UNIVERSITY OF CALIFORNIA

Los Angeles

Structured Learning and Decision-Making for Medical Informatics

A dissertation submitted in partial satisfaction
of the requirements for the degree
Doctor of Philosophy in Electrical Engineering

by

Onur Atan

2018

ABSTRACT OF THE DISSERTATION

Structured Learning and
Decision-Making for Medical Informatics

by

Onur Atan
Doctor of Philosophy in Electrical Engineering
University of California, Los Angeles, 2018
Professor Mihaela van der Schaar, Chair

Clinicians are routinely faced with the practical challenge of integrating high-dimensional data in order to make the most appropriate clinical decision from a large set of possible actions for a given patient. Current clinical decisions continue to rely on clinical practice guidelines, which are aimed at a representative patient rather than an individual patient who may display other characteristics. Unfortunately, if it were necessary to learn everything from the limited medical data, the problem would be completely intractable because of the high-dimensional feature space and large number of medical decisions. My thesis aims to design and analyze algorithms that learn and exploit the structure in the medical data – for instance, structures among the features (relevance relations) or decisions (correlations). The proposed algorithms have much in common with the works in online and counterfactual learning literature but unique challenges in the medical informatics lead to numerous key differences from existing state of the art literature in Machine Learning (ML) and require key innovations to deal with large number of features and treatments, heterogeneity of the patients, sequential decision-making, and so on.

The dissertation of Onur Atan is approved.

William Hsu

Arash A. Amini

Stanley Osher

Aydogan Ozcan

Mihaela van der Schaar, Committee Chair

University of California, Los Angeles

2018

To my dear family, Hayri, Hafize, Volkan Atan,

Without you, none of this would've been possible.

TABLE OF CONTENTS

1 Introduction **1**

1.1 Motivation 1

1.2 Contribution of Dissertation 2

1.2.1 Chapter 2: Global Bandits 2

1.2.2 Chapter 3: Constructing Effective Policies from Observational Datasets with Many Features 3

1.2.3 Chapter 4: Counterfactual Policy Optimization Using Domain Adversarial Neural Networks 4

1.2.4 Chapter 5: Online Decision Making with Costly Observations 4

1.2.5 Chapter 6: Sequential Patient Allocation for Randomized Controlled Trials 4

2 Global Bandits **6**

2.1 Introduction 6

2.2 Contribution and Key Results 8

2.3 Related Work 8

2.3.1 Non-informative MAB 8

2.3.2 Group-informative MAB 9

2.3.3 Globally-informative MAB 9

2.4 Problem Formulation 10

2.4.1 Arms, Reward Functions and Informativeness 10

2.4.2 Definition of the Regret 12

2.5 Weighted-Arm Greedy Policy (WAGP) 12

2.6 Regret Analysis of the WAGP 14

2.6.1 Preliminaries for the Regret Analysis 14
2.6.2 Worst-case Regret Bounds for the WAGP 15
2.6.3 Parameter Dependent Regret Bounds for the WAGP 17
2.6.4 Lower Bound on the Worst-case Regret 19
2.7 The Best of the UCB and the WAGP (BUW) 19
2.8 Appendices 21
2.8.1 Preliminaries 21
2.8.2 Proof of Proposition 1 21
2.8.3 Preliminary Results 21
2.8.4 Proof of Theorem 1 23
2.8.5 Proof of Theorem 2 24
2.8.6 Proof of Theorem 3 25
2.8.7 Proof of Theorem 4 25
2.8.8 Proof of Theorem 5 26
2.8.9 Proof of Theorem 6 27
2.8.10 Proof of Theorem 7 28
2.8.11 Auxiliary Lemma 31

3 Constructing Effective Policies from Observational Datasets with Many Features 32

3.1 Introduction 32
3.2 Related Work 34
3.3 Data 37
3.4 The Algorithm 39
3.4.1 True Propensities 39

3.4.2 Relevance . . . 40
3.4.3 Policy Optimization . . . 46
3.4.4 Policy Neural Network (PONN) objective . . . 47
3.4.5 Unknown Propensities . . . 48
3.5 Pseudo-code for the Algorithm PONN-B . . . 49
3.6 Extension: Relevant Feature Selection with Fine Gradations . . . 50
3.7 Numerical Results . . . 53
3.7.1 Dataset . . . 53
3.7.2 Comparisons . . . 53
3.8 Appendix . . . 56

4 Counterfactual Policy Optimization Using Domain-Adversarial Neural Networks . . . 62

4.1 Introduction . . . 62
4.2 Related Work . . . 63
4.3 Problem Setup . . . 65
4.3.1 Observational Data . . . 65
4.3.2 Definition of Policy Outcome . . . 66
4.4 Counterfactual Estimation Bounds . . . 67
4.5 Counterfactual Policy Optimization (CPO) . . . 73
4.5.1 Domain Adversarial Neural Networks . . . 73
4.6 Numerical Results . . . 76
4.6.1 Experimental Setup . . . 76
4.6.2 Benchmarks . . . 77
4.6.3 Results . . . 78

5 Online Decision Making with Costly Observations 82

5.1 Introduction 82

5.2 Related Work 83

5.2.1 MAB Literature 84

5.2.2 MDP literature 84

5.2.3 Budgeted Learning 85

5.3 Contextual Multi-armed Bandits with Costly Observations 85

5.3.1 Problem Formulation 85

5.3.2 Simultaneous Optimistic Observation Selection (Sim-OOS) Algorithm 90

5.3.3 Regret Bounds for the Sim-OOS algorithm 92

5.4 Multi-armed Bandits with Sequential Costly Observations 93

5.4.1 Problem Formulation 93

5.4.2 Sequential Optimistic Observation Selection (Seq-OOS) 96

5.4.3 Regret Bounds of the Seq-OOS 98

5.5 Illustrative Results 99

5.6 Proofs 100

5.7 Appendices 113

5.7.1 Probability of Confidence Intervals Violation for Sim-OOS 113

5.7.2 Probability of Confidence Intervals Violation for Seq-OOS 114

5.7.3 L_1 deviation of true and empirical distributions 115

5.7.4 Summation bound 116

6 Sequential Patient Allocation for Randomized Controlled Trials 117

6.1 Introduction 117

6.2 Related Work 120

6.3 Randomized Clinical Trial (RCT) Design 121
6.3.1 Exponential Families and Jeffrey's Prior 122
6.3.2 MDP Model for RCT Design 123
6.3.3 Dynamic Programming (DP) solution 126
6.4 A greedy solution for fixed recruitment: Optimistic Knowledge Gradient (Opt-KG) 127
6.5 Extension to Treatment Efficacy Identification in RCTs 129
6.6 Experiments 132
6.6.1 Results on Treatment Efficacy Identification 133
6.6.2 Results on ATE estimation 136

7 Concluding Remarks 139

References 140

LIST OF FIGURES

2.1 Illustration of the minimum suboptimality gap and the suboptimality distance. . 15

3.1 Neural network architecture 45

3.2 Effect of the hyperparameter on the accuracy of our algorithm 56

4.1 Neural network model based on [GUA16] 72

4.2 The effect of domain loss in DACPOL performance 79

4.3 The effect of selection bias in DACPOL performance 80

4.4 The effect of irrelevant features in DACPOL vs POEM 81

5.1 Illustration of sequential policy 95

5.2 Performance of Sim-OOS and Seq-OOS 99

6.1 Stages of Clinical Trials 118

6.2 Error Comparisons with Benchmarks 133

6.3 Total error rates for different parameter 134

6.4 Tradeoffs between Type-I and Type-II errors 135

6.5 RMSE improvement of Opt-KG with respect to UA 137

6.6 RMSE performance of Opt-KG with respect to UA and TS 138

6.7 RMSE performance of recruiting M patients in the first step 138

LIST OF TABLES

2.1 Comparison with related works 10

2.2 Frequently used notations in regret analysis 16

3.1 Success rates of two treatments for kidney stones [BPC13] 33

3.2 Performance in the Breast Cancer Experiment 55

4.1 Comparison with the related literature 63

4.2 Loss Comparisons for Breast Cancer Dataset; Means and 95% Confidence Intervals 78

6.1 RCT Examples in the Literature 119

6.2 Error Metric for different patient m 135

6.3 Improvement score for different budgets 135

6.4 Comparison of trial length for a confidence level 136

ACKNOWLEDGMENTS

I am deeply grateful to my advisor Professor Mihaela van der Schaar, without whom this thesis does not exist. I benefited and learned tremendously from her passion and enthusiasm for high-quality research, her unique taste of excellent research topics, and her perpetual energy. I would like to thank Prof. William Zame and Prof. Cem Tekin for their support and help through the course of my PhD. I would also like to thank the other members of my dissertation committee, Prof. Stanley Osher, Prof. Aydogan Ozcan, Prof. Arash Amini and Prof. William Hsu for their time and efforts in evaluating my work.

I would also like to thank my labmates Kartik Ahuja, Basar Akbay, Ahmed Alaa, Jinsung Yoon, Changhee Lee for helping me to think through many of my research ideas, and for helping to peer review my papers before submission.

Finally, I must thank my family, my mom and dad, and my brother Volkan, for their continued love and support for me during the course of my PhD career. I would like to dedicate my thesis to my family.

VITA

2013 B.S. (Electrical and Electronics Engineering), Bilkent University

2014 M.S. (Electrical Engineering), University of California, Los Angeles

2013-present Graduate Student Researcher, Electrical Engineering, UCLA

PUBLICATIONS

1. **Onur Atan**, Cem Tekin, Mihaela van der Schaar, "Global Bandits," To appear in *IEEE Transactions on Neural Networks and Learning Systems.*

2. **Onur Atan**, Cem Tekin, Mihaela van der Schaar, "Global Bandits with Hölder Continuity," In *Artificial Intelligence and Statistics*, pp. 28-36, 2015.

3. **Onur Atan**, James Jordon, Mihaela van der Schaar, "Deep-Treat: Learning Optimal Personalized Treatments From Observational Data Using Neural Networks", In *AAAI Conference on Artificial Intelligence*, pp. 2071-2078, 2018.

4. **Onur Atan**, William Zame, Qiaojun Feng, Mihaela van der Schaar, "Constructing Effective Personalized Policies Using Counterfactual Inference from Biased Data Sets with Many Features", In *arXiv:1612.08082* (under review in Machine Learning).

