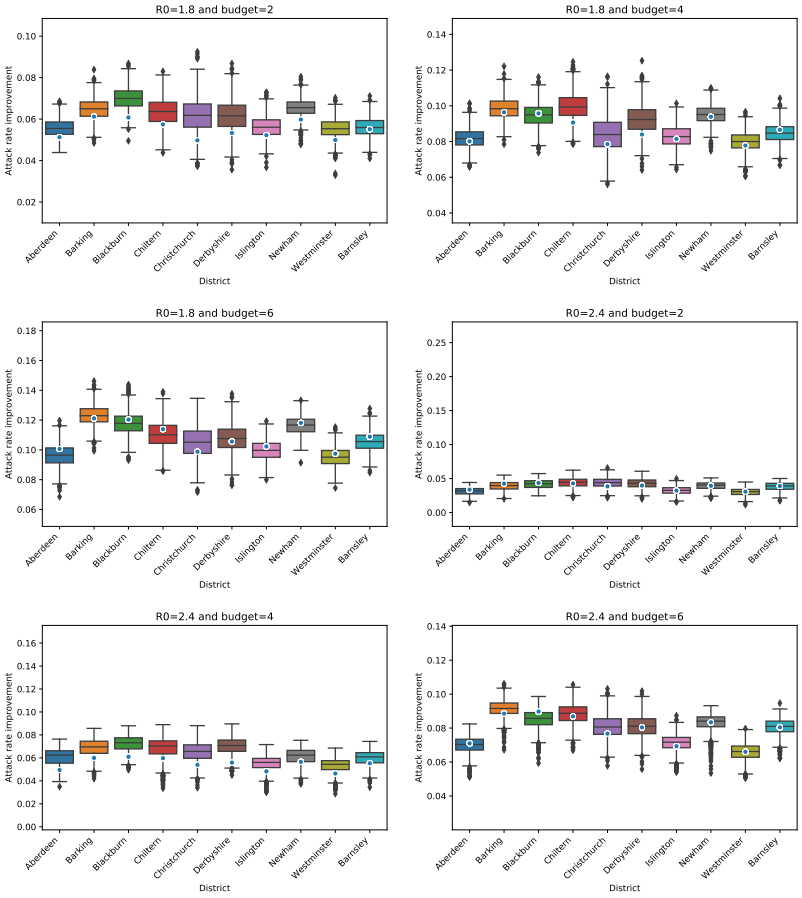


Figure A.21: Comparing PPO to the ground truth for $R_0 \in \{1.8, 2.4\}$ and $\acute{b} \in \{2, 4, 6\}$.

Curriculum vitae

Personal information

Pieter Jules Karel LIBIN, male, born in Leuven, Belgium (02/12/1982),
married to Karen Goedeweeck (born 31/07/1985),
with one daughter Arwen Libin (born 09/05/2016)

Education

Professional bachelor in Applied Informatics

Rega Department, Katholieke Hogeschool Leuven
Graduated cum laude in 2003

Master in Applied Informatics

Vrije Universiteit Brussel
Graduated summa cum laude in 2014
Master thesis: "Applying graphical modeling techniques to virological data"

Professional history

Since 2016 until 2019:

FWO pre-doctoral fellow

Vrije Universiteit Brussel, Department of computer science

Since 2015 until 2016:

PhD student
Vrije Universiteit Brussel, Department of computer science

Since 2012 until 2014:

Scientific programmer
Rega Institute for Medical Research, Katholieke Universiteit Leuven

2009 until 2012:

Software developer
Emweb

2006 until 2009:

Scientific programmer
Rega Institute for Medical Research, Katholieke Universiteit Leuven

2003 until 2006:

Software developer
LMS International

Academic awards

- Doctoral research grant for strategic basic research, by the National science foundation (FWO), 01/01/2015
- Grant for participation in a conference abroad (AAMAS conference, May, 2019), by the National science foundation (FWO), 02/04/2019
- Visionary paper award, for the paper titled "Efficient evaluation of influenza mitigation strategies using preventive bandits", by the Adaptive Learning Agents workshop, 29/03/2017
- Best paper award, for the paper titled "Interactive Multi-Objective Reinforcement Learning in Multi-Armed Bandits for Any Utility Function", by the Adaptive Learning Agents workshop, 15/07/2018
- Best student paper award, for the paper titled "IPC-Net: 3D Point-Cloud Segmentation Using Deep Inter-Point Convolutional Layers", International Conference on Tools with Artificial Intelligence, 07/11/2018

Teaching experience

- Teaching assistant for the course "Automata and computability" (Prof. Dr. Ann Nowé), at the Vrije Universiteit Brussel, during three academic years (2015-2018).
- Teaching assistant for the course "Fundamentals of computer science" (Prof. Dr. Ann Nowé), at the Vrije Universiteit Brussel, during two academic years (2015-2017).

Master students

- Laurens Hernalsteen, Vrije Universiteit Brussel, 2015-2016, with a project titled "Modelling HIV disease progression" (promotor: Ann Nowé)
- Nassim Versbraegen, Vrije Universiteit Brussel, 2016-2017, with a project titled "Modelling the undiagnosed HIV epidemic" (promotor: Ann Nowé)
- Wenjia Wang, Vrije Universiteit Brussel, 2017-2018, with a project titled "Anytime m-top Exploration for Preventive Bandits" (promotor: Ann Nowé)
- Thomas Cloostermans, Vrije Universiteit Brussel, 2018-2020, with a project titled "Bayesian Optimization for Conflict Resolution in Air Traffic Control" (joint advisorship with Timothy Verstraeten, promotor: Ann Nowé)

Journal publications (peer-reviewed)

Note: * denotes equal contribution.

1. Pineda-Pena, A., Pingarilho, M., Li, G., Vrancken, B., **Libin, P.**, Gomes, P., Camacho, R., Theys, K., Abecasis, A. and the Portuguese HIV-1 Resistance Study Group, "Drivers of HIV-1 transmission: The Portuguese case", PLOS ONE, vol 14, issue 9, 2019. [Peer reviewed, 2-yearly JCR impact factor 2018: 2.776]

2. Abbas Jariani, Christopher Warth, Koen Deforche, **Libin P.**, Alexei J Drummond, Andrew Rambaut, Frederick A Matsen Iv, Kristof Theys, "SANTA-SIM: simulating viral sequence evolution dynamics under selection and recombination", Virus Evolution, vol. 5, issue 1, 2019. [Peer reviewed, 2-yearly JCR impact factor 2018: 5.408]

3. Vagner Fonseca*, **Pieter J K Libin***, Kristof Theys*, Nuno R Faria, Marcio R T Nunes, Maria I Restovic, Murilo Freire, Marta Giovanetti, Lize Cuypers, Ann Nowé, Ana Abecasis, Koen Deforche, Gilberto A Santiago, Isadora C de Siqueira, Emmanuel J San, Kaliane C B Machado, Vasco Azevedo, Ana Maria Bispode Filippis, Rivaldo Venâncio da Cunha, Oliver G Pybus, Anne-Mieke Vandamme, Luiz C J Alcantara,

Tulio de Oliveira, "A computational method for the identification of Dengue, Zika and Chikungunya virus species and genotypes", PLoS Neglected Tropical Diseases, vol. 13, issue 5, 2019. [Peer reviewed, 2-yearly JCR impact factor 2018: 4.487] (* denotes equal contribution)

4. Anna Schultze, Carlo Torti, Alessandro Cozzi-Lepri, Anne-Mieke Vandamme, Maurizio Zazzi, Helen Sambatakou, Andrea De Luca, Anna Maria Geretti, Anders Sonnerborg, Lidia Ruiz, Laura Monno, Simona Di Giambenedetto, Andrea Gori, Giuseppe Lapadula, **European Transmitted Drug Resistance collaboration (EU-TDR)**[2], "The effect of primary drug resistance on CD4 cell decline and the viral load set-point in HIV positive individuals before the start of antiretroviral therapy", AIDS, vol. 33, issue 2, p 315 - 326, 2019. [Peer reviewed, 2-yearly JCR impact factor 2018: 4.499]

5. Cuypers L*, **Libin P***, Simmonds P, Nowé A, Muñoz-Jordán J, Alcantara L, Vandamme AM, Santiago G, Theys K., "Time to Harmonize Dengue Nomenclature and Classification.", Viruses, vol. 10, issue 10, 2018. [Peer reviewed, 2-yearly JCR impact factor 2018: 3.811] (* denotes equal contribution)

6. **Libin, P.**, Deforche, K., Theys, K. and Abecasis, A.,"VIRULIGN: fast codon-correct alignment and annotation of viral genomes.", Bioinformatics, vol 35, issue 10, p1763-1765, 2018. [Peer reviewed, 2-yearly JCR impact factor 2018: 4.53]

7. Kristof Theys*, **Libin P.***, Andrea-Clemencia Pineda-Pena,Ann Nowé,Anne-mieke Vandamme,Ana B Abecasis, "The impact of HIV-1 within-host evolution on transmission dynamics", Current Opinion in Virology, vol. 28, p 92 - 101, 2018. [Peer reviewed, 2-yearly JCR impact factor 2018: 5.4] (* denotes equal contribution)

8. Lize Cuypers, **Libin. P.**,Yoeri Schrooten, Kristof Theys, Velia Chiara Di Maio,Valeria Cento,Maja M Lunar, Frederik Nevens, Mario Poljak, Francesca Ceccherini-Silberstein, Ann Nowé, Kristel Van Laethem, Anne-Mieke Vandamme, "Exploring resistance pathways for first-generation NS3/4A protease inhibitors boceprevir and telaprevir using Bayesian network learning", Infection Genetics and Evolution, vol. 53, p. 15 - 23, 2017. [Peer reviewed, 2-yearly JCR impact factor 2018: 2.611]

9. Simona Paraschiv, Leontina Banica, Ionelia Nicolae, Iulia Niculescu, Adrian Abagiu, Raluca Jipa, Andrea-Clemencia Pineda-Pena, Marta Pingarilho, Emil Neaga, Kristof Theys, **Pieter Libin**, Dan Otelea, Ana Abecasis, "Epidemic dispersion of HIV and HCV in a population of co-infected Romanian injecting drug users", PLoS ONE, vol. 12, issue 10, 2017. [Peer reviewed, 2-yearly JCR impact factor 2018: 2.776]

[2]Author as a member of the European Transmitted Drug Resistance collaboration work group.

10. **Pieter Libin**, Ewout Vanden Eynden, Francesca Incardona, Ann Nowé, Antonia Bezenchek, Anders Sönnerborg, Anne-Mieke Vandamme, EucoHIV Study Group, Kristof Theys, Guy Baele, "PhyloGeoTool: interactively exploring large phylogenies in an epidemiological context", Bioinformatics, vol. 33, issue 24, p. 3993 - 3995, 2017. [Peer reviewed, 2-yearly JCR impact factor 2018: 4.531]

11. Kristof Theys, **Pieter Libin**, Kai Dallmeier, Andrea Clemencia Pineda-Pena, Anne-Mieke Vandamme, Lize Cuypers, Ana B Abecasis, "Zika genomics urgently need standardized and curated reference sequences", PLoS Pathogens, vol. 13, issue 9, 2017. [Peer reviewed, 2-yearly JCR impact factor 2018: 6.463]

12. Sinaye Ngcapu, Kristof Theys, **Pieter Libin**, Vincent C Marconi, Henry Sunpath,Thumbi Ndung'u, Michelle L Gordon, "Characterization of Nucleoside Reverse Transcriptase Inhibitor-Associated Mutations in the RNase H Region of HIV-1 Subtype C Infected Individuals", Viruses, vol. 9, issue 11, 2017. [Peer reviewed, 2-yearly JCR impact factor 2018: 3.811]

13. Lize Cuypers, Guandgi Li, Christoph Neumann-Haefelin, Supinya Piampongsant, **Pieter Libin**, Kristel Van Laethem, Anne-Mieke Vandamme, Kristof Theys, "Mapping the genomic diversity of HCV subtypes 1a and 1b: Implications of structural and immunological constraints for vaccine and drug development", Virus Evolution, vol. 2, issue 2, 2016. [Peer reviewed, 2-yearly JCR impact factor 2018: 5.408]

14. Lize Cuypers, Joke Snoeck, Lien Kerremans, **Pieter Libin**, Raf Crabbé, Sonia Van Dooren, Grégoire Vuagniaux, Anne-Mieke Vandamme, "HCV1b genome evolution under selective pressure of the cyclophilin inhibitor alisporivir during the DEB-025-HCV-203 phase II clinical trial", Infection Genetics and Evolution, vol. 44, p. 169 - 181, 2016. [Peer reviewed, 2-yearly JCR impact factor 2018: 2.611]

15. Lize Cuypers, Guangi Li, **Pieter Libin**, Supinya Piampongsant, Anne-Mieke Vandamme, Kristof Theys, "Genetic Diversity and Selective Pressure in Hepatitis C Virus Genotypes 1–6: Significance for Direct-Acting Antiviral Treatment and Drug Resistance", Viruses, vol. 7, issue 9, p. 5018 - 5039, 2015. [Peer reviewed, 2-yearly JCR impact factor 2018: 3.811]

16. Kristof Theys, Ana Abecasis, **Pieter Libin**, Perpétua Gomes, Joaquim Cabanas,Ricardo Camacho, Kristel Van Laethem, "Discordant predictions of residual activity could impact dolutegravir prescription upon raltegravir failure", Journal of Clinical Virology, vol. 70, p. 120 - 127, 2015 [Peer reviewed, 2-yearly JCR impact factor 2018: 3.02]

Conference proceedings

Note: * denotes equal contribution.

1. Verstraeten, T., **Libin, P.** and Nowé, A., "Fleet Control using Coregionalized Gaussian Process Policy Iteration", European Conference on Artificial Intelligence, Manuscript accepted, 2020.
2. **Libin, P.**, Verstraeten, T., Roijers, D. M., Wang, W., Theys, K., and Nowé, A., "Thompson sampling for m-top Exploration", International Conference on Tools with Artificial Intelligence, p. 1414-1420, 2019.
3. **Libin, P.***, Versbraegen, N.*, Abecasis, A. B., Gomes, P., Lenaerts, T., and Nowé, A., "Towards a phylogenetic measure to quantify HIV incidence.", Benelux Conference on Artificial Intelligence, 2019. (* denotes equal contribution)
4. Méndez-Hernández, B. M., Rodríguez-Bazan, E. D., Martinez-Jimenez, Y., **Libin, P.***, and Nowé, A.*, "A Multi-objective Reinforcement Learning Algorithm for JSSP", International Conference on Artificial Neural Networks, p. 567-584, 2019. (* denotes equal contribution)
5. **Libin, P.J.K.**, Verstraeten, T., Roijers, D.M., Grujic, J., Theys, K., Lemey, P. and Nowé, A., "Bayesian Best-Arm Identification for Selecting Influenza Mitigation Strategies". In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 456-471), 2018.
6. Felipe Gomez Marulanda, **Pieter Libin**, Timothy Verstraeten, Ann Nowe, "Deep hybrid approach for 3D plane segmentation", European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, vol. 27, p529-534, 2019.
7. Marulanda, F. G., **Libin, P.**, Verstraeten, T., and Nowé, A., "IPC-Net: 3D Point-Cloud Segmentation Using Deep Inter-Point Convolutional Layers", International Conference on Tools with Artificial Intelligence, p. 293-301, 2018. (* denotes equal contribution)
8. **Pieter Libin***, Laurens Hernalsteen*, Kristof Theys, Perpetua Gomes, Ana Abecasis, Ann Nowe, "Bayesian inference of set-point viral load transmission models", Benelux Artificial Intelligence Conference, p.: 107 - 12, 2018.
9. **Pieter Libin**, Timothy Verstraeten, Kristof Theys, Diederik Roijers, Peter Vrancx, Ann Nowe, "Efficient evaluation of influenza mitigation strategies using preventive bandits", Autonomous Agents and Multiagent Systems, p. 67 - 85, 2017.

Conference presentations

1. T. Verstraeten, E. Bargiacchi, **P. Libin**, D. Roijers, A. Nowé, Thompson Sampling for Factored Multi-Agent Bandits, AAMAS 2020, 9-13/05/2020, Auckland, New-Zealand. Poster presentation. [Extended abstract]

2. **Pieter Libin**, Benjamin Linard, Nassim Versbraegen, Peter Simmonds, Donald Smith, Ana Abecasis, Ann Nowé, Fabio Pardi, Kristof Theys, PLATYPI: fast phylogenetic genotype classification and recombination detection, HIV Dynamics/Evolution, 24-27/03/2019, Lisbon, Portugal [Conference abstract]

3. **Pieter Libin**, Koen Deforche, Ana B. Abecasis, and Kristof Theys, VIRULIGN: fast codon-correct alignment and annotation of viral genomes, 23rd international bioinformatics workshop on virus evolution and molecular epidemiology, 26-31/08/2018 September, 2014; Berlin, Germany. Poster. [Conference abstract]

4. Diederik M Roijers, Luisa M Zintgraf, **Pieter Libin**, Ann Nowé, Interactive Multi-Objective Reinforcement Learning in Multi-Armed Bandits for Any Utility Function, Adaptive Learning Agents (ALA) workshop at AAMAS/ICML/IJCAI, 14-15/06/2018, Stockholm, Sweden. Oral presentation [Workshop paper, peer reviewed, Best paper award]

5. Versbraegen, N., **Libin, P.**, Abecasis, A., Gomes, P., Lenaerts, T., Nowe, A., Inferring HIV prevalence from genetic data. Benelux Bioinformatics Conference, 14-15/12/17, Leuven, Belgium. Poster presentation. [Conference abstract]

6. Ngcapu, S., Theys, K., **Libin, P.**, Marconi, V. C., Sunpath, H., Ndung'u, T., Gordon, M. L., Characterization of Nucleoside Reverse Transcriptase Inhibitor-Associated Mutations in the RNase H Region of HIV-1 Subtype C Infected Individuals, Benelux Bioinformatics Conference, 13-14 december 2017, Leuven, Belgium. Poster Presentation. [Conference abstract]

7. Lize Cuypers, **Pieter Libin**, Ana Abecasis, Anne-Mieke Vandamme, Kristof Theys, Zika virus genetic diversity and selective pressure: importance for diagnostics, vaccines and therapeutics, European Congress on Tropical Medicine and International Health, 16-20/10/2017, Antwerp, Belgium. Oral presentation. [Conference abstract]

8. **Libin P.J.K.**, Vanden Eynden E., Incardona F., Nowé A., Bezenchek A., EucoHIV study group, Sönnerborg A., Vandamme A., Theys K. and Baele G., Interactively exploring the global Dengue phylogeny with PhyloGeoTool, European Congress on Tropical Medicine and International Health, 16-20/10/2017, Antwerp, Belgium. Oral presentation. [Conference abstract]

9. Cuypers L., **Libin P.J.K.**, Simmonds P., Nowé A., Muñoz-Jordán J.L., Alcântara L.C.J., Vandamme A., Santiago G.A., Theys K., Time to Harmonize Dengue Nomenclature and Classification, European Congress on Tropical Medicine and International Health, 16-20/10/2017, Antwerp, Belgium. Oral presentation. [Conference abstract]

10. Pingarilho, M., Pineda-Pena, A., Gomes, P., Amaral-Alves, D., Pimentel, V., **Libin, P.**, Theys, K., Martins, M. , Dias, S., Vandamme, A., Transmitted drug-resistance in newly diagnosed HIV drug-naive individuals in Portugal. International workshop on HIV transmission- principles of intervention, 21-22/07/17, Paris, France. Poster presentation. [Conference abstract]

11. Pingarilho, M., Pineda-Pena, A., Gomes, P., **Libin, P.**, Theys, K., Martins, M., Dias, S., Vandamme, A., Camacho, R. J., Abecasis, A. B., Molecular epidemiology of HIV infection in portuguese migrant population, EUROPEAN MEETING ON HIV & HEPATITIS, 7-9/06/17, Rome, Italy. Poster presentation. [Conference abstract]

12. **Pieter Libin**, Timothy Verstraeten, Kristof Theys, Diederik Roijers, Peter Vrancx, Ann Nowe, Efficient evaluation of influenza mitigation strategies using preventive bandits, 2017 Adaptive Learning Agents (ALA) workshop @ AAMAS, 8-9/05/17, Sao Paolo, Brazil. Oral presentation [Workshop paper, peer reviewed]