5. Cem Tekin, **Onur Atan**, Mihaela van der Schaar, "Discover the expert: Context-adaptive expert selection for medical diagnosis", In *IEEE transactions on emerging topics in computing*, pp. 220-234, 2015.

6. **Onur Atan**, William Hsu, Cem Tekin, Mihaela van der Schaar, "A data-driven approach for matching clinical expertise to individual cases," In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference* , pp. 2105-2109, 2015.

7. Yannick Meier, Jie Xu, **Onur Atan**, Mihaela van der Schaar, "Predicting Grades," In *IEEE transactions on signal processing*, pp. 959-972, 2016.

8. Yannick Meier, Jie Xu, **Onur Atan**, Mihaela van der Schaar, "Personalized grade prediction: A data mining approach," In *Data Mining (ICDM), 2015 IEEE International Conference on.* , pp. 907-912, 2015.

9. Sabrina Muller, **Onur Atan**, Mihaela van der Schaar, Anja Kelin, "Context-aware proactive content caching with service differentiation in wireless networks", In *IEEE Transactions on Wireless Communications*, pp. 1024-1036, 2017.

CHAPTER 1

Introduction

1.1 Motivation

Should a cancer patient get chemotherapy or radiotherapy? Does a hospitalized kidney disease patient need a dialysis or hemofiltration? How do these different treatment options impact the wellness of the patient over time? Such treatment planning and management questions have always been in the heart of daily clinical practice; however, the current practice is short of rigorous domain-knowledge guidelines that can accurately prescribe a well-managed plan of treatments for an individual patient over time. Specific examples of such kind of clinical problems include: deciding on a breast cancer treatment plan, deciding on a sequence of treatments over time for adults with primary immune thrombocytopenia, and selecting the right renal replacement therapy for a kidney disease patient whose illness is progressing. A common trait shared by all those problems is the need to estimate the response of an individual patient to a specific intervention over time. Accurate treatment response estimates would indeed allow for more efficient treatment plans that maximize the long-term clinical benefit for individual patients.

With the unprecedentedly vast availability of modern electronic health records (EHRs) in most major hospitals and health-care facilities, constructing data-driven models for treatment response has become more viable than ever before. The main focus of existing literature is on determining the optimal treatment online or estimating the treatment effects from observational EHR data with the outcomes of the treatments that are not undertaken (counterfactuals) missing. However, there are structures among the features (relevance relations) or decisions (correlations) in the EHR data. For example, treatments with similar

chemical composition may have similar effects on the patients, and treatments effects are only dependent on the relevant features and using irrelevant features could hurt the performance of the algorithm significantly. The existing research does not address the structures in medical data sufficiently in order to make (treatment) decisions.

1.2 Contribution of Dissertation

My dissertation develops novel practical algorithms for learning and exploiting the relevance relations among the features and correlations in order to estimate treatment effects and make optimal medical decisions. I answer the following questions:

- Can we develop systematic algorithms in order to estimate personalized treatment policy from a limited EHR data with high-dimensional feature space?
- What happens if there are many choices for the treatment? Do the correlations among the treatments help us learn faster?
- What happens if the clinician has the choice of obtaining more information (by ordering medical tests) by paying a cost (money, inconvenience, etc.) before treatment?
- Can we develop sequential patient recruitment/allocation algorithms to improve the learning power in Randomized Clinical Trials?

1.2.1 Chapter 2: Global Bandits

Multi-armed bandits (MAB) model sequential decision making problems, in which a learner sequentially chooses arms with unknown reward distributions in order to maximize its cumulative reward. Most of the prior work on MAB assumes that the reward distributions of each arm are independent. But in a wide variety of decision problems in medical informatics including drug dosage control, the expected rewards of different arms are correlated, so that selecting one arm provides information about the expected rewards of other arms as well. This chapter (based on [ATS18, ATS15a]) proposes and analyzes a class of models of such

decision problems, which we call *global bandits*. In the case in which rewards of all arms are deterministic functions of a single unknown parameter, this chapter constructs a greedy policy that achieves *bounded regret*, with a bound that depends on the single true parameter of the problem. Hence, this policy selects suboptimal arms only finitely many times with probability one. For this case we also obtain a bound on regret that is *independent of the true parameter*; this bound is sub-linear, with an exponent that depends on the informativeness of the arms. This chapter also proposes a variant of the greedy policy that achieves $\tilde{\mathcal{O}}(\sqrt{T})$ worst-case and $\mathcal{O}(1)$ parameter dependent regret.

1.2.2 Chapter 3: Constructing Effective Policies from Observational Datasets with Many Features

This chapter (based on [AZF16]) proposes a novel approach for constructing effective personalized policies when the observed data lacks counter-factual information, is biased and possesses many features. The approach is applicable in a wide variety of settings from healthcare to advertising to education to finance. These settings have in common that the decision maker can observe, for each previous instance, an array of features of the instance, the action taken in that instance, and the reward realized – but not the rewards of actions that were not taken: the counterfactual information. Learning in such settings is made even more difficult because the observed data is typically biased by the existing policy (that generated the data) and because the array of features that might affect the reward in a particular instance – and hence should be taken into account in deciding on an action in each particular instance – is often vast. The approach presented here estimates propensity scores for the observed data, infers counterfactuals, identifies a (relatively small) number of features that are (most) relevant for each possible action and instance, and prescribes a policy to be followed. Comparison of the proposed algorithm against state-of-art algorithms on actual datasets demonstrates that the proposed algorithm achieves a significant improvement in performance.

1.2.3 Chapter 4: Counterfactual Policy Optimization Using Domain Adversarial Neural Networks

This chapter (based on [**?**, AJS18]) extends the previous chapter on policy optimization from observational studies. It presents theoretical bounds on estimation errors of counterfactuals from observational data by making connections to domain adaptation theory. It also presents a principled way of choosing optimal policies using domain adversarial neural networks. The experiments on semi-synthetic breast cancer shows that our algorithm significantly outperforms the existing methods.

1.2.4 Chapter 5: Online Decision Making with Costly Observations

The decision-maker needs to learn simultaneously what observations to make and what actions to take. This chapter (based on [AS16]) incorporates the information acquisition decision into an online learning framework. We propose two different algorithms for this dual learning problem: Sim-OOS and Seq-OOS where observations are made simultaneously and sequentially, respectively. A regret that is sublinear in time is proved for both cases. The developed framework and algorithms can be used in medical informatics in which collecting information prior to making decisions is costly. Our algorithms are validated in a breast cancer example setting in which we show substantial performance gains for our proposed algorithms.

1.2.5 Chapter 6: Sequential Patient Allocation for Randomized Controlled Trials

Randomized Controlled Trials (RCTs) have became to gold standard comparing the effect and value of treatment(s) with respect to control group. Most of the existing RCTs allocate the patients into treatment groups by simply *repeated fair coin tossing*. However, we show that this approach is not optimal when the standard deviations of treatment groups differ. In this chapter, we formulate this problem as Markov Decision Process (MDP) by assuming a Bayesian prior on the parameters of the treatment outcomes and propose Optimistic

Knowledge Gradient (Opt-KG), an approximate solution to MDP, to allocate the patients into treatment groups so that both estimation error of subgroup-level treatment effects or convex combination of type-I and type-II errors are minimized. This chapter illustrates our algorithm and its effectiveness on a synthetic dataset and verifies that Opt-KG significantly outperforms the repeated fair coin tossing as the deviations of treatment groups differ.

CHAPTER 2

Global Bandits

2.1 Introduction

Multi-armed bandits (MAB) provide powerful models and algorithms for sequential decision-making problems in which the expected reward of each arm (action) is unknown. The goal in MAB problems is to design online learning algorithms that maximize the total reward, which turns out to be equivalent to minimizing the regret, where the regret is defined as the difference between the total expected reward obtained by an oracle that always selects the best arm based on complete knowledge of arm reward distributions, and that of the learner, who does not know the expected arm rewards beforehand. Classical K-armed MAB [LR85] does not impose any dependence between the expected arm rewards. But in a wide variety of examples such as drug dosage, the expected rewards of different arms are correlated, so that selecting one arm provides information about the expected rewards of other arms as well. This chapter proposes and analyzes such a MAB model, which we call *Global Bandits* (GB).

In GB, the expected reward of each arm is a function of a single global parameter. It is assumed that the learner knows these functions but does not know the true value of the parameter. For this problem, this chapter proposes a greedy policy, which constructs an estimate of the global parameter by taking a weighted average of parameter estimates computed separately from the reward observations of each arm. The proposed policy achieves *bounded regret*, where the bound depends on the value of the parameter. This implies that the greedy policy learns the optimal arm, i.e., the arm with the highest expected reward, in finite time. In addition, a worst-case (parameter independent) bound on the regret of the

greedy policy is obtained. This bound is sub-linear in time, and its time exponent depends on the *informativeness of the arms*, which is a measure of the strength of correlation between expected arm rewards.

GBs encompass the model studied in [MRT09], in which it is assumed that the expected reward of each arm is a *linear function* of a single global parameter. This is a special case of the more general model we consider in this paper, in which the expected reward of each arm is a Hölder continuous, possibly non-linear function of a single global parameter. On the technical side, non-linear expected reward functions significantly complicates the learning problem. When the expected reward functions are linear, then the information one can infer about the expected reward of arm X by an additional single sample of the reward from arm Y is independent of the history of previous samples from arm Y.[1] However, if reward functions are non-linear, then the additional information that can be inferred about the expected reward of arm X by a single sample of the reward from arm Y is biased. Therefore, the previous samples from arm X and arm Y needs to be incorporated to ensure that this bias asymptotically converges to 0.

Many applications can be formalized as GBs. An example is clinical trials involving similar drugs (e.g., drugs with a similar chemical composition) or treatments which may have similar effects on the patients.

Example 1: Let y_t be the dosage level of the drug for patient t and x_t be the response of patient t. The relationship between the drug dosage and patient response is modeled in [LR78] as $x_t = M(y_t; \theta_*) + \epsilon_t$, where $M(\cdot)$ is the response function, θ_* is the slope if the function is linear or the elasticity if the function is exponential or logistic, and ϵ_t is i.i.d. zero mean noise. For this model, θ_* becomes the global parameter and the set of drug dosage levels becomes the set of arms.

[1]The additional information about the expected reward of arm X that can be inferred from obtaining sample reward r from arm Y is the same as the additional information about the expected reward of arm X that could be inferred from obtaining the sample reward $L(r)$ from arm X itself, where L is a linear function that depends only on the reward functions themselves.

2.2 Contribution and Key Results

The main contributions in this chapter can be summarized as follows:

- It proposes a non-linear parametric model for MABs, which we refer to as GBs, and a greedy policy, referred to as *Weighted Arm Greedy Policy* (WAGP), which achieves bounded regret.
- It defines the concept of *informativeness*, which measures how well one can estimate the expected reward of an arm by using rewards observed from the other arms, and then, prove a sublinear in time worst-case regret bound for WAGP that depends on the informativeness.
- It also proposes another learning algorithm called the *Best of UCB and WAGP (BUW)*, which fuses the decisions of the UCB1 [ACF02a] and WAGP in order to achieve $\tilde{\mathcal{O}}(\sqrt{T})$[2] worst-case and $\mathcal{O}(1)$ parameter dependent regrets.

2.3 Related Work

There is a wide strand of literature on MABs including finite armed stochastic MAB [LR85, ACF02a, Aue02, GC11], Bayesian MAB [KOG12, Tho33, AG12a, KKE13, BL13], contextual MAB [LZ08, Sli14a, AG13] and distributed MAB [TS15a, XTZ15, LZ13]. Depending on the extent of informativeness of the arms, MABs can be categorized into three: non-informative, group informative and globally informative MABs.

2.3.1 Non-informative MAB

In this chapter, a MAB is called as *non-informative* if the reward observations of any arm do not reveal any information about the rewards of the other arms. Example of non-informative

[2] $\mathcal{O}(\cdot)$ is the Big O notation, $\tilde{\mathcal{O}}(\cdot)$ is the same as $\mathcal{O}(\cdot)$ except it hides terms that have polylogarithmic growth.

MABs include finite armed stochastic [LR85, ACF02a] and non-stochastic [ACF95] MABs. Lower bounds derived for these settings point out to the impossibility of bounded regret.

2.3.2 Group-informative MAB

A MAB is called *group-informative* if the reward observations from an arm provides information about a group of other arms. Examples include linear contextual bandits [LCL10, CLR11], multi-dimensional linear bandits [APS11, RT10, DHK08, APS12, CK11] and combinatorial bandits [CWY13, GKJ12]. In these works, the regret is sublinear in time and in the number of arms. For example, [APS11] assumes a reward structure that is linear in an unknown parameter and shows a regret bound that scales linearly with the dimension of the parameter. It is not possible to achieve bounded regret in any of the above settings since multiple arms are required to be selected at least logarithmically many times in order to learn the unknown parameters.

Another related work [MS11] studies a setting that interpolates between the bandit (partial feedback) and experts (full feedback) settings. In this setting, the decision-maker obtains not only the reward of the selected arm but also an unbiased estimate of the rewards of a subset of the other arms, where this subset is determined by a graph. This is not possible in the proposed setting due to the non-linear reward structure and bandit feedback.

2.3.3 Globally-informative MAB

A MAB is called *globally-informative* if the reward observations from an arm provide information about the rewards of *all* the arms [LM14, MRT09]. GB belongs to the class of globally-informative MAB and includes the linearly-parametrized MAB [MRT09] as a subclass. Hence, the results obtained in this chapter reduces to the results of [MRT09] for the special case when expected arm rewards are linear in the parameter.

Table 2.1 summarizes the GB model and theoretical results, and compares them with the existing literature in the parametric MAB models. Although GB is more general than the model in [MRT09], both WAGP and BUW achieves bounded parameter-dependent regret,

	GB (our work)	[APS11, RT10, DHK08, APS12, CK11]	[MRT09]	[FCG11]
Parameter dimension	Single	Multi	Single	Multi
Reward functions	Non-linear	Linear	Linear	Generalized linear
Worst-case regret	BUW: $\tilde{\mathcal{O}}(\sqrt{T})$, WAGP: $\mathcal{O}(T^{1-\frac{\gamma}{2}})$	$\tilde{\mathcal{O}}(\sqrt{T})$	$\mathcal{O}(\sqrt{T})$	$\tilde{\mathcal{O}}(\sqrt{T})$
Parameter dependent regret	BUW: $\mathcal{O}(1)$, WAGP: $\mathcal{O}(1)$	$\mathcal{O}(\log T)$	$\mathcal{O}(1)$	$\mathcal{O}\left((\log T)^3\right)$

Table 2.1: Comparison with related works

and BUW is able to achieve the same worst-case regret as the policy in [MRT09]. On the other hand, although the linear MAB models are more general than GB, it is not possible to achieve bounded regret in these models.

2.4 Problem Formulation

2.4.1 Arms, Reward Functions and Informativeness

There are K arms indexed by the set $\mathcal{K} := \{1, \ldots, K\}$. The global parameter is denoted by θ_*, which belongs to the parameter set Θ that is taken to be the unit interval for simplicity of exposition. The random variable $X_{k,t}$ denotes the reward of arm k at time t. $X_{k,t}$ is drawn independently from a distribution $\nu_k(\theta_*)$ with support $\mathcal{X}_k \subseteq [0, 1]$. The expected reward of arm k is a Hölder continuous, invertible function of θ_*, which is given by $\mu_k(\theta_*) := \mathrm{E}_{\nu_k(\theta_*)}[X_{k,t}]$, where $\mathrm{E}_\nu[\cdot]$ denotes the expectation taken with respect to distribution ν. This is formalized in the following assumption.

Assumption 1. *(i) For each $k \in \mathcal{K}$ and $\theta, \theta' \in \Theta$ there exists $D_{1,k} > 0$ and $1 < \gamma_{1,k}$, such that*

$$|\mu_k(\theta) - \mu_k(\theta')| \geq D_{1,k}|\theta - \theta'|^{\gamma_{1,k}}.$$

(ii) For each $k \in \mathcal{K}$ and $\theta, \theta' \in \Theta$ there exists $D_{2,k} > 0$ and $0 < \gamma_{2,k} \leq 1$, such that

$$|\mu_k(\theta) - \mu_k(\theta')| \leq D_{2,k}|\theta - \theta'|^{\gamma_{2,k}}.$$

The first assumption ensures that the reward functions are monotonic and the second assumption, which is also known as Hölder continuity, ensures that the reward functions are

smooth. These assumptions imply that the reward functions are invertible and the inverse reward functions are also Hölder continuous. Moreover, they generalize the model proposed in [MRT09], and allow us to model real-world scenarios described in Examples 1 and 2, and propose algorithms that achieve bounded regret.

Some examples of the reward functions that satisfy Assumption 1 are: (i) exponential functions such as $\mu_k(\theta) = a\exp(b\theta)$ where $a > 0$, (ii) linear and piecewise linear functions, and (iii) sub-linear and super-linear functions in θ which are invertible in Θ such as $\mu_k(\theta) = a\theta^\gamma$ where $\gamma > 0$ and $\Theta = [0, 1]$.

Proposition 1. *Define* $\underline{\mu}_k = \min_{\theta\in\Theta} \mu_k(\theta)$ *and* $\overline{\mu}_k = \max_{\theta\in\Theta} \mu_k(\theta)$. *Under Assumption 1, the following are true: (i) For all* $k \in \mathcal{K}$, $\mu_k(\cdot)$ *is invertible. (ii) For all* $k \in \mathcal{K}$ *and for all* $x, x' \in [\underline{\mu}_k, \overline{\mu}_k]$,

$$|\mu_k^{-1}(x) - \mu_k^{-1}(x')| \leq \bar{D}_{1,k}|x - x'|^{\bar{\gamma}_{1,k}}$$

where $\bar{\gamma}_{1,k} = \frac{1}{\gamma_{1,k}}$ *and* $\bar{D}_{1,k} = \left(\frac{1}{D_{1,k}}\right)^{\frac{1}{\gamma_{1,k}}}$.

Invertibility of the reward functions allows us to use the rewards obtained from an arm to estimate the expected rewards of other arms. Let $\bar{\gamma}_1$ and γ_2 be the minimum exponents and $\bar{D}_1$, D_2 be the maximum constants, that is

$$\begin{aligned} \bar{\gamma}_1 &= \min_{k\in\mathcal{K}} \bar{\gamma}_{1,k},\ \gamma_2 = \min_{k\in\mathcal{K}} \gamma_{2,k}, \\ \bar{D}_1 &= \max_{k\in\mathcal{K}} \bar{D}_{1,k},\ D_2 = \max_{k\in\mathcal{K}} D_{2,k}. \end{aligned}$$

Definition 1. *The informativeness of arm* k *is defined as* $\gamma_k := \bar{\gamma}_{1,k}\gamma_{2,k}$. *The informativeness of the GB instance is defined as* $\gamma := \bar{\gamma}_1\gamma_2$.

The informativeness of arm k measures the extent of information that can be obtained about the expected rewards of other arms from the rewards observed from arm k. As shown in later sections, when the informativeness is high, one can form better estimates of the expected rewards of other arms by using the rewards observed from arm k.

2.4.2 Definition of the Regret

The learner knows $\mu_k(\cdot)$ for all $k \in \mathcal{K}$ but does not know θ_*. At each time t it selects one of the arms, denoted by I_t, and receives the random reward $X_{I_t,t}$. The learner's goal is to maximize its cumulative reward up to any time T.

Let $\mu^*(\theta) := \max_{k\in\mathcal{K}} \mu_k(\theta)$ be the maximum expected reward and $\mathcal{K}^*(\theta) := \{k \in \mathcal{K} : \mu_k(\theta) = \mu^*(\theta)\}$ be the optimal set of arms for parameter θ. In addition, let $k^*(\theta)$ denote an arm that is optimal for parameter θ. The policy that selects one of the arms in $\mathcal{K}^*(\theta_*)$ is referred to as the *oracle* policy. The learner incurs a regret (loss) at each time it deviates from the oracle policy. Define the one-step regret at time t as the difference between the expected reward of the oracle policy and the learner, which is given by $r_t(\theta_*) := \mu^*(\theta_*) - \mu_{I_t}(\theta_*)$. Based on this, the cumulative regret of the learner by time T (also referred to as the regret hereafter) is defined as

$$\mathrm{Reg}(\theta_*, T) := \mathbb{E}\left[\sum_{t=1}^{T} r_t(\theta_*)\right].$$

Maximizing the reward is equivalent to minimizing the regret. In the seminal work by Lai and Robbins [LR78], it is shown that the regret becomes infinite as T grows for the classical K-armed bandit problem. On the other hand, $\lim_{T\to\infty} \mathrm{Reg}(\theta_*, T) < \infty$ will imply that the learner deviates from the oracle policy only finitely many times.