13. Pingarilho, M., Pineda-Pena, A., Gomes, P., Pimentel, V., **Libin, P.**, Theys, K., Martins, M. D. R., Dias, S. F., Vandamme, A., Camacho, R., Abecasis, A., BEST HOPE - Cohort of HIV newly diagnosed patients in Portugal, 1 April 2017, Lisbon, Portugal. Poster presentation. [Conference abstract]

14. Cuypers, L., **Libin, P.**, Abecasis, A., Vandamme, A., Theys, K., Zika virus genetic diversity and selective pressure: important for diagnostics, vaccines and therapeutics, International Conference on Zika Virus, 22-25/02/17, Washington, United States. Poster presentation. [Conference abstract]

15. Alcantara, L., Faria, N., **Libin, P.**, Nunes, M., Fonseca, V., Restovic, M. I., Freire, M., Giovanetti, M., Theys, K., Cuypers, L., Analyzing the epidemic size distributions of an individual-based influenza model, First International Conference on Zika Virus, 22-25 February 2017, Washington, United States. Poster presentation. [Conference abstract]

16. Mendez-Hernandez, B. M., Martinez-Jimenez, Y., Rodriguez-Bazan, E., **Libin, P.**, Nowe, A., CIPS: a new method to handle robustness in the Job Shop Scheduling problem, ORBEL, 2-3 February 2017, Brussels, Belgium. Poster presentation. [Conference abstract]

17. **Libin, P.**, Versbraegen, N., Cuypers, L., Theys, K., and Nowé, A., An automated maximum likelihood method for classifying virus sequences. European Conference on Computational Biology, 3-7 September 2016, Den Haag, The Netherlands. Poster presentation. [Conference abstract]

18. Cuypers, L., Li, G., Neumann-Haefelin, C., **Libin, P.**, Van Laethem, K., Vandamme, A., Theys, K., The host immune system may not be the main driving factor of hepatitis C virus genetic diversity. Vilnius International Summit on Communicable Diseases, 26/07/16. Vilnius, Lithuania. Poster presentation. [Conference abstract]

19. Cuypers, L., **Libin, P.**, Nowé, A., Vandamme, A., Santiago, G., Theys, K., Classification and nomenclature of Dengue: challenges of the present and defining the future, 26/06/16, Vilnius, Lithuania. Poster presentation. [Conference abstract]

20. Switzer, W. M., Pan, Y., Saduvala, N., Zhang, T., Hernandez, A., **Libin, P.**, Struck, D., de Oliveira, T., Vandamme, A., Wertheim, J., Comparing Three HIV-1 Subtyping Tools With a Novel Phylogenetic-based Method, CROI 2016, 22-25/02/16, Boston, United States. Poster presentation. [Conference abstract]

21. Cuypers, L., Neumann-Haefelin, C., **Libin, P.**, Van Laethem, K., Vandamme, A., Theys, K., Mapping HCV1a and HCV1b full-length genomes in the context of structural and immunological constraints and drug resistance-related positions, HepDART, 6-10/12/15, Walilea, United States. Poster presentation. [Conference abstract]

22. Cuypers, L., Li, G., **Libin, P.**, Vandamme, A., Theys, K., Frequency of consensus residues in hepatitis C virus genotypes 1 to 6: significance for DAA treatment and drug resistance, Viral Infections of Regional Significance, 3-5/10/15, Moscow, Russian Federation. Poster presentation. [Conference abstract]

23. **P. Libin**, K. Deforche, K. Theys, A.-M. Vandamme. ASANOD: Automated Sequence Anomaly Detection tool. 25th European Congress of Clinical Microbiology and Infectious Diseases, 25-28 April, 2015; Copenhagen, Denmark. Oral presentation. [Conference abstract]

24. K. Theys, A. Abecasis, **P. Libin**, P. Gomes, J. Cabanas, R. Camacho, K. Van Laethem. Discordant predictions could impact dolutegravir use upon raltegravir failure. Conference on Retroviruses and Opportunistic Infections (CROI) 2015, 23-26 February, 2015; Seattle, Washington, United States. Poster. [Conference abstract]

Bibliography

Abbink, P., K. E. Stephenson, and D. H. Barouch
2018. Zika virus vaccines. *Nature reviews Microbiology*, 16(10):594.

Agrawal, S. and N. Goyal
2012. Analysis of Thompson sampling for the multi-armed bandit problem. In *Conference on Learning Theory*, Pp. 31–39.

Aguiar, M. and N. Stollenwerk
2017. Dengvaxia efficacy dependency on serostatus: a closer look at more recent data. *Clinical Infectious Diseases*.

Aitchison, J.
1983. Principal component analysis of compositional data. *Biometrika*, 70(1):57–65.

Aitchison, J.
1992. On criteria for measures of compositional difference. *Mathematical Geology*, 24(4):365–379.

Aitchison, J.
1994. Principles of compositional data analysis. *Lecture Notes-Monograph Series*, Pp. 73–81.

Aitchison, J. and V. Pawlowsky-Glahn
1997. The one-hour course in compositional data analysis or compositional data analysis is simple. In *Proceedings of IAMG*, volume 97, Pp. 3–35.

Akiner, M. M., B. Demirci, G. Babuadze, V. Robert, and F. Schaffner
2016. Spread of the invasive mosquitoes aedes aegypti and aedes albopictus in the black sea region increases risk of chikungunya, dengue, and zika outbreaks in europe. *PLoS neglected tropical diseases*, 10(4):e0004664.

Alberts, B., A. Johnson, J. Lewis, M. Raff, K. Roberts, and P. Walter
2002. *Molecular Biology of the Cell. 4th edition*. Garland Science.

Alderton, S., E. T. Macleod, N. E. Anderson, K. Schaten, J. Kuleszo, M. Simuunza, S. C. Welburn, and P. M. Atkinson
2016. A multi-host agent-based model for a zoonotic, vector-borne disease. a case study on trypanosomiasis in eastern province, zambia. *PLoS neglected tropical diseases*, 10(12):e0005252.

Allen, E. J., L. J. Allen, A. Arciniega, and P. E. Greenwood
2008. Construction of equivalent stochastic differential equation models. *Stochastic analysis and applications*, 26(2):274–297.

Andersen, K. G., A. Rambaut, W. I. Lipkin, E. C. Holmes, and R. F. Garry
2020. The proximal origin of sars-cov-2. *Nature Medicine*, Pp. 1–3.

Anderson, D. F.
2007. A modified next reaction method for simulating chemical systems with time dependent propensities and delays. *The Journal of chemical physics*, 127(21):214107.

Anderson, D. F.
2008. Incorporating postleap checks in tau-leaping. *The Journal of chemical physics*, 128(5):054103.

Anderson, R. M. and R. M. May
1992. *Infectious diseases of humans: dynamics and control*. Oxford university press.

Andrews, G. E., R. Askey, and R. Roy
1999. *Special functions*, volume 71. Cambridge university press.

Apolloni, A., C. Poletto, and V. Colizza
2013. Age-specific contacts and travel patterns in the spatial spread of 2009 h1n1 influenza pandemic. *BMC infectious diseases*, 13(1):176.

Aranda, C., M. J. Martínez, T. Montalvo, R. Eritja, J. Navero-Castillejos, E. Herreros, E. Marqués, R. Escosa, I. Corbella, E. Bigas, et al.
2018. Arbovirus surveillance: first dengue virus detection in local aedes albopictus mosquitoes in europe, catalonia, spain, 2015. *Eurosurveillance*, 23(47).

Arulkumaran, K., M. P. Deisenroth, M. Brundage, and A. A. Bharath
2017. Deep reinforcement learning: A brief survey. *IEEE Signal Processing Magazine*, 34(6):26–38.

Aubry, M., A. Teissier, M. Huart, S. Merceron, J. Vanhomwegen, C. Roche, A.-L. Vial, S. Teururai, S. Sicard, S. Paulous, et al.
2017. Zika virus seroprevalence, french polynesia, 2014–2015. *Emerging infectious diseases*, 23(4):669.

Audibert, J.-Y. and S. Bubeck
2010. Best arm identification in multi-armed bandits. In *COLT-23th Conference on Learning Theory*.

Auer, P., N. Cesa-Bianchi, and P. Fischer
2002. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47(2-3):235–256.

Azman, A. S. and J. Lessler
2015. Reactive vaccination in the presence of disease hotspots. *Proceedings of the Royal Society B: Biological Sciences*, 282(1798):20141341.

Baguelin, M., A. J. Van Hoek, M. Jit, S. Flasche, P. J. White, and W. J. Edmunds
2010. Vaccination against pandemic influenza a/h1n1v in england: a real-time economic evaluation. *Vaccine*, 28(12):2370–2384.

Balcan, D., H. Hu, B. Goncalves, P. Bajardi, C. Poletto, J. J. Ramasco, D. Paolotti, N. Perra, M. Tizzoni, W. Van den Broeck, et al.
2009. Seasonal transmission potential and activity peaks of the new influenza a (h1n1): a monte carlo likelihood analysis based on human mobility. *BMC medicine*, 7(1):45.

Ball, F., T. Britton, T. House, V. Isham, D. Mollison, L. Pellis, and G. S. Tomba
2015. Seven challenges for metapopulation models of epidemics, including households models. *Epidemics*, 10:63–67.

Bargiacchi, E., T. Verstraeten, D. M. Roijers, and A. Nowé
2020. Model-based multi-agent reinforcement learning with cooperative prioritized sweeping. *arXiv preprint 2001.07527*.

Barrat, A., M. Barthelemy, R. Pastor-Satorras, and A. Vespignani
2004. The architecture of complex weighted networks. *Proceedings of the national academy of sciences*, 101(11):3747–3752.

Barrett, A. D.
2017. Yellow fever live attenuated vaccine: a very successful live attenuated vaccine but still we have problems controlling the disease. *Vaccine*, 35(44):5951–5955.

Barthélemy, M., C. Godreche, and J.-M. Luck
2010. Fluctuation effects in metapopulation models: percolation and pandemic threshold. *Journal of theoretical biology*, 267(4):554–564.