2.5 Weighted-Arm Greedy Policy (WAGP)

A greedy policy called the Weighted-Arm Greedy Policy (WAGP) is proposed in this section. The pseudocode of WAGP is given in Algorithm 1. The WAGP consists of two phases: arm selection phase and parameter update phase.

Let $N_k(t)$ denote the number of times arm k is selected until time t, and $\hat{X}_{k,t}$ denote the reward estimate, $\hat{\theta}_{k,t}$ denote the global parameter estimate and $w_k(t)$ denote the weight of arm k at time t. Initially, all the counters and estimates are set to zero. In the arm selection phase at time $t > 1$, the WAGP selects the arm with the highest estimated expected reward: $I_t \in \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\theta}_{t-1})$ where $\hat{\theta}_{t-1}$ is the estimate of the global parameter calculated at the

Algorithm 1 The WAGP

Inputs: $\mu_k(\cdot)$ for each arm k
Initialization: $w_k(0) = 0, \hat{\theta}_{k,0} = 0, \hat{X}_{k,0} = 0, N_k(0) = 0$ for all $k \in \mathcal{K}$, $t = 1$
while $t > 0$ **do**
  if $t = 1$ **then**
    Select arm I_1 uniformly at random from $\mathcal{K}$
  else
    Select arm $I_t \in \arg\max_{k \in \mathcal{K}} \mu_k(\hat{\theta}_{t-1})$ (break ties randomly)
  end if
  $\hat{X}_{k,t} = \hat{X}_{k,t-1}$ for all $k \in \mathcal{K} \setminus I_t$
  $\hat{X}_{I_t,t} = \frac{N_{I_t}(t-1)\hat{X}_{I_t,t-1} + X_{I_t,t}}{N_{I_t}(t-1)+1}$
  $\hat{\theta}_{k,t} = \arg\min_{\theta \in \Theta} |\mu_k(\theta) - \hat{X}_{k,t}|$ for all $k \in \mathcal{K}$
  $N_{I_t}(t) = N_{I_t}(t-1) + 1$
  $N_k(t) = N_k(t-1)$ for all $k \in \mathcal{K} \setminus I_t$
  $w_k(t) = N_k(t)/t$ for all $k \in \mathcal{K}$
  $\hat{\theta}_t = \sum_{k=1}^{K} w_k(t)\hat{\theta}_{k,t}$
end while

end of time $t-1$.[3,4]

In the parameter update phase the WAGP updates: (i) the estimated reward of selected arm I_t, denoted by $\hat{X}_{I_t,t}$, (ii) the global parameter estimate of the selected arm I_t, denoted by $\hat{\theta}_{I_t,t}$, (iii) the global parameter estimate $\hat{\theta}_t$, and (iv) the counters $N_k(t)$. The reward of estimate of arm I_t is updated as:

$$\hat{X}_{I_t,t} = \frac{N_{I_t}(t-1)\hat{X}_{I_t,t-1} + X_{I_t,t}}{N_{I_t}(t-1)+1}.$$

The reward estimates of the other arms are not updated. The WAGP constructs estimates of the global parameter from the rewards of all the arms and combines their estimates using a weighted sum. The WAGP updates $\hat{\theta}_{I_t,t}$ of arm I_t in a way that minimizes the distance between $\hat{X}_{I_t,t}$ and $\mu_{I_t}(\theta)$, i.e., $\hat{\theta}_{I_t,t} = \arg\min_{\theta \in \Theta} |\mu_{I_t}(\theta) - \hat{X}_{I_t,t}|$. Then, the WAGP sets the global parameter estimate as $\hat{\theta}_t = \sum_{k=1}^{K} w_k(t)\hat{\theta}_{k,t}$ where $w_k(t) = N_k(t)/t$. Hence, the WAGP gives more weights to the arms with more reward observations since the confidence on their estimates are higher.

[3]The ties are broken randomly.

[4]For $t = 1$, the WAGP selects a random arm since there is no prior reward observation that can be used to estimate θ_*.

2.6 Regret Analysis of the WAGP

2.6.1 Preliminaries for the Regret Analysis

In this subsection, the tools that will be used in deriving the regret bounds for the WAGP will be defined. Consider any arm $k \in \mathcal{K}$. Its *optimality region* is defined as

$$\Theta_k := \{\theta \in \Theta : k \in \mathcal{K}^*(\theta)\}.$$

Note that Θ_k can be written as union of intervals in each of which arm k is optimal. Each such interval is called an *optimality interval*. Clearly, the following holds: $\bigcup_{k\in\mathcal{K}} \Theta_k = \Theta$. If $\Theta_k = \emptyset$ for an arm k, this implies that there exists no global parameter value for which arm k is optimal. Since there exists an arm k' such that $\mu_{k'}(\theta) > \mu_k(\theta)$ for any $\theta \in \Theta$ for an arm with $\Theta_k = \emptyset$, the greedy policy will discard arm k after $t = 1$. Therefore, without loss of generality, it is assumed that $\Theta_k \neq \emptyset$ for all $k \in \mathcal{K}$. The *suboptimality gap* of arm $k \in \mathcal{K}$ given global parameter $\theta_* \in \Theta$ is defined as $\delta_k(\theta_*) := \mu^*(\theta_*) - \mu_k(\theta_*)$. The *minimum suboptimality gap* given global parameter $\theta_* \in \Theta$ is defined as $\delta_{\min}(\theta_*) := \min_{k\in\mathcal{K}\setminus\mathcal{K}^*(\theta_*)} \delta_k(\theta_*)$.

Let $\Theta^{\text{sub}}(\theta_*)$ be the suboptimality region of the global parameter θ_*, which is defined as the subset of the parameter space in which none of the arms in $\mathcal{K}^*(\theta_*)$ is optimal, i.e.,

$$\Theta^{\text{sub}}(\theta_*) := \Theta \setminus \bigcup_{k'\in\mathcal{K}^*(\theta_*)} \Theta_{k'}.$$

As shown later, the global parameter estimate will converge to θ_*. However, if θ_* lies close to $\Theta^{\text{sub}}(\theta_*)$, the global parameter estimate may fall into the suboptimality region for a large number of times, thereby resulting in a large regret. In order to bound the expected number of times this happens, we define the *suboptimality distance* as the smallest distance between the global parameter and the suboptimality region.

Definition 2. *For a given global parameter θ_*, the suboptimality distance is defined as*

$$\Delta_{min}(\theta_*) := \begin{cases} \inf_{\theta'\in\Theta^{sub}(\theta_*)} |\theta_* - \theta'| & \text{if } \Theta^{sub}(\theta_*) \neq \emptyset \\ 1 & \text{if } \Theta^{sub}(\theta_*) = \emptyset \end{cases}$$

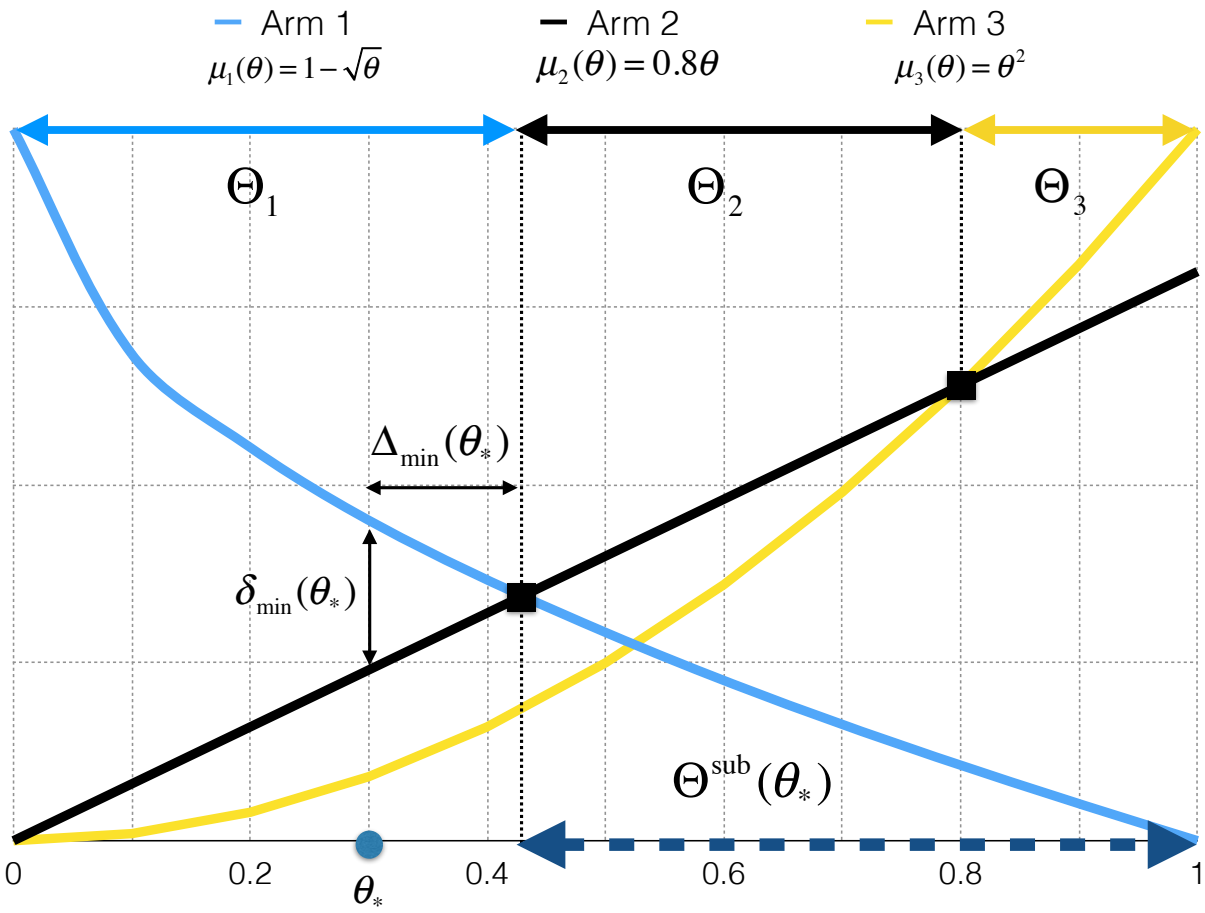


Figure 2.1: Illustration of the minimum suboptimality gap and the suboptimality distance.