Basta, N. E., D. L. Chao, M. E. Halloran, L. Matrajt, and I. M. Longini
2009. Strategies for pandemic and seasonal influenza vaccination of schoolchildren in the United States. *American journal of epidemiology*, 170(6):679–686.

Bechhofer, R. E.
1958. A sequential multiple-decision procedure for selecting the best one of several normal populations with a common unknown variance, and its use with various experimental designs. *Biometrics*, 14(3):408–429.

Bedford, J., J. Farrar, C. Ihekweazu, G. Kang, M. Koopmans, and J. Nkengasong
2019. A new twenty-first century science for effective epidemic response. *Nature*, 575(7781):130–136.

Behbehani, A. M.
1983. The smallpox story: life and death of an old disease. *Microbiological reviews*, 47(4):455.

Belshe, R. B.
2005. The origins of pandemic influenza—lessons from the 1918 virus. *New England Journal of Medicine*, 353(21):2209–2211.

Belshe, R. B.
2010. The need for quadrivalent vaccine against seasonal influenza. *Vaccine*, 28:D45–D53.

Bertsekas, D. P.
1997. Nonlinear programming. *Journal of the Operational Research Society*, 48(3):334–334.

Bertsekas, D. P. and J. N. Tsitsiklis
2002. *Introduction to probability*, volume 1. Athena Scientific Belmont, MA.

Billingsley, P.
2008. *Probability and measure*. John Wiley and Sons.

Blondel, V. D., J.-L. Guillaume, R. Lambiotte, and E. Lefebvre
2008. Fast unfolding of communities in large networks. *Journal of statistical mechanics: theory and experiment*, 2008(10):P10008.

Böhmer, W., V. Kurin, and S. Whiteson
2019. Deep coordination graphs. *arXiv preprint arXiv:1910.00091*.

Boyer, N. and P. Marcellin
2000. Pathogenesis, diagnosis and management of hepatitis c. *Journal of hepatology*, 32:98–112.

Breman, J. G. and I. Arita
1980. The confirmation and maintenance of smallpox eradication. *New England Journal of Medicine*, 303(22):1263–1273.

Britton, T.
2010. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35.

Britton, T. and G. Scalia Tomba
2019. Estimation in emerging epidemics: Biases and remedies. *Journal of the Royal Society Interface*, 16(150):20180670.

Brown, S. T., J. H. Tai, R. R. Bailey, P. C. Cooley, W. D. Wheaton, M. A. Potter, R. E. Voorhees, M. LeJeune, J. J. Grefenstette, D. S. Burke, et al.
2011. Would school closure for the 2009 h1n1 influenza epidemic have been worth the cost?: a computational simulation of pennsylvania. *BMC public health*, 11(1):353.

Brys, T.
2016. *Reinforcement Learning with Heuristic Information*. PhD thesis.

Bubeck, S., R. Munos, and G. Stoltz
2009. Pure exploration in multi-armed bandits problems. In *International conference on Algorithmic learning theory*, Pp. 23–37. Springer.

Bubeck, S., R. Munos, and G. Stoltz
2011. Pure exploration in finitely-armed and continuous-armed bandits. *Theoretical Computer Science*, 412(19):1832–1852.

Bubeck, S., T. Wang, and N. Viswanathan
2013. Multiple identifications in multi-armed bandits. In *International Conference on Machine Learning*, Pp. 258–265.

Cai, X.
2007. Exact stochastic simulation of coupled chemical reactions with delays. *The Journal of chemical physics*, 126(12):124108.

Caminade, C., J. M. Medlock, E. Ducheyne, K. M. McIntyre, S. Leach, M. Baylis, and A. P. Morse
2012. Suitability of european climate for the asian tiger mosquito aedes albopictus: recent trends and future scenarios. *Journal of the Royal Society Interface*, 9(75):2708–2717.

Cao, Y., D. T. Gillespie, and L. R. Petzold
2006. Efficient step size selection for the tau-leaping simulation method. *The Journal of chemical physics*, 124(4):044109.

Caruana, R. and A. Niculescu-Mizil
2006. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, Pp. 161–168. ACM.

Cauchemez, S., N. M. Ferguson, C. Wachtel, A. Tegnell, G. Saour, B. Duncan, and A. Nicoll
2009. Closure of schools during an influenza pandemic. *The Lancet infectious diseases*, 9(8):473–481.

Cauchemez, S., A.-J. Valleron, P.-Y. Boelle, A. Flahault, and N. M. Ferguson
2008. Estimating the impact of school closure on influenza transmission from sentinel data. *Nature*, 452(7188):750.

Chao, D. L., M. E. Halloran, V. J. Obenchain, and I. M. Longini Jr
2010. FluTE, a publicly available stochastic influenza epidemic simulation model. *PLoS Computational Biology*, 6(1):e1000656.

Chao, D. L., S. B. Halstead, M. E. Halloran, and I. M. Longini
2012. Controlling Dengue with Vaccines in Thailand. *PLoS Neglected Tropical Diseases*, 6(10).

Chapelle, O. and L. Li
2011. An empirical evaluation of thompson sampling. In *Advances in neural information processing systems*, Pp. 2249–2257.

Chen, I., R. Cooney, R. G. Feachem, A. Lal, and W. Mpanju-Shumbusho
2018. The lancet commission on malaria eradication. *The Lancet*, 391(10130):1556–1558.

Chowell, G., S. Echevarría-Zuno, C. Viboud, L. Simonsen, J. Tamerius, M. A. Miller, and V. H. Borja-Aburto
2011. Characterizing the epidemiology of the 2009 influenza a/h1n1 pandemic in mexico. *PLoS medicine*, 8(5).

Ciavarella, C., L. Fumanelli, S. Merler, C. Cattuto, and M. Ajelli
2016. School closure policies at municipality level for mitigating influenza spread: a model-based evaluation. *BMC infectious diseases*, 16(1):576.

Cinlar, E.
2013. *Introduction to stochastic processes*. Courier Corporation.

Clardy, J., M. A. Fischbach, and C. R. Currie
2009. The natural history of antibiotics. *Current biology*, 19(11):R437–R441.

Coffee, M., M. N. Lurie, and G. P. Garnett
2007. Modelling the impact of migration on the hiv epidemic in south africa. *Aids*, 21(3):343–350.

Collobert, R. and J. Weston
2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, Pp. 160–167.

Copeland, D. L., R. Basurto-Davila, W. Chung, A. Kurian, D. B. Fishbein, P. Szymanowski, J. Zipprich, H. Lipman, M. S. Cetron, M. I. Meltzer, et al.
2013. Effectiveness of a school district closure for pandemic influenza a (h1n1) on acute respiratory illnesses in the community: a natural experiment. *Clinical infectious diseases*, 56(4):509–516.

Cuypers, L., P. Libin, P. Simmonds, A. Nowé, J. Muñoz-Jordán, L. Alcantara, A.-M. Vandamme, G. Santiago, and K. Theys
2018. Time to harmonize dengue nomenclature and classification. *Viruses*, 10(10):569.

Dawood, F. S., A. D. Iuliano, C. Reed, M. I. Meltzer, D. K. Shay, P.-Y. Cheng, D. Bandaranayake, R. F. Breiman, W. A. Brooks, P. Buchy, et al.
2012. Estimated global mortality associated with the first 12 months of 2009 pandemic influenza a h1n1 virus circulation: a modelling study. *The Lancet infectious diseases*, 12(9):687–695.

De Luca, G., K. Van Kerckhove, P. Coletti, C. Poletto, N. Bossuyt, N. Hens, and V. Colizza
2018. The impact of regular school closure on seasonal influenza epidemics: a data-driven spatial transmission model for belgium. *BMC infectious diseases*, 18(1):29.

De Silva, U., J. Warachit, S. Waicharoen, and M. Chittaganpitch
2009. A preliminary analysis of the epidemiology of influenza a (h1n1) v virus infection in thailand from early outbreak data, june-july 2009. *Eurosurveillance*, 14(31):19292.

Defferrard, M., X. Bresson, and P. Vandergheynst
2016. Convolutional neural networks on graphs with fast localized spectral filtering. In *Advances in neural information processing systems*, Pp. 3844–3852.

Diekmann, O., H. Heesterbeek, and T. Britton
2012. *Mathematical tools for understanding infectious disease dynamics*, volume 7. Princeton University Press.

Diekmann, O., J. Heesterbeek, and M. Roberts
2009. The construction of next-generation matrices for compartmental epidemic models. *Journal of the Royal Society Interface*, P. rsif20090386.

Dorigatti, I., S. Cauchemez, A. Pugliese, and N. M. Ferguson
2012. A new approach to characterising infectious disease transmission dynamics from sentinel surveillance: application to the Italian 2009/2010 A/H1N1 influenza pandemic. *Epidemics*, 4(1):9–21.

Drummond, A. J., A. Rambaut, B. Shapiro, and O. G. Pybus
2005. Bayesian coalescent inference of past population dynamics from molecular sequences. *Molecular biology and evolution*, 22(5):1185–1192.

Dudas, G., L. M. Carvalho, T. Bedford, A. J. Tatem, G. Baele, N. R. Faria, D. J. Park, J. T. Ladner, A. Arias, D. Asogun, et al.
2017. Virus genomes reveal factors that spread and sustained the ebola epidemic. *Nature*, 544(7650):309.

Duong, V., L. Lambrechts, R. E. Paul, S. Ly, R. S. Lay, K. C. Long, R. Huy, A. Tarantola, T. W. Scott, A. Sakuntabhai, et al.
2015. Asymptomatic humans transmit dengue virus to mosquitoes. *Proceedings of the National Academy of Sciences*, 112(47):14688–14693.

D'Silva, J. P. and M. C. Eisenberg
2017. Modeling spatial invasion of ebola in west africa. *Journal of theoretical biology*, 428:65–75.

Eames, K. T., N. L. Tilston, E. Brooks-Pollock, and W. J. Edmunds
2012. Measured dynamic social contact patterns explain the spread of h1n1v influenza. *PLoS computational biology*, 8(3):e1002425.

Earn, D. J., D. He, M. B. Loeb, K. Fonseca, B. E. Lee, and J. Dushoff
2012. Effects of school closure on incidence of pandemic influenza in alberta, canada. *Annals of internal medicine*, 156(3):173–181.

Eggo, R. M., S. Cauchemez, and N. M. Ferguson
2010. Spatial dynamics of the 1918 influenza pandemic in england, wales and the united states. *Journal of the Royal Society Interface*, 8(55):233–243.

Enserink, M.
2004. Crisis underscores fragility of vaccine production system. *Science*, 306(5695):385.

Eubank, S., V. Kumar, M. Marathe, A. Srinivasan, and N. Wang
2006. Structure of social contact networks and their impact on epidemics. *DIMACS Series in Discrete Mathematics and Theoretical Computer Science*, 70(0208005):181.

Even-Dar, E., S. Mannor, and Y. Mansour
2002. PAC bounds for multi-armed bandit and markov decision processes. In *International Conference on Computational Learning Theory*, Pp. 255–270. Springer.

Even-Dar, E., S. Mannor, and Y. Mansour
2006. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *Journal of machine learning research*, 7(Jun):1079–1105.

Féraud, R., R. Allesiardo, T. Urvoy, and F. Clérot
2016. Random forest for the contextual bandit problem. In *Artificial Intelligence and Statistics*, Pp. 93–101.

Ferguson, N. M., D. A. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke
2006. Strategies for mitigating an influenza pandemic. *Nature*, 442(7101):448.

Ferguson, N. M., D. A. T. Cummings, S. Cauchemez, C. Fraser, and Others
2005. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209.

Fischer, D., S. Thomas, M. Neteler, N. Tjaden, and C. Beierkuhnlein
2014. Climatic suitability of aedes albopictus in europe referring to climate change projections: comparison of mechanistic and correlative niche modelling approaches. *Eurosurveillance*, 19(6).

Foerster, J. N., G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson
2018. Counterfactual multi-agent policy gradients. In *Thirty-second AAAI conference on artificial intelligence*.

Fortunato, S.
2010. Community detection in graphs. *Physics reports*, 486(3-5):75–174.

Fraser, C., D. A. T. Cummings, D. Klinkenberg, D. S. Burke, and N. M. Ferguson
2011. Influenza transmission in households during the 1918 pandemic. *American journal of epidemiology*, 174(5):505–514.

Fraser, C., C. A. Donnelly, S. Cauchemez, W. P. Hanage, M. D. Van Kerkhove, T. D. Hollingsworth, J. Griffin, R. F. Baggaley, H. E. Jenkins, E. J. Lyons, et al.
2009. Pandemic potential of a strain of influenza a (h1n1): early findings. *science*, 324(5934):1557–1561.

Fumanelli, L., M. Ajelli, P. Manfredi, A. Vespignani, and S. Merler
2012. Inferring the structure of social contacts from demographic data in the analysis of infectious diseases spread. *PLoS computational biology*, 8(9):e1002673.

Fumanelli, L., M. Ajelli, S. Merler, N. M. Ferguson, and S. Cauchemez
2016. Model-Based Comprehensive Analysis of School Closure Policies for Mitigating Influenza Epidemics and Pandemics. *PLoS Computational Biology*, 12(1).

Fung, I. C.-H., M. Gambhir, J. W. Glasser, H. Gao, M. L. Washington, A. Uzicanin, and M. I. Meltzer
2015. Modeling the effect of school closures in a pandemic scenario: exploring two different contact matrices. *Clinical Infectious Diseases*, 60(suppl_1):S58–S63.

Gabillon, V., M. Ghavamzadeh, and A. Lazaric
2012. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Advances in Neural Information Processing Systems*, Pp. 3212–3220.

Galimand, M., E. Carniel, and P. Courvalin
2006. Resistance of yersinia pestis to antimicrobial agents. *Antimicrobial agents and chemotherapy*, 50(10):3233–3236.

Garivier, A. and E. Kaufmann
2016. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory*, Pp. 998–1027.

George, D. B., C. T. Webb, M. L. Farnsworth, T. J. O'Shea, R. A. Bowen, D. L. Smith, T. R. Stanley, L. E. Ellison, and C. E. Rupprecht
2011. Host and viral ecology determine bat rabies seasonality and maintenance. *Proceedings of the National Academy of Sciences*, 108(25):10208–10213.

Geretti, A. M.
2006. Hiv-1 subtypes: epidemiology and significance for hiv management. *Current opinion in infectious diseases*, 19(1):1–7.

Germann, T. C., H. Gao, M. Gambhir, A. Plummer, M. Biggerstaff, C. Reed, and A. Uzicanin
2019. School dismissal as a pandemic influenza response: When, where and for how long? *Epidemics*, P. 100348.

Germann, T. C., K. Kadau, I. M. Longini, and C. A. Macken
2006. Mitigation strategies for pandemic influenza in the United States. *Proceedings of the National Academy of Sciences*, 103(15):5935–5940.

Gibson, M. A. and J. Bruck
2000. Efficient exact stochastic simulation of chemical systems with many species and many channels. *The journal of physical chemistry A*, 104(9):1876–1889.

Gillespie, D. T.
1977. Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361.

Glass, R. J., L. M. Glass, W. E. Beyeler, and H. J. Min
2006. Targeted social distancing design for pandemic influenza. *Emerging infectious diseases*, 12(11):1671–1681.

Gog, J. R., S. Ballesteros, C. Viboud, L. Simonsen, O. N. Bjornstad, J. Shaman, D. L. Chao, F. Khan, and B. T. Grenfell
2014. Spatial transmission of 2009 pandemic influenza in the us. *PLoS computational biology*, 10(6):e1003635.

Gopalan, A., S. Mannor, and Y. Mansour
2014. Thompson sampling for complex online problems. In *International Conference on Machine Learning*, Pp. 100–108.

Grover, A., T. Markov, P. Attia, N. Jin, N. Perkins, B. Cheong, M. Chen, Z. Yang, S. Harris, W. Chueh, and S. Ermon
2018. Best arm identification in multi-armed bandits with delayed feedback. In *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, A. Storkey and F. Perez-Cruz, eds., volume 84 of *Proceedings of Machine Learning Research*, Pp. 833–842, Playa Blanca, Lanzarote, Canary Islands. PMLR.

Grubaugh, N. D., J. T. Ladner, P. Lemey, O. G. Pybus, A. Rambaut, E. C. Holmes, and K. G. Andersen
2019. Tracking virus outbreaks in the twenty-first century. *Nature microbiology*, 4(1):10–19.

Guha, S. and K. Munagala
2014. Stochastic regret minimization via thompson sampling. In *Conference on Learning Theory*, Pp. 317–338.

Haber, M. J., D. K. Shay, X. M. Davis, R. Patel, X. Jin, E. Weintraub, E. Orenstein, and W. W. Thompson
2007. Effectiveness of interventions to reduce contact rates during a simulated influenza pandemic. *Emerging infectious diseases*, 13(4):581.

Hadinegoro, S. R., J. L. Arredondo-García, M. R. Capeding, C. Deseda, T. Chotpitayasunondh, R. Dietze, H. H. M. Ismail, H. Reynales, K. Limkittikul, D. M. Rivera-Medina, H. N. Tran, A. Bouckenooghe, D. Chansinghakul, M. Cortés, K. Fanouillere, R. Forrat, C. Frago, S. Gailhardou, N. Jackson, F. Noriega, E. Plennevaux, T. A. Wartel, B. Zambrano, and M. Saville
2015. Efficacy and Long-Term Safety of a Dengue Vaccine in Regions of Endemic Disease. *New England Journal of Medicine*, 373(13):1195–1206.

Halder, N., J. K. Kelso, and G. J. Milne
2010. Developing guidelines for school closure interventions to be used during a future influenza pandemic. *BMC infectious diseases*, 10(1):221.

Halloran, M. E., N. M. Ferguson, S. Eubank, I. M. Longini, D. A. T. Cummings, B. Lewis, S. Xu, C. Fraser, A. Vullikanti, T. C. Germann, and Others
2008. Modeling targeted layered containment of an influenza pandemic in the United States. *Proceedings of the National Academy of Sciences*, 105(12):4639–4644.

Halloran, M. E., I. M. Longini, A. Nizam, and Y. Yang
2002. Containing bioterrorist smallpox. *Science (New York, N.Y.)*, 298(5597):1428–1432.

Hanski, I. et al.
1999. *Metapopulation ecology*. Oxford University Press.

Hartfield, M. and S. Alizon
2013. Introducing the outbreak threshold in epidemiology. *PLoS Pathog*, 9(6):e1003277.

Hastings, W. K.
1970. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57.

Hemelaar, J., E. Gouws, P. D. Ghys, and S. Osmanov
2006. Global and regional distribution of hiv-1 genetic subtypes and recombinants in 2004. *Aids*, 20(16):W13–W23.

Herbeck, J. T., J. E. Mittler, G. S. Gottlieb, and J. I. Mullins
2014. An hiv epidemic model based on viral load dynamics: value in assessing empirical trends in hiv virulence and community viral load. *PLoS computational biology*, 10(6):e1003673.

Hernandez-Leal, P., B. Kartal, and M. E. Taylor
2019. A survey and critique of multiagent deep reinforcement learning. *Autonomous Agents and Multi-Agent Systems*, Pp. 1–48.

Heymann, A., G. Chodick, B. Reichman, E. Kokia, and J. Laufer
2004. Influence of school closure on the incidence of viral respiratory diseases among children and on health care utilization. *The Pediatric infectious disease journal*, 23(7):675–677.

Hinton, G. E., T. J. Sejnowski, and T. A. Poggio
1999. *Unsupervised learning: foundations of neural computation*. MIT press.

Hoffman, M., B. Shahriari, and N. Freitas
2014. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Artificial Intelligence and Statistics*, Pp. 365–374.

Hofmann, W. P. and S. Zeuzem
2011. A new standard of care for the treatment of chronic hcv infection. *Nature Reviews Gastroenterology & Hepatology*, 8(5):257.

Holmes, E. C.
2004. The phylogeography of human viruses. *Molecular ecology*, 13(4):745–756.

Holzmann, H. and U. Wiedermann
2019. Mandatory vaccination: suited to enhance vaccination coverage in europe? *Eurosurveillance*, 24(26):1900376.

Honda, J. and A. Takemura
2014. Optimality of Thompson Sampling for Gaussian Bandits Depends on Priors. In *AISTATS*, Pp. 375–383.