From the definition of the suboptimality distance it is evident that the proposed policy always selects an optimal arm in $\mathcal{K}^*(\theta_*)$ when $\hat{\theta}_t$ is within $\Delta_{\min}(\theta_*)$ of θ_*. For notational brevity, we also use $\Delta_* := \Delta_{\min}(\theta_*)$ and $\delta_* := \delta_{\min}(\theta_*)$. An illustration of the suboptimality gap and the suboptimality distance is given in Fig. 2.1 for the case with 3 arms and reward functions $\mu_1(\theta) = 1 - \sqrt{\theta}$, $\mu_2(\theta) = 0.8\theta$ and $\mu_3(\theta) = \theta^2$, $\theta \in [0, 1]$.

The notations frequently used in the regret analysis is highlighted in Table 2.2.

2.6.2 Worst-case Regret Bounds for the WAGP

First, we show that parameter estimate of the WAGP converges in the mean-squared sense.

$\mathcal{K}^*(\theta_*)$	set of optimal arms for θ_*
$\mu^*(\theta_*)$	expected reward of optimal arms
I_t	selected arm at time t
$\hat{\theta}_t$	global parameter estimate at time t
$\delta_* = \delta_{\min}(\theta_*)$	minimum suboptimality gap
$\Delta_* = \Delta_{\min}(\theta_*)$	minimum suboptimality distance
Θ_k	optimality region of arm k
$\Theta^{\text{sub}}(\theta_*)$	suboptimality region of θ_*
γ	informativeness of the arms

Table 2.2: Frequently used notations in regret analysis

Theorem 1. *Under Assumption 1, the global parameter estimate converges to true value of global parameter in mean-squared sense, i.e.,* $\lim_{t\to\infty} \mathbb{E}\left[|\hat{\theta}_t - \theta_*|^2\right] = 0$.

The following theorem bounds the expected one-step regret of the WAGP.

Theorem 2. *Under Assumption 1, we have for WAGP* $\mathbb{E}\left[r_t(\theta_*)\right] \leq \mathcal{O}(t^{-\frac{\gamma}{2}})$.

Theorem 2 proves that the expected one-step regret of the WAGP converges to zero.[5] This is a *worst-case* bound in the sense that it holds for any θ_*. Using this result, the following worst-case regret bound for the WAGP is derived.

Theorem 3. *Under Assumption 1, the worst-case regret of WAGP is*

$$\sup_{\theta_* \in \Theta} Reg(\theta_*, T) \leq \mathcal{O}(K^{\frac{\gamma}{2}} T^{1-\frac{\gamma}{2}}).$$

Note that the worst-case regret bound is sublinear both in the time horizon T and the number of arms K. Moreover, it depends on the informativeness γ. When the reward functions are linear or piecewise linear, we have $\gamma = 1$, which is an extreme case of our model; hence, the worst-case regret is $\mathcal{O}(\sqrt{T})$, which matches with (i) the worst-case regret bound of the standard MAB algorithms in which a linear estimator is used [BC12], and (ii) the bounds obtained for the linearly parametrized bandits [MRT09].

[5]The asymptotic notation is only used for a succinct representation, to hide the constants and highlight the time dependence. This bound holds not just asymptotically but for any finite t.

2.6.3 Parameter Dependent Regret Bounds for the WAGP

This section shows a bound on the parameter dependent regret of the WAGP. First, several constants that will appear in the regret bound are introduced.

Definition 3. *$C_1(\Delta_*)$ is the smallest integer τ such that $\tau \geq \left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \frac{\log(\tau)}{2}$ and $C_2(\Delta_*)$ is the smallest integer τ such that $\tau \geq \left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \log(\tau)$.*

Closed form expressions for these constants can be obtained in terms of the *glog* function [Kal01], for which the following equivalence holds: $y = \text{glog}(x)$ if and only if $x = \frac{\exp(y)}{y}$. Then, we have

$$C_1(\Delta_*) = \left\lceil \frac{1}{2}\left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \text{glog}\left(\frac{1}{2}\left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}}\right) \right\rceil,$$

$$C_2(\Delta_*) = \left\lceil \left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}} \text{glog}\left(\left(\frac{\bar{D}_1 K}{\Delta_*}\right)^{\frac{2}{\bar{\gamma}_1}}\right) \right\rceil, .$$

Next, we define the expected regret incurred between time steps T_1 and T_2 given θ_* as $R_{\theta_*}(T_1, T_2) := \sum_{t=T_1}^{T_2} \mathbb{E}\left[r_t(\theta_*)\right]$. The following theorem bounds the parameter dependent regret of the WAGP.

Theorem 4. *Under Assumption 1, the regret of the WAGP is bounded as follows:*
(i) For $1 \leq T < C_1(\Delta_*)$, the regret grows sublinearly in time, i.e.,

$$R_{\theta_*}(1, T) \leq S_1 + S_2 T^{1-\frac{\gamma}{2}}$$

where S_1 and S_2 are constants that are independent of the global parameter θ_*, whose exact forms are given in Appendix 2.8.6.

(ii) For $C_1(\Delta_*) \leq T < C_2(\Delta_*)$, the regret grows logarithmically in time, i.e.,

$$R_{\theta_*}(C_1(\Delta_*), T) \leq 1 + 2K \log\left(\frac{T}{C_1(\Delta_*)}\right).$$

(iii) For $T \geq C_2(\Delta_*)$, the growth of the regret is bounded, i.e.,

$$R_{\theta_*}(C_2(\Delta_*), T) \leq K\frac{\pi^2}{3}.$$

Thus, we have $\lim_{T\to\infty} Reg(\theta_*, T) < \infty$, i.e., $Reg(\theta_*, T) = \mathcal{O}(1)$.

Theorem 4 shows that the regret is inversely proportional to the suboptimality distance Δ_*, which depends on θ_*. The regret bound contains three regimes of growth: Initially the regret grows sublinearly until time threshold $C_1(\Delta_*)$. After this, it grows logarithmically until time threshold $C_2(\Delta_*)$. Finally, the growth of the regret is bounded after time threshold $C_2(\Delta_*)$. In addition, since $\lim_{\Delta_*\to 0} C_1(\Delta_*) = \infty$, in the worst-case, the bound given in Theorem 4 reduces to the one given in Theorem 3. It is also possible to calculate a Bayesian risk bound for the WAGP by assuming a prior over the global parameter space. This risk bound is given to be $\mathcal{O}(\log T)$ when $\gamma = 1$ and $\mathcal{O}(T^{1-\gamma})$ when $\gamma < 1$ (see [ATS15b]).

Theorem 5. *The sequence of arms selected by the WAGP converges to the optimal arm almost surely, i.e.,* $\lim_{t\to\infty} I_t \in \mathcal{K}^*(\theta_*)$ *with probability 1.*

Theorem 5 implies that a suboptimal arm is selected by the WAGP only finitely many times. This is the major difference between GB and the classical MAB [LR85,ACF02a,RV15], in which every arm needs to be selected infinitely many times asymptotically by any *good* learning algorithm.

Remark 1. *Assumption 1 ensures that the parameter dependent regret is bounded. When this assumption is relaxed, bounded regret may not be achieved, and the best possible regret becomes logarithmic in time. For instance, consider the case when the reward functions are constant over the global parameter space, i.e.,* $\mu_k(\theta_*) = m_k$ *for all* $\theta_* \in [0, 1]$ *where* m_k *is a constant. This makes the reward functions non-invertible. In this case, the learner cannot use the rewards obtained from the other arms when estimating the rewards of arm* k*. Thus, it needs to learn* m_k *of each arm separately, which results in logarithmic in time regret when a policy like UCB1 [ACF02a] is used. This issue still exists even when there are only finitely many possible solutions to* $\mu_k(\theta_*) = x$ *for some* x*, in which case some of the arms should be selected at least logarithmically many times to rule out the incorrect global parameters.*

2.6.4 Lower Bound on the Worst-case Regret

Theorem 3 shows that the worst-case regret of the WAGP is $\mathcal{O}(T^{1-\frac{\gamma}{2}})$, which implies that the regret decreases with γ.

Theorem 6. *For $T \geq 8$ and any policy, the parameter dependent regret is lower bounded by $\Omega(1)$ and the worst-case regret is lower bounded by $\Omega(\sqrt{T})$.*

The theorem above raises a natural question: Can we achieve both $\tilde{\mathcal{O}}(\sqrt{T})$ worst-case regret (like the UCB based MAB algorithms [ACF02a]) and bounded parameter dependent regret by using a combination of UCB and WAGP policies? This question in the affirmative in the next section is answered in affirmative way.

2.7 The Best of the UCB and the WAGP (BUW)

In the this section, the Best of the UCB and the WAGP (BUW), which combines UCB1 and the WAGP to achieve bounded parameter dependent and $\mathcal{O}(\sqrt{T})$ worst-case regrets is proposed. In the worst-case, the WAGP achieves $\mathcal{O}(T^{1-\frac{\gamma}{2}})$ regret, which is weaker than $\tilde{\mathcal{O}}(\sqrt{T})$ worst-case regret of UCB1. On the other hand, the WAGP achieves bounded parameter dependent regret whereas UCB1 achieves a logarithmic parameter dependent regret. In this section, an algorithm which combines these two algorithms and achieves both $\tilde{\mathcal{O}}(\sqrt{T})$ worst-case regret and bounded parameter dependent regret is proposed.