Huber, J. H., M. L. Childs, J. M. Caldwell, and E. A. Mordecai
2018. Seasonal temperature variation influences climate suitability for dengue, chikungunya, and zika transmission. *PLoS neglected tropical diseases*, 12(5):e0006451.

Iqbal, S. and F. Sha
2019. Actor-attention-critic for multi-agent reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*.

Jamieson, K., M. Malloy, R. Nowak, and S. Bubeck
2014. lil'ucb: An optimal exploration algorithm for multi-armed bandits. In *Conference on Learning Theory*, Pp. 423–439.

Jamieson, K. and A. Talwalkar
2016. Non-stochastic best arm identification and hyperparameter optimization. In *Artificial Intelligence and Statistics*, Pp. 240–248.

Jamieson, K. G., L. Jain, C. Fernandez, N. J. Glattard, and R. Nowak
2015. Next: A system for real-world development, evaluation, and application of active learning. In *Advances in Neural Information Processing Systems*, Pp. 2656–2664.

Jaynes, E. T.
1968. Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3):227–241.

Jennison, C., I. M. Johnstone, and B. W. Turnbull
1982. Asymptotically optimal procedures for sequential adaptive selection of the best of several normal means. *Statistical decision theory and related topics III*, 2:55–86.

Jernigan, J. A., D. S. Stephens, D. A. Ashford, C. Omenaca, M. S. Topiel, M. Galbraith, M. Tapper, T. L. Fisk, S. Zaki, T. Popovic, et al.
2001. Bioterrorism-related inhalational anthrax: the first 10 cases reported in the united states. *Emerging infectious diseases*, 7(6):933.

Jiang, J., C. Dun, and Z. Lu
2018. Graph convolutional reinforcement learning for multi-agent cooperation. *arXiv preprint arXiv:1810.09202*.

Jiang, J. and Z. Lu
2018. Learning attentional communication for multi-agent cooperation. In *Advances in Neural Information Processing Systems 31*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, eds., Pp. 7254–7264. Curran Associates, Inc.

Johnston, M. I. and A. S. Fauci
2008. An hiv vaccine—challenges and prospects. *New England Journal of Medicine*, 359(9):888–890.

Jun, K.-S. and R. D. Nowak
2016. Anytime exploration for multi-armed bandits using confidence information. In *33rd International Conference on Machine Learning*, Pp. 974–982.

Kadane, J. B. et al.
2016. Sums of possibly associated bernoulli variables: The conway–maxwell-binomial distribution. *Bayesian Analysis*, 11(2):403–420.

Kalyanakrishnan, S., A. Tewari, P. Auer, and P. Stone
2012. PAC subset selection in stochastic multi-armed bandits. In *ICML*, volume 12, Pp. 655–662.

Karim, S. S. A. and Q. A. Karim
2011. Antiretroviral prophylaxis: a defining moment in hiv control. *The Lancet*, 378(9809):e23–e25.

Karnin, Z., T. Koren, and O. Somekh
2013. Almost optimal exploration in multi-armed bandits. In *International Conference on Machine Learning*, Pp. 1238–1246.

Kasaie, P., S. A. Berry, M. S. Shah, E. S. Rosenberg, K. W. Hoover, T. L. Gift, H. Chesson, J. Pennington, D. German, C. P. Flynn, et al.
2018. Impact of providing pre-exposure prophylaxis for hiv at clinics for sexually transmitted infections in baltimore city: an agent-based model. *Sexually transmitted diseases*, 45(12):791.

Kaufmann, E., O. Cappé, and A. Garivier
2016. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42.

Kaufmann, E. and S. Kalyanakrishnan
2013. Information complexity in bandit subset selection. In *Conference on Learning Theory*, Pp. 228–251.

Keeling, M., M. Woolhouse, R. May, G. Davies, and B. T. Grenfell
2003. Modelling vaccination strategies against foot-and-mouth disease. *Nature*, 421(6919):136.

Keeling, M. J. and P. Rohani
2011. *Modeling infectious diseases in humans and animals*. Princeton University Press.

Kermack, W. O. and A. G. McKendrick
1927. A contribution to the mathematical theory of epidemics. *Proceedings of the royal society of london. Series A, Containing papers of a mathematical and physical character*, 115(772):700–721.

Khan, A., E. Tolstaya, A. Ribeiro, and V. Kumar
2019. Graph policy gradients for large scale robot control. *arXiv preprint arXiv:1907.03822*.

King, A. A., M. Domenech de Cellès, F. M. Magpantay, and P. Rohani
2015. Avoidable errors in the modelling of outbreaks of emerging pathogens, with special reference to ebola. *Proceedings of the Royal Society B: Biological Sciences*, 282(1806):20150347.

Kissler, S. M., J. R. Gog, C. Viboud, V. Charu, O. N. Bjørnstad, L. Simonsen, and B. T. Grenfell
2019. Geographic transmission hubs of the 2009 influenza pandemic in the united states. *Epidemics*, 26:86–94.

Klepac, P., S. Kissler, and J. Gog
2018. Contagion! the bbc four pandemic–the model behind the documentary. *Epidemics*.

Klingen, T. R., S. Reimering, C. A. Guzmán, and A. C. McHardy
2018. In silico vaccine strain prediction for human influenza viruses. *Trends in microbiology*, 26(2):119–131.

Kocák, T., M. Valko, R. Munos, and S. Agrawal
2014. Spectral thompson sampling. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*.

Kraemer, M., N. Golding, D. Bisanzio, S. Bhatt, D. Pigott, S. Ray, O. Brady, J. Brownstein, N. Faria, D. Cummings, et al.
2019. Utilizing general human movement models to predict the spread of emerging infectious diseases in resource poor settings. *Scientific reports*, 9(1):5151.

Kubiak, R. J. and A. R. McLean
2012. Why was the 2009 influenza pandemic in england so small? *PloS one*, 7(2):e30223.

Kühnert, D., C.-H. Wu, and A. J. Drummond
2011. Phylogenetic and epidemic modeling of rapidly evolving infectious diseases. *Infection, genetics and evolution*, 11(8):1825–1841.

Kullback, S. and R. A. Leibler
1951. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.

Lam, S. K., A. Pitrou, and S. Seibert
2015. Numba: A llvm-based python jit compiler. In *Proceedings of the Second Workshop on the LLVM Compiler Infrastructure in HPC*, P. 7. ACM.

LeCun, Y., Y. Bengio, and G. Hinton
2015. Deep learning. *nature*, 521(7553):436.

Leicht, E. A. and M. E. Newman
2008. Community structure in directed networks. *Physical review letters*, 100(11):118703.

Lekone, P. E. and B. F. Finkenstädt
2006. Statistical inference in a stochastic epidemic seir model with control intervention: Ebola as a case study. *Biometrics*, 62(4):1170–1177.

Lemey, P., A. Rambaut, A. J. Drummond, and M. A. Suchard
2009. Bayesian phylogeography finds its roots. *PLoS computational biology*, 5(9):e1000520.

Leung, N. H., C. Xu, D. K. Ip, and B. J. Cowling
2015. The fraction of influenza virus infections that are asymptomatic: a systematic review and meta-analysis. *Epidemiology (Cambridge, Mass.)*, 26(6):862.

Lewis, N. S., C. A. Russell, P. Langat, T. K. Anderson, K. Berger, F. Bielejec, D. F. Burke, G. Dudas, J. M. Fonville, R. A. Fouchier, et al.
2016. The global antigenic diversity of swine influenza a viruses. *elife*, 5:e12217.

Lewis, T. G.
2011. *Network science: Theory and applications*. John Wiley & Sons.

Liljeros, F., C. R. Edling, L. A. N. Amaral, H. E. Stanley, and Y. Åberg
2001. The web of human sexual contacts. *Nature*, 411(6840):907.

Lima, V. D., K. Johnston, R. S. Hogg, A. R. Levy, P. R. Harrigan, A. Anema, and J. S. Montaner
2008. Expanded access to highly active antiretroviral therapy: a potentially powerful strategy to curb the growth of the hiv epidemic. *The Journal of infectious diseases*, 198(1):59–67.

Lipsitch, M., T. Cohen, M. Murray, and B. R. Levin
2007. Antiviral resistance and the control of pandemic influenza. *PLoS medicine*, 4(1):e15.

Liu, E. and X. Yan
2019. New parameter-free mobility model: Opportunity priority selection model. *Physica A: Statistical Mechanics and its Applications*, 526:121023.

Liu, Y., W. Wang, Y. Hu, J. Hao, X. Chen, and Y. Gao
2019. Multi-agent game abstraction via graph attention neural network. *arXiv preprint arXiv:1810.09202*.

Lloyd-Smith, J. O., S. J. Schreiber, P. E. Kopp, and W. M. Getz
2005. Superspreading and the effect of individual variation on disease emergence. *Nature*, 438(7066):355–359.

Longini, I. M., A. Nizam, S. Xu, K. Ungchusak, W. Hanshaoworakul, D. A. Cummings, and M. E. Halloran
2005. Containing pandemic influenza at the source. *Science*, 309(5737):1083–1087.

Luk, J., P. Gross, and W. W. Thompson
2001. Observations on mortality during the 1918 influenza pandemic. *Clinical Infectious Diseases*, 33(8):1375–1378.

Lunn, D., C. Jackson, N. Best, D. Spiegelhalter, and A. Thomas
2012. *The BUGS book: A practical introduction to Bayesian analysis*. Chapman and Hall/CRC.

MacDonald, N. E. et al.
2015. Vaccine hesitancy: Definition, scope and determinants. *Vaccine*, 33(34):4161–4164.

Marini, G., G. Guzzetta, C. A. M. Toledo, M. Teixeira, R. Rosà, and S. Merler
2019. Effectiveness of ultra-low volume insecticide spraying to prevent dengue in a non-endemic metropolitan area of brazil. *PLoS computational biology*, 15(3):e1006831.

Markel, H., H. B. Lipman, J. A. Navarro, A. Sloan, J. R. Michalsen, A. M. Stern, and M. S. Cetron
2007. Nonpharmaceutical interventions implemented by us cities during the 1918-1919 influenza pandemic. *Jama*, 298(6):644–654.

Marseille, E., P. B. Hofmann, and J. G. Kahn
2002. Hiv prevention before haart in sub-saharan africa. *The Lancet*, 359(9320):1851–1856.

Masucci, A. P., J. Serras, A. Johansson, and M. Batty
2013. Gravity versus radiation models: On the importance of scale and heterogeneity in commuting flows. *Physical Review E*, 88(2):022812.

Maxmen, A.
2019. Science under fire: Ebola researchers fight to test drugs and vaccines in a war zone. *Nature*, 572(7767):16.

Mazzoli, M., A. Molas, A. Bassolas, M. Lenormand, P. Colet, and J. J. Ramasco
2019. Field theory for recurrent mobility. *Nature communications*, 10(1):1–10.

Medlock, J. and A. P. Galvani
2009. Optimizing influenza vaccine distribution. *Science*, 325(5948):1705–1708.

Mellor, J. C.
2014. *Decision Making Using Thompson Sampling*. PhD thesis, University of Manchester.

Meltzer, M. I., N. J. Cox, and K. Fukuda
1999. The economic impact of pandemic influenza in the united states: priorities for intervention. *Emerging infectious diseases*, 5(5):659.

Merler, S., M. Ajelli, A. Pugliese, and N. M. Ferguson
2011. Determinants of the spatiotemporal dynamics of the 2009 h1n1 pandemic in europe: implications for real-time modelling. *PLoS computational biology*, 7(9):e1002205.

Messina, J. P., O. J. Brady, N. Golding, M. U. Kraemer, G. W. Wint, S. E. Ray, D. M. Pigott, F. M. Shearer, K. Johnson, L. Earl, et al.
2019. The current and future global distribution and population at risk of dengue. *Nature microbiology*, P. 1.

Metcalf, C. J. E. and J. Lessler
2017. Opportunities and challenges in modeling emerging infectious diseases. *Science*, 357(6347):149–152.

Miller, E., K. Hoschler, P. Hardelid, E. Stanford, N. Andrews, and M. Zambon
2010. Incidence of 2009 pandemic influenza a h1n1 infection in england: a cross-sectional serological study. *The Lancet*, 375(9720):1100–1108.

Mills, C. E., J. M. Robins, and M. Lipsitch
2004. Transmissibility of 1918 pandemic influenza. *Nature*, 432(7019):904.

Mills, H. L. and S. Riley
2014. The spatial resolution of epidemic peaks. *PLoS computational biology*, 10(4):e1003561.

Milne, G. J., N. Halder, and J. K. Kelso
2013. The cost effectiveness of pandemic influenza interventions: a pandemic severity based analysis. *PloS one*, 8(4).

Minsky, M.
1954. Neural nets and the brain model problem (ph. d. dissertation). *Princeton Univ*.

Mitchell, T. M. et al.
1997. Machine learning.

Mnih, V., K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, et al.
2015. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

Moore, R. D. and R. E. Chaisson
1999. Natural history of hiv infection in the_era of combination antiretroviral therapy. *Aids*, 13(14):1933–1942.

Moreira, L. A., I. Iturbe-Ormaetxe, J. A. Jeffery, G. Lu, A. T. Pyke, L. M. Hedges, B. C. Rocha, S. Hall-Mendelin, A. Day, M. Riegler, et al.
2009. A wolbachia symbiont in aedes aegypti limits infection with dengue, chikungunya, and plasmodium. *Cell*, 139(7):1268–1278.

Mossong, J., N. Hens, M. Jit, P. Beutels, K. Auranen, R. Mikolajczyk, M. Massari, S. Salmaso, G. S. Tomba, J. Wallinga, J. Heijne, M. Sadkowska-Todys, M. Rosinska, and W. J. Edmunds
2008. Social contacts and mixing patterns relevant to the spread of infectious diseases. *PLOS Medicine*, 5(3):1–1.

Nair, V. and G. E. Hinton
2010. Rectified linear units improve restricted boltzmann machines. In *Proceedings of the 27th international conference on machine learning (ICML-10)*, Pp. 807–814.

Nakagawa, F., R. K. Lodwick, C. J. Smith, R. Smith, V. Cambiano, J. D. Lundgren, V. Delpech, and A. N. Phillips
2012. Projected life expectancy of people with hiv according to timing of diagnosis. *Aids*, 26(3):335–343.

Nelson, M. I., L. Simonsen, C. Viboud, M. A. Miller, and E. C. Holmes
2007. Phylogenetic analysis reveals the global migration of seasonal influenza a viruses. *PLoS pathogens*, 3(9):e131.

Nicholls, H.
2006. Pandemic influenza: the inside story. *PLoS Biology*, 4(2):e50.

Nishiura, H., G. Chowell, M. Safan, and C. Castillo-Chavez
2010. Pros and cons of estimating the reproduction number from early epidemic growth rate of influenza a (h1n1) 2009. *Theoretical Biology and Medical Modelling*, 7(1):1.

Oliphant, T.
2006. *Guide to NumPy*.

Osaki, S.
2012. *Applied stochastic system modeling*. Springer Science & Business Media.

Osband, I., D. Russo, and B. Van Roy
2013. (more) efficient reinforcement learning via posterior sampling. In *Advances in Neural Information Processing Systems*, Pp. 3003–3011.

Pandey, A., K. E. Atkins, J. Medlock, N. Wenzel, J. P. Townsend, J. E. Childs, T. G. Nyenswah, M. L. Ndeffo-Mbah, and A. P. Galvani
2014. Strategies for containing ebola in west africa. *Science*, 346(6212):991–995.

Parkhill, J., B. Wren, N. Thomson, R. Titball, M. Holden, M. Prentice, M. Sebaihia, K. James, C. Churcher, K. Mungall, et al.
2001. Genome sequence of yersinia pestis, the causative agent of plague. *Nature*, 413(6855):523.

Paules, C. and K. Subbarao
2017. Influenza. *The Lancet*, Pp. 697–708.

Petersen, L. R., D. J. Jamieson, A. M. Powers, and M. A. Honein
2016. Zika virus. *New England Journal of Medicine*, 374(16):1552–1563.

Pigott, D. M., N. Golding, A. Mylne, Z. Huang, A. J. Henry, D. J. Weiss, O. J. Brady, M. U. Kraemer, D. L. Smith, C. L. Moyes, et al.
2014. Mapping the zoonotic niche of ebola virus disease in africa. *Elife*, 3:e04395.

Pilosof, S., M. A. Porter, M. Pascual, and S. Kéfi
2017. The multilayer nature of ecological networks. *Nature Ecology & Evolution*, 1(4):0101.

Piot, P., H. J. Larson, K. L. O'Brien, J. N'kengasong, E. Ng, S. Sow, and B. Kampmann
2019. Immunization: vital progress, unfinished agenda. *Nature*, 575(7781):119–129.

Potter, G. E., M. S. Handcock, I. M. Longini Jr, and M. E. Halloran
2012. Estimating within-school contact networks to understand influenza transmission. *The annals of applied statistics*, 6(1):1.

Powell, W. B. and I. O. Ryzhov
2012. *Optimal learning*, volume 841. John Wiley & Sons.

Prem, K., A. R. Cook, and M. Jit
2017. Projecting social contact matrices in 152 countries using contact surveys and demographic data. *PLOS Computational Biology*, 13(9):1–21.

Prentice, M. B. and L. Rahalison
2007. Plague. *The Lancet*, 369(9568):1196–1207.

Probert, W. J., S. Lakkur, C. J. Fonnesbeck, K. Shea, M. C. Runge, M. J. Tildesley, and M. J. Ferrari
2019. Context matters: using reinforcement learning to develop human-readable, state-dependent outbreak response policies. *Philosophical Transactions of the Royal Society B*, 374(1776):20180277.

Qin, C., D. Klabjan, and D. Russo
2017. Improving the expected improvement algorithm. In *Advances in Neural Information Processing Systems*, Pp. 5381–5391.

Radivojević, M. and J. Grujić
2017. Community structure of copper supply networks in the prehistoric balkans: An independent evaluation of the archaeological record from the 7th to the 4th millennium bc. *Journal of Complex Networks*, 6(1):106–124.

Rappuoli, R., M. Pizza, G. Del Giudice, and E. De Gregorio
2014. Vaccines, new opportunities for a new society. *Proceedings of the National Academy of Sciences*, 111(34):12288–12293.

Rashid, T., M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson
2018. QMIX: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, J. Dy and A. Krause, eds., volume 80 of *Proceedings of Machine Learning Research*, Pp. 4295–4304, Stockholmsmässan, Stockholm Sweden. PMLR.

Rasmussen, D. A., O. Ratmann, and K. Koelle
2011. Inference for nonlinear epidemiological models using genealogies and time series. *PLoS computational biology*, 7(8):e1002136.

Reichardt, J. and S. Bornholdt
2006. Statistical mechanics of community detection. *Physical Review E*, 74(1):016110.

Robert, A., S. Funk, and A. J. Kucharski
2019. The measles crisis in europe—the need for a joined-up approach. *The Lancet*, 393(10185):2033.

Robert, C.
2007. *The Bayesian choice: from decision-theoretic foundations to computational implementation*. Springer Science and Business Media.

Roijers, D. M., P. Vamplew, S. Whiteson, and R. Dazeley
2013. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48:67–113.

Roijers, D. M., L. M. Zintgraf, and A. Nowé
2017. Interactive thompson sampling for multi-objective multi-armed bandits. In *International Conference on Algorithmic Decision Theory*, Pp. 18–34. Springer.

Rosenblatt, F.
1958. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams
1986. Learning representations by back-propagating errors. *nature*, 323(6088):533–536.

Russell, S. J. and P. Norvig
2016. *Artificial intelligence: a modern approach*. Malaysia; Pearson Education Limited,.

Russo, D.
2016. Simple bayesian algorithms for best arm identification. In *Conference on Learning Theory*, Pp. 1417–1418.

Russo, D. and B. Van Roy
2013. Eluder dimension and the sample complexity of optimistic exploration. In *Advances in Neural Information Processing Systems*, Pp. 2256–2264.

Russo, D. and B. Van Roy
2016. An information-theoretic analysis of thompson sampling. *The Journal of Machine Learning Research*, 17(1):2442–2471.

Sassi, F.
2006. Calculating qalys, comparing qaly and daly calculations. *Health policy and planning*, 21(5):402–408.