The main idea for such an algorithm follows from Theorem 4. Recall that Theorem 4 shows that the WAGP achieves $O(T^{1-\frac{\gamma}{2}})$ regret when $1 < T < C_1(\Delta_*)$. If the BUW could follow the recommendations of UCB1 when $T < C_1(\Delta_*)$ and the recommendations of the WAGP when $T \geq C_1(\Delta_*)$, then it will achieve a worst-case regret bound of $\tilde{\mathcal{O}}(\sqrt{T})$ and bounded parameter-dependent regret. The problem with this approach is that the suboptimality distance Δ_* is unknown a priori. This problem can be solved by using a data-dependent estimate $\tilde{\Delta}_t$ where $\Delta_* > \tilde{\Delta}_t$ holds with high probability. The data-dependent estimate $\tilde{\Delta}_t$ is given as

$$\tilde{\Delta}_t = \hat{\Delta}_t - \bar{D}_1 K \left(\frac{\log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}}$$

where

$$\hat{\Delta}_t = \Delta_{\min}(\hat{\theta}_t) = \begin{cases} \inf_{\theta' \in \Theta^{\text{sub}}(\hat{\theta}_t)} |\hat{\theta}_t - \theta'| & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) \neq \emptyset \\ 1 & \text{if } \Theta^{\text{sub}}(\hat{\theta}_t) = \emptyset \end{cases}$$

Algorithm 2 The BUW

Inputs: T, $\mu_k(\cdot)$ for each arm k.
Initialization: Select each arm once for $t = 1, 2, \ldots, K$, compute $\hat{\theta}_{k,K}$, $N_k(K)$, $\hat{\mu}_k$, $\hat{X}_{k,K}$ for all $k \in \mathcal{K}$, and $\hat{\theta}_K$, $\hat{\Delta}_K$, $\tilde{\Delta}_K$, $t = K+1$
while $t \geq K+1$ **do**
  if $t < C_2 \left(\max \left(0, \tilde{\Delta}_{t-1} \right) \right)$ **then**
    $I_t \in \arg\max_{k \in \mathcal{K}} \hat{X}_{k,t-1} + \sqrt{\frac{2\log(t-1)}{N_k(t-1)}}$
  else
    $I_t \in \arg\max_{k \in \mathcal{K}} \mu_k(\hat{\theta}_{t-1})$
  end if
  Update $\hat{X}_{I_t,t}$, $N_k(t)$, $w_k(t)$, $\hat{\theta}_{k,t}$, $\hat{\theta}_t$ as in the WAGP
  Solve

$$\hat{\Delta}_t = \begin{cases} \inf_{\theta' \in \Theta^{\mathrm{sub}}(\hat{\theta}_t)} |\hat{\theta}_t - \theta'| & \text{if } \Theta^{\mathrm{sub}}(\hat{\theta}_t) \neq \emptyset \\ 1 & \text{if } \Theta^{\mathrm{sub}}(\hat{\theta}_t) = \emptyset \end{cases}$$

  $\tilde{\Delta}_t = \hat{\Delta}_t - \bar{D}_1 K \left(\frac{\log t}{t} \right)^{\frac{\bar{\gamma}_1}{2}}$
end while

The pseudo-code for the BUW is given in Fig. 2. The regret bounds for the BUW are given in the following theorem.

Theorem 7. *Under Assumption 1, the worst-case regret of the BUW is bounded as follows:*

$$\sup_{\theta_* \in \Theta} Reg(\theta_*, T) \leq \tilde{\mathcal{O}}(\sqrt{KT}).$$

Under Assumption 1, the parameter dependent regret of the BUW is bounded as follows:

(i) For $1 \leq T < C_2(\Delta_*/3)$*, the regret grows logarithmically in time, i.e.,*

$$R_{\theta_*}(1, T) \leq \left[8 \sum_{k:\mu_k < \mu^*} \frac{\log T}{\delta_k} \right] + K \left(1 + \pi^2 \right).$$

(ii) For $T \geq C_2(\Delta_*/3)$*, the growth of the regret is bounded, i.e.,*

$$R_{\theta_*}(C_2(\Delta_*/3), T) \leq K\pi^2.$$

The BUW achieves the lower bound given in Theorem 7, that is $\mathcal{O}(1)$ parameter-dependent regret and $\tilde{\mathcal{O}}(\sqrt{T})$ worst-case regret.

2.8 Appendices

2.8.1 Preliminaries

In all the proofs given below. Let $\boldsymbol{w}(t) = (w_1(t), \ldots, w_K(t))$ be the vector of weights and $\boldsymbol{N}(t) = (N_1(t), \ldots, N_k(t))$ be the vector of counters at time t. Then, $\boldsymbol{w}(t) = \boldsymbol{N}(t)/t$. Since $\boldsymbol{N}(t)$ depends on the history, they are both random variables that depend on the sequence of obtained rewards.

2.8.2 Proof of Proposition 1

(i) Let k and $\theta \neq \theta'$ be arbitrary. Then, by Assumption 1,

$$|\mu_k(\theta) - \mu_k(\theta')| \geq D_{1,k}|\theta - \theta'|^{\gamma_{1,k}} > 0$$

and hence $\mu_k(\theta) \neq \mu_k(\theta')$.

(ii) Suppose $x = \mu_k(\theta)$ and $x' = \mu_k(\theta')$ for some arbitrary θ and θ'. Then, by Assumption 1,

$$|x - x'| \geq D_{1,k}|\mu_k^{-1}(x) - \mu_k^{-1}(x')|^{\gamma_{1,k}}.$$

2.8.3 Preliminary Results

Lemma 1. *For the WAGP the following relation between $\hat{\theta}_t$ and θ_* holds with probability one: $|\hat{\theta}_t - \theta_*| \leq \sum_{k=1}^{K} w_k(t)\bar{D}_1|\hat{X}_{k,t} - \mu_k(\theta_*)|^{\bar{\gamma}_1}$.*

Proof. Before deriving a bound of gap between the global parameter estimate and true global parameter at time t, let $\tilde{\mu}_k^{-1}(x) = \arg\min_{\theta\in\Theta} |\mu_k(\theta) - x|$. By monotonicity of $\mu_k(\cdot)$ and Proposition 1, the following holds $|\tilde{\mu}_k^{-1}(x) - \tilde{\mu}_k^{-1}(x')| \leq \bar{D}_1|x - x'|^{\bar{\gamma}_1}$. Then,

$$\begin{aligned}
|\theta_* - \hat{\theta}_t| &= |\sum_{k=1}^{K} w_k(t)\hat{\theta}_{k,t} - \theta_*| = \sum_{k=1}^{K} w_k(t)\left|\theta_* - \hat{\theta}_{k,t}\right| \\
&\leq \sum_{k=1}^{K} w_k(t)|\tilde{\mu}_k^{-1}(\hat{X}_{k,t}) - \tilde{\mu}_k^{-1}(\tilde{\mu}_k(\theta_*))| \\
&\leq \sum_{k=1}^{K} w_k(t)\bar{D}_1|\hat{X}_{k,t} - \mu_k(\theta_*)|^{\bar{\gamma}_1},
\end{aligned}$$

where there are two cases for the first inequality to check. The first case is $\hat{X}_{k,t} \in \mathcal{X}_k$ where the statement immediately follows. The second case is $\hat{X}_{k,t} \notin \mathcal{X}_k$, where the global parameter estimator $\hat{\theta}_{k,t}$ is either 0 or 1. □

Lemma 2. *The one-step regret of the WAGP is bounded by* $r_t(\theta_*) = \mu^*(\theta_*) - \mu_{I_t}(\theta_*) \leq 2D_2|\theta_* - \hat{\theta}_{t-1}|^{\gamma_2}$ *with probability one, for* $t \geq 2$.

Proof. Note that $I_t \in \arg\max_{k\in\mathcal{K}} \mu_k(\hat{\theta}_{t-1})$. Therefore, we have

$$\mu_{I_t}(\hat{\theta}_{t-1}) - \mu_{k^*(\theta_*)}(\hat{\theta}_{t-1}) \geq 0. \tag{2.1}$$

Since $\mu^*(\theta_*) = \mu_{k^*(\theta_*)}(\theta_*)$, we have

$$\begin{aligned}
&\mu^*(\theta_*) - \mu_{I_t}(\theta_*) \\
&= \mu_{k^*(\theta_*)}(\theta_*) - \mu_{I_t}(\theta_*) \\
&\leq \mu_{k^*(\theta_*)}(\theta_*) - \mu_{I_t}(\theta_*) + \mu_{I_t}(\hat{\theta}_{t-1}) - \mu_{k^*(\theta_*)}(\hat{\theta}_{t-1}) \\
&= \mu_{k^*(\theta_*)}(\theta_*) - \mu_{k^*(\theta_*)}(\hat{\theta}_{t-1}) + \mu_{I_t}(\hat{\theta}_{t-1}) - \mu_{I_t}(\theta_*) \\
&\leq 2D_2|\theta_* - \hat{\theta}_{t-1}|^{\gamma_2},
\end{aligned}$$

where the first inequality follows from (2.1) and the second inequality follows from Assumption 1. □

Let $\mathcal{G}_{\theta_*,\hat{\theta}_t}(x) := \{|\theta_* - \hat{\theta}_t| > x\}$ be the event that the distance between the global parameter estimate and its true value exceeds x. Similarly, let $\mathcal{F}^k_{\theta_*,\hat{\theta}_t}(x) := \{|\hat{X}_{k,t} - \mu_k(\theta_*)| > x\}$ be the event that the distance between the sample mean reward estimate of arm k and the true expected reward of arm k exceeds x.

Lemma 3. *For WAGP we have*

$$\mathcal{G}_{\theta_*,\hat{\theta}_t}(x) \subseteq \bigcup_{k=1}^{K} \mathcal{F}^k_{\theta_*,\hat{\theta}_t}\left(\left(\frac{x}{\bar{D}_1 w_k(t) K}\right)^{\frac{1}{\bar{\gamma}_1}}\right)$$

with probability one, for $t \geq 2$.