Savitzky, A. and M. J. Golay
1964. Smoothing and differentiation of data by simplified least squares procedures. *Analytical chemistry*, 36(8):1627–1639.

Schick, V., D. Herbenick, M. Reece, S. A. Sanders, B. Dodge, S. E. Middlestadt, and J. D. Fortenberry
2010. Sexual behaviors, condom use, and sexual health of americans over 50: implications for sexual health promotion for older adults. *The journal of sexual medicine*, 7:315–329.

Schlaifer, R. and H. Raiffa
1961. *Applied statistical decision theory*.

Schmid, B. V. and M. Kretzschmar
2012. Determinants of sexual network structure and their impact on cumulative network measures. *PLoS computational biology*, 8(4):e1002470.

Schulman, J., S. Levine, P. Abbeel, M. Jordan, and P. Moritz
2015. Trust region policy optimization. In *International conference on machine learning*, Pp. 1889–1897.

Schulman, J., F. Wolski, P. Dhariwal, A. Radford, and O. Klimov
2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Scott, S. L.
2010. A modern bayesian look at the multi-armed bandit. *Applied Stochastic Models in Business and Industry*, 26(6):639–658.

Silver, D., A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al.
2016. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484.

Simini, F., M. C. González, A. Maritan, and A.-L. Barabási
2012. A universal model for mobility and migration patterns. *Nature*, 484(7392):96.

Singh, M., P. Sarkhel, G. J. Kang, A. Marathe, K. Boyle, P. Murray-Tuite, K. M. Abbas, and S. Swarup
2019. Impact of demographic disparities in social distancing and vaccination on influenza epidemics in urban and rural regions of the united states. *BMC infectious diseases*, 19(1):221.

Slepoy, A., A. P. Thompson, and S. J. Plimpton
2008. A constant-time kinetic monte carlo algorithm for simulation of large biochemical reaction networks. *The journal of chemical physics*, 128(20):05B618.

Soulsby, R. L. and J. A. Thomas
2012. Insect population curves: modelling and application to butterfly transect data. *Methods in Ecology and Evolution*, 3(5):832–841.

Spinney, L.
2017. *Pale rider: the Spanish flu of 1918 and how it changed the world*. PublicAffairs.

Spyrou, M. A., M. Keller, R. I. Tukhbatova, C. L. Scheib, E. A. Nelson, A. A. Valtueña, G. U. Neumann, D. Walker, A. Alterauge, N. Carty, et al.
2019. Phylogeography of the second plague pandemic revealed through analysis of historical yersinia pestis genomes. *Nature Communications*, 10(1):1–13.

Staras, S. A., S. C. Dollard, K. W. Radford, W. D. Flanders, R. F. Pass, and M. J. Cannon
2006. Seroprevalence of cytomegalovirus infection in the united states, 1988–1994. *Clinical Infectious Diseases*, 43(9):1143–1151.

Stein, M.
1987. Large sample properties of simulations using latin hypercube sampling. *Technometrics*, 29(2):143–151.

Steinberg, D.
2009. Cart: classification and regression trees. In *The top ten algorithms in data mining*, Pp. 193–216. Chapman and Hall/CRC.

Stöhr, K.
2002. Influenza: WHO cares. *The Lancet infectious diseases*, 2(9):517.

Stouffer, S. A.
1940. Intervening opportunities: a theory relating mobility and distance. *American sociological review*, 5(6):845–867.

Sutton, R. and A. Barto
1998. *Reinforcement learning: An introduction*, volume 1. Cambridge Univ Press.

Taubenberger, J. K. and D. M. Morens
2006. 1918 influenza: the mother of all pandemics. *Emerging Infectious Diseases*, 12(1):15.

Thompson, W. R.
1933. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3/4):285–294.

Thompson, W. W., D. K. Shay, E. Weintraub, L. Brammer, N. Cox, L. J. Anderson, and K. Fukuda
2003. Mortality associated with influenza and respiratory syncytial virus in the united states. *Jama*, 289(2):179–186.

Thorndike, E.
1911. *Animal intelligence: Experimental studies*. Routledge.

Tomba, G. S. and J. Wallinga
2008. A simple explanation for the low impact of border control as a countermeasure to the spread of an infectious disease. *Mathematical biosciences*, 214(1-2):70–72.

Towers, S. and Z. Feng
2012. Social contact patterns and control strategies for influenza in the elderly. *Mathematical biosciences*, 240(2):241–249.

Traag, V. A., L. Waltman, and N. J. van Eck
2019. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9.

Treanor, J.
2004. Influenza vaccine—outmaneuvering antigenic shift and drift. *New England Journal of Medicine*, 350(3):218–220.

Truscott, J. and N. M. Ferguson
2012. Evaluating the adequacy of gravity models as a description of human mobility for epidemic modelling. *PLoS computational biology*, 8(10):e1002699.

Tuite, A. R., A. L. Greer, M. Whelan, A.-L. Winter, B. Lee, P. Yan, J. Wu, S. Moghadas, D. Buckeridge, B. Pourbohloul, et al.
2010. Estimated epidemiologic parameters and morbidity associated with pandemic h1n1 influenza. *Can Med Assoc J*, 182(2):131–136.

Tuyl, F.
2017. A note on priors for the multinomial model. *The American Statistician*, 71(4):298–301.

Ulmer, J. B., U. Valley, and R. Rappuoli
2006. Vaccine manufacturing: challenges and solutions. *Nature biotechnology*, 24(11):1377.

Vamplew, P., R. Dazeley, A. Berry, R. Issabekov, and E. Dekker
2011. Empirical evaluation methods for multiobjective reinforcement learning algorithms. *Machine learning*, 84(1-2):51–80.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin
. Attention is all you need. In *Advances in Neural Information Processing Systems 30*, I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.

Viboud, C., L. Simonsen, R. Fuentes, J. Flores, M. A. Miller, and G. Chowell
2016. Global mortality impact of the 1957–1959 influenza pandemic. *The Journal of infectious diseases*, 213(5):738–745.

Vinyals, O., I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, et al.
2019. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, Pp. 1–5.

Volz, E. M. and I. Siveroni
2018. Bayesian phylodynamic inference with complex models. *PLoS computational biology*, 14(11):e1006546.

Vynnycky, E. and R. White
2010. *An introduction to infectious disease modelling*. OUP oxford.

Wallinga, J., P. Teunis, and M. Kretzschmar
2006. Using data on social contacts to estimate age-specific transmission parameters for respiratory-spread infectious agents. *American journal of epidemiology*, 164(10):936–944.

Wang, L. and J. T. Wu
2018. Characterizing the dynamics underlying global spread of epidemics. *Nature communications*, 9(1):218.

Watkins, C. J. C. H.
1989. *Learning from delayed rewards*. PhD thesis.

Watts, D. J., R. Muhamad, D. C. Medina, and P. S. Dodds
2005. Multiscale, resurgent epidemics in a hierarchical metapopulation model. *Proceedings of the National Academy of Sciences of the United States of America*, 102(32):11157–11162.

Wein, L. M., D. L. Craft, and E. H. Kaplan
2003. Emergency response to an anthrax attack. *Proceedings of the National Academy of Sciences*, 100(7):4346–4351.

WHO
2004. WHO guidelines on the use of vaccines and antivirals during influenza pandemics.

WHO et al.
2017a. Measles vaccines: Who position paper–april 2017–note de synthèse de l'oms sur les vaccins contre la rougeole–avril 2017. *Weekly Epidemiological Record= Relevé épidémiologique hebdomadaire*, 92(17):205–227.

WHO et al.
2017b. *Ten years in public health, 2007–2017: report by Dr Margaret Chan, Director-General, World Health Organization*. World Health Organization.

Willem, L., S. Stijven, E. Vladislavleva, J. Broeckhove, P. Beutels, and N. Hens
2014. Active Learning to Understand Infectious Disease Models and Improve Policy Making. *PLoS Comput Biol*, 10(4):e1003563.

Willem, L., F. Verelst, J. Bilcke, N. Hens, and P. Beutels
2017. Lessons from a decade of individual-based models for infectious disease transmission: a systematic review (2006-2015). *BMC infectious diseases*, 17(1):612.

Williams, R. J.
1992. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine learning*, 8(3-4):229–256.

Wolfson, L. J., P. M. Strebel, M. Gacic-Dobo, E. J. Hoekstra, J. W. McFarland, B. S. Hersh, M. Initiative, et al.
2007. Has the 2005 measles mortality reduction goal been achieved? a natural history modelling study. *The Lancet*, 369(9557):191–200.

Wu, J. T., S. Riley, C. Fraser, and G. M. Leung
2006. Reducing the impact of the next influenza pandemic using household-based public health interventions. *PLoS medicine*, 3(9):e361.

Yaesoubi, R. and T. Cohen
2011. Dynamic health policies for controlling the spread of emerging infections: influenza as an example. *PloS one*, 6(9).

Yaesoubi, R. and T. Cohen
2013. Identifying dynamic tuberculosis case-finding policies for hiv/tb coepidemics. *Proceedings of the National Academy of Sciences*, 110(23):9457–9462.

Yaesoubi, R. and T. Cohen
2016. Identifying cost-effective dynamic policies to control epidemics. *Statistics in medicine*, 35(28):5189–5209.

Yan, X.-Y., W.-X. Wang, Z.-Y. Gao, and Y.-C. Lai
2017. Universal model of individual and population mobility on diverse spatial scales. *Nature communications*, 8(1):1639.

Yanez, A., C. Hayes, and F. Glavin
. Towards the control of epidemic spread: Designing reinforcement learning environments.

Yang, Y., J. D. Sugimoto, M. E. Halloran, N. E. Basta, D. L. Chao, L. Matrajt, G. Potter, E. Kenah, and I. M. Longini
2009. The transmissibility and control of pandemic influenza a (h1n1) virus. *Science*, 326(5953):729–733.

Yu, Y.
2018. Towards sample efficient reinforcement learning. In *IJCAI*, Pp. 5739–5743.

Zhu, N., D. Zhang, W. Wang, X. Li, B. Yang, J. Song, X. Zhao, B. Huang, W. Shi, R. Lu, et al.
2020. A novel coronavirus from patients with pneumonia in china, 2019. *New England Journal of Medicine*.