Proof. Observe that

$$\begin{aligned}
&\{|\theta_* - \hat{\theta}_t| \leq x\} \\
&\supseteq \left\{ \sum_{k=1}^{K} w_k(t) \bar{D}_1 |\hat{X}_{k,t} - \mu_k(\theta_*)|^{\bar{\gamma}_1} \leq x \right\} \\
&\supseteq \bigcap_{k=1}^{K} \left\{ |\hat{X}_{k,t} - \mu_k(\theta_*)| \leq \left(\frac{x}{w_k(t) \bar{D}_1 K} \right)^{1/\bar{\gamma}_1} \right\},
\end{aligned}$$

where the first inequality follows from Lemma 1. Then,

$$\begin{aligned}
&\{|\theta_* - \hat{\theta}_t| > x\} \subseteq \\
&\qquad \bigcup_{k=1}^{K} \left\{ |\hat{X}_{k,t} - \mu_k(\theta_*)| > \left(\frac{x}{w_k(t) \bar{D}_1 K} \right)^{1/\bar{\gamma}_1} \right\}.
\end{aligned}$$

□

2.8.4 Proof of Theorem 1

Using Lemma 1, the mean-squared error can be bounded as

$$\begin{aligned}
&\mathbb{E}\left[|\theta_* - \hat{\theta}_t|^2 \right] \\
&\leq \mathbb{E}\left[\left(\sum_{k=1}^{K} \bar{D}_1 w_k(t) |\hat{X}_{k,t} - \mu_k(\theta_*)|^{\bar{\gamma}_1} \right)^2 \right] \\
&\leq K \bar{D}_1^2 \sum_{k=1}^{K} \mathbb{E}\left[w_k^2(t) |\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1} \right],
\end{aligned} \tag{2.2}$$

where the inequality follows from the fact that $\left(\sum_{k=1}^{K} a_k \right)^2 \leq K \sum_{k=1}^{K} a_k^2$ for any $a_k > 0$. Then,

$$\begin{aligned}
&\mathbb{E}\left[|\theta_* - \hat{\theta}_t|^2 \right] \\
&\leq K \bar{D}_1^2 \mathbb{E}\left[\sum_{k=1}^{K} w_k^2(t) \mathbb{E}\left[|\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1} | \boldsymbol{w}(t) \right] \right] \\
&\leq K \bar{D}_1^2 \mathbb{E}\left[\sum_{k=1}^{K} w_k^2(t) \int_{x=0}^{\infty} \Pr(|\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1} \geq x | \boldsymbol{w}(t)) dx \right],
\end{aligned} \tag{2.3}$$

where the second inequality follows from the fundamental theorem of expectation. Then, the bound on inner expectation is

$$\begin{aligned}
&\int_{x=0}^{\infty} \Pr(|\hat{X}_{k,t} - \mu_k(\theta_*)|^{2\bar{\gamma}_1} \geq x|\boldsymbol{w}(t))dx \\
&\leq \int_{x=0}^{\infty} 2\exp(-x^{\frac{1}{\bar{\gamma}_1}} N_k(t))\, dx. \\
&= 2\bar{\gamma}_1 \Gamma(\bar{\gamma}_1) N_k(t)^{-\bar{\gamma}_1},
\end{aligned}$$

where $\Gamma(\cdot)$ is Gamma function. Then,

$$\begin{aligned}
\mathbb{E}[|\theta_* - \hat{\theta}_t|^2] &\leq 2K\bar{\gamma}_1 \bar{D}_1^2 \Gamma(\bar{\gamma}_1) \mathbb{E}\left[\sum_{k=1}^{K} \frac{N_k(t)^{2-\bar{\gamma}_1}}{t^2}\right] \\
&\leq 2K\bar{\gamma}_1 \bar{D}_1^2 \Gamma(\bar{\gamma}_1) t^{-\bar{\gamma}_1},
\end{aligned}$$

where the last inequality follows from the fact that $\mathbb{E}[\sum_{k=1}^{K} N_k^{2-\bar{\gamma}_1}(t)/t^2] \leq t^{-\bar{\gamma}_1}$ for any $N_k(t)$ since $\sum_{k=1}^{K} N_k(t) = t$ and $\bar{\gamma}_1 \leq 1$.

2.8.5 Proof of Theorem 2

By Lemma 2 and Jensen's inequality,

$$\mathbb{E}[r_{t+1}(\theta_*)] \leq 2D_2 \mathrm{E}\left[|\theta_* - \hat{\theta}_t|\right]^{\gamma_2}. \tag{2.4}$$

Also by Lemma 1 and Jensen's inequality,

$$\begin{aligned}
&\mathbb{E}\left[|\theta_* - \hat{\theta}_t|\right] \\
&\leq \bar{D}_1 \mathbb{E}\left[\sum_{k=1}^{K} w_k(t) \mathbb{E}\left[|\hat{X}_{k,t} - \mu_k(\theta_*)|\ |\boldsymbol{w}(t)\right]^{\bar{\gamma}_1}\right],
\end{aligned} \tag{2.5}$$

where $\mathbb{E}[\cdot|\cdot]$ denotes the conditional expectation. Using Hoeffding's inequality, for each $k \in \mathcal{K}$,

$$\begin{aligned}
&\mathbb{E}\left[|\hat{X}_{k,t} - \mu_k(\theta_*)|\ |\boldsymbol{w}(t)\right] \\
&= \int_{x=0}^{1} \Pr\left(|\hat{X}_{k,t} - \mu_k(\theta_*)| > x|\boldsymbol{w}(t)\right) \mathrm{d}x \\
&\leq \int_{x=0}^{\infty} 2\exp(-2x^2 N_k(t))\, \mathrm{d}x \leq \sqrt{\frac{\pi}{2N_k(t)}}.
\end{aligned} \tag{2.6}$$

Combining (2.5) and (2.6), the following is obtained:

$$\mathbb{E}[|\theta_* - \hat{\theta}_t|] \leq \bar{D}_1 (\frac{\pi}{2})^{\frac{\bar{\gamma}_1}{2}} \frac{1}{t^{\frac{\bar{\gamma}_1}{2}}} \mathbb{E}\left[\sum_{k=1}^{K} w_k(t)^{1-\frac{\bar{\gamma}_1}{2}}\right]. \tag{2.7}$$

Since $w_k(t) \leq 1$ for all $k \in \mathcal{K}$, and $\sum_{k=1}^{K} w_k(t) = 1$ for any possible $\boldsymbol{w}(t)$, we have $\mathbb{E}[\sum_{k=1}^{K} w_k(t)^{1-\frac{\bar{\gamma}_1}{2}}] \leq K^{\frac{\bar{\gamma}_1}{2}}$. Then, combining (2.4) and (2.7), we have

$$\mathbb{E}[r_{t+1}(\theta_*)] \leq 2\bar{D}_1^{\gamma_2} D_2 \frac{\pi}{2}^{\frac{\bar{\gamma}_1\gamma_2}{2}} K^{\frac{\bar{\gamma}_1\gamma_2}{2}} \frac{1}{t^{\frac{\bar{\gamma}_1\gamma_2}{2}}}.$$

2.8.6 Proof of Theorem 3

This bound is consequence of Theorem 2 and the inequality given in bound where for $\gamma > 0$ and $\gamma \neq 1$, $\sum_{t=1}^{T} 1/t^{\gamma} \leq 1 + \frac{(T^{1-\gamma}-1)}{1-\gamma}$, i.e.,

$$\mathrm{Reg}(\theta_*, T) \leq 2 + \frac{2\bar{D}_1^{\gamma_2} D_2 \frac{\pi}{2}^{\frac{\gamma_1\gamma_2}{2}} K^{\frac{\bar{\gamma}_1\gamma_2}{2}}}{1 - \frac{\bar{\gamma}_1\gamma_2}{2}} T^{1-\frac{\bar{\gamma}_1\gamma_2}{2}}.$$

2.8.7 Proof of Theorem 4

In order to complete the proof, the probability of the event that $\{I_t \notin \mathcal{K}^*(\theta_*)\}$ needs to be bounded. Since at time $t+1$, the arm with the highest $\mu_k(\hat{\theta}_t)$ is selected by the WAGP, $\hat{\theta}_t$ should lie in $\Theta \setminus \Theta_{k^*(\theta_*)}$ for a suboptimal arm to be selected. Therefore,

$$\begin{aligned} &\{I_{t+1} \notin \mathcal{K}^*(\theta_*)\} \\ &= \{\hat{\theta}_t \in \Theta \setminus \Theta_{k^*(\theta_*)}\} \subseteq \mathcal{G}_{\theta_*, \hat{\theta}_t}(\Delta_*). \end{aligned} \tag{2.8}$$

By Lemma 3 and (2.8),

$$\begin{aligned} &\Pr(I_{t+1} \notin \mathcal{K}^*(\theta_*)) \\ &\leq \sum_{k=1}^{K} \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}\left(\mathcal{F}^k_{\theta_*, \hat{\theta}_t}\left(\left(\frac{\Delta_*}{w_k(t)\bar{D}_1 K}\right)^{\frac{1}{\bar{\gamma}_1}}\right)\right) | \boldsymbol{N}(t)\right]\right] \\ &\leq \sum_{k=1}^{K} 2\mathbb{E}\left[\exp\left(-2\left(\frac{\Delta_*}{w_k(t)\bar{D}_1 K}\right)^{\frac{2}{\bar{\gamma}_1}} w_k(t) t\right)\right] \\ &\leq 2K \exp\left(-2\left(\frac{\Delta_*}{\bar{D}_1 K}\right)^{\frac{2}{\bar{\gamma}_1}} t\right), \end{aligned} \tag{2.9}$$

where $\mathbb{I}(\cdot)$ is indicator function which is 1 if the statement is correct and 0 otherwise, the first inequality follows from a union bound, the second inequality is obtained by using the Chernoff-Hoeffding bound, and the last inequality is obtained by using Lemma 4. Then, $\Pr(I_{t+1} \notin \mathcal{K}^*(\theta_*)) \leq 1/t$ for $t > C_1(\Delta_*)$ and $\Pr(I_{t+1} \notin \mathcal{K}^*(\theta_*)) \leq 1/t^2$ for $t > C_2(\Delta_*)$. The bound in the first regime is the result of Theorem 3. The bounds in the second and third regimes are obtained by summing the probability given in (2.9) from $C_1(\Delta_*)$ to T and $C_2(\Delta_*)$ to T, respectively.

2.8.8 Proof of Theorem 5

Let $(\Omega, \mathcal{F}, P)$ denote probability space, where Ω is the sample set and $\mathcal{F}$ is the σ-algebra that the probability measure P is defined on. Let $\omega \in \Omega$ denote a sample path. Then, the following needs to be proved: There exists event $N \in \mathcal{F}$ such that $P(N) = 0$ and if $\omega \in N^c$, then $\lim_{t\to\infty} I_t(\omega) \in \mathcal{K}^*(\theta_*)$. Define the event $\mathcal{E}_t := \{I_t \neq k^*(\theta_*)\}$. We show in the proof of Theorem 4 that $\sum_{t=1}^{T} P(\mathcal{E}_t) < \infty$. By Borel-Cantelli lemma,

$$\Pr(\mathcal{E}_t \text{ infinitely often}) = \Pr(\limsup_{t\to\infty} \mathcal{E}_t) = 0.$$

Define $N := \limsup_{t\to\infty} \mathcal{E}_t$, where $\Pr(N) = 0$. Then,

$$N^{\mathrm{c}} = \liminf_{t\to\infty} \mathcal{E}_t^{\mathrm{c}},$$

where $\Pr(N^{\mathrm{c}}) = 1 - \Pr(N) = 1$, which means that $I_t \in \mathcal{K}^*(\theta_*)$ for all but a finite number of time t.

2.8.9 Proof of Theorem 6

Consider a problem instance with two arms with reward functions $\mu_1(\theta) = \theta^\gamma$ and $\mu_2(\theta) = 1 - \theta^\gamma$, where γ is an odd positive integer and rewards are Bernoulli distributed with $X_{1,t} \sim \mathrm{Ber}(\mu_1(\theta))$ and $X_{2,t} \sim \mathrm{Ber}(\mu_2(\theta))$. Then, optimality regions are $\Theta_1 = [2^{-\frac{1}{\gamma}}, 1]$ and $\Theta_2 = [0, 2^{-\frac{1}{\gamma}}]$. Note that $\gamma_2 = 1$ and $\gamma_1 = 1/\gamma$ for this case. It can be shown that

$$|\mu_k(\theta) - \mu_k(\theta')| \leq D_2|\theta - \theta'|$$
$$|\mu_k^{-1}(x) - \mu_k^{-1}(x')| \leq \bar{D}_1|x - x'|^{1/\gamma}$$

Let $\theta^* = 2^{-\frac{1}{\gamma}}$. Consider following two cases with $\theta_1^* = \theta^* + \Delta$ and $\theta_2^* = \theta^* - \Delta$. The optimal arm is 1 in the first case and 2 in the second case. In the first case, one step loss due to choosing arm 2 is lower bounded by

$$\begin{aligned}
&(\theta^* + \Delta)^\gamma - (1 - (\theta^* + \Delta)^\gamma) \\
&= 2(\theta^* + \Delta)^\gamma - 1 \\
&= 2((\theta^*)^\gamma + \binom{\gamma}{1}(\theta^*)^{\gamma-1}\Delta + \binom{\gamma}{2}(\theta^*)^{\gamma-2}\Delta^2 + \ldots) - 1 \\
&\geq 2\gamma 2^{\frac{1-\gamma}{\gamma}}\Delta.
\end{aligned}$$

Similarly, in the second case, the loss due to choosing arm 1 is $2\gamma 2^{\frac{1-\gamma}{\gamma}}\Delta + \sum_{i=2}^{\gamma}\binom{\gamma}{i}(\theta^*)^{(\gamma-i)}(-\Delta)^i$. Let $A_1(\Delta) = 2\gamma 2^{\frac{1-\gamma}{\gamma}}\Delta + \sum_{i=2}^{\gamma}\binom{\gamma}{i}(\theta^*)^{(\gamma-i)}(-\Delta)^i$.

Define two processes $\nu_1 = \mathrm{Ber}(\mu_1(\theta^* + \Delta)) \otimes \mathrm{Ber}(\mu_2(\theta^* + \Delta))$ and $\nu_2 = \mathrm{Ber}(\mu_1(\theta^* - \Delta)) \otimes \mathrm{Ber}(\mu_2(\theta^* - \Delta))$ where $x \otimes y$ denotes the product distribution of x and y. Let $\Pr_\nu$ denote probability associated with distribution ν. Then, the following holds:

$$\begin{aligned}
&\mathrm{Reg}(\theta^* + \Delta, T) + \mathrm{Reg}(\theta_* - \Delta, T) \\
&\quad \geq A_1(\Delta) \sum_{t=1}^{T} \Big(\Pr_{\nu_1^{\otimes t}}(I_t = 2) + \Pr_{\nu_2^{\otimes t}}(I_t = 1) \Big), \qquad (2.10)
\end{aligned}$$

where $\nu^{\otimes t}$ is the t times product distribution of ν. Using well-known lower bounding techniques for the minimax risk of hypothesis testing [TZ09], the following holds:

$$\begin{aligned}
&\mathrm{Reg}(\theta^* + \Delta, T) + \mathrm{Reg}(\theta^* - \Delta, T) \qquad (2.11) \\
&\quad \geq A_1(\Delta) \sum_{t=1}^{T} \exp(-\mathrm{KL}(\nu_1^{\otimes t}, \nu_2^{\otimes t})), \qquad (2.12)
\end{aligned}$$

where

$$\begin{aligned}
\mathrm{KL}(\nu_1^{\otimes t}, \nu_2^{\otimes t}) = t\Big(&\mathrm{KL}(\mathrm{Ber}(\mu_1(\theta^* + \Delta)), \mathrm{Ber}(\mu_1(\theta^* - \Delta)) \\
&+ \mathrm{KL}(\mathrm{Ber}(\mu_2(\theta^* + \Delta)), \mathrm{Ber}(\mu_2(\theta^* - \Delta))\Big). \qquad (2.13)
\end{aligned}$$

Define $A_2 = (1 - \exp(\frac{-4D_2^2\Delta^2 T}{(\theta^* - \Delta)^\gamma(1-(\theta^* - \Delta)^\gamma)}))(\theta^* - \Delta)^\gamma(1 - (\theta^* - \Delta)^\gamma)$.By using the fact $\mathrm{KL}(p, q) \leq$

$\frac{(p-q)^2}{q(1-q)}$ [RZ10], we can further bound (2.12) by

$$\begin{aligned}
&\mathrm{Reg}(\theta^* + \Delta, T) + \mathrm{Reg}(\theta^* - \Delta, T) \\
&\geq A_1(\Delta) \sum_{t=1}^{T} \exp\left(-\frac{4D_2 t \Delta^2}{(\theta^* - \Delta)^\gamma (1 - (\theta^* - \Delta)^\gamma)} \right) \\
&\geq A_1(\Delta) \frac{A_2}{4D_2\Delta^2}
\end{aligned}$$

where $A_2 \in (0,1)$ for any $\Delta \in (0, \max(\theta^*, 1-\theta^*))$. Hence, the lower bound for the parameter dependent regret is $\Omega(1)$. In order to show the lower bound for the worst-case regret, observe that

$$\begin{aligned}
&\mathrm{Reg}(\theta^* + \Delta, T) + \mathrm{Reg}(\theta^* - \Delta, T) \\
&\geq \frac{2\gamma 2^{\frac{1-\gamma}{\gamma}} A_2}{4D_2\Delta} + \sum_{i=2}^{\gamma} \binom{\gamma}{i} (-\Delta)^{i-2} (\theta^*)^{\gamma-i}.
\end{aligned}$$

By choosing $\Delta = \frac{1}{\sqrt{T}}$, it is shown that for large T, $A_2 = 0.25(1 - \exp(-16D_2^2))$. Hence, worst-case lower bound is $\Omega(\sqrt{T})$.

2.8.10 Proof of Theorem 7

Without loss of generality, assume that a unique arm is optimal for $\hat{\theta}_t$ and θ_*. First, the following is shown: $|\hat{\theta}_t - \theta_*| = \epsilon$ implies $|\hat{\Delta}_t - \Delta_*| \leq \epsilon$. There are four possible cases for $\hat{\Delta}_t$:

- θ_* and $\hat{\theta}_t$ lie in the same optimality interval of the optimal arm, and Δ_* and $\hat{\Delta}_t$ are computed with respect to the same endpoint of that interval.
- θ_* and $\hat{\theta}_t$ lie in the same optimality interval and Δ_* and $\hat{\Delta}_t$ are computed with respect to the different endpoints of that interval.
- θ_* and $\hat{\theta}_t$ lie in adjacent optimality intervals.
- θ_* and $\hat{\theta}_t$ lie in non-adjacent optimality intervals.

In the first case, $|\hat{\theta}_t - \theta_*| = |\hat{\Delta}_t - \Delta_*| = \epsilon$. In the second case, $\hat{\Delta}_t$ can not be larger than $\Delta_* + \epsilon$ since in that case $\hat{\theta}_t$ would be computed with respect to the same endpoint of

that interval. Similarly, $\hat{\Delta}_t$ can not be smaller than $\Delta_* - \epsilon$ since in that case θ_* would be computed with respect to the same endpoint of that interval. In the third and fourth cases, since $|\hat{\theta}_t - \theta_*| = \epsilon$, $\hat{\Delta}_t \leq \epsilon - \Delta_*$, and hence the difference between $\hat{\Delta}_t$ and Δ_* is smaller than ϵ.

Second, the following is shown: $|\hat{\Delta}_t - \Delta_*| < \bar{D}_1 \left(\frac{2K \log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}}$ holds with high probability.

$$
\begin{aligned}
&\Pr\left(|\hat{\Delta}_t - \Delta_*| \geq \bar{D}_1 \left(\frac{K \log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}}\right) \\
&\leq \Pr\left(|\hat{\theta}_t - \theta_*| \geq \bar{D}_1 \left(\frac{K \log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}}\right) \\
&\leq \sum_{k=1}^{K} 2\mathbb{E}\left[\exp\left(-2\left(\frac{\bar{D}_1 K \left(\frac{\log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}}}{\bar{D}_1 K w_k(t)}\right)^{\frac{2}{\bar{\gamma}_1}} N_k(t)\right) \middle| N_k(t)\right] \\
&\leq \sum_{k=1}^{K} 2\mathbb{E}\left[\exp\left(-2 w_k(t)^{1-\frac{2}{\bar{\gamma}_1}} \log t\right) \middle| w_k(t)\right] \\
&\leq 2Kt^{-2}, \qquad (2.14)
\end{aligned}
$$

where the second inequality follows from Lemma 3 and Chernoff-Hoeffding inequality and third inequality by Lemma 4. Then, at time t, with probability at least $1 - 2Kt^{-2}$, the following holds:

$$
\Delta_* - 2\bar{D}_1 K \left(\frac{\log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}} \leq \tilde{\Delta}_t. \qquad (2.15)
$$

Also, note that if $t \geq C_2(\Delta_*/3)$, then $2\bar{D}_1 K \left(\frac{\log t}{t}\right)^{\frac{\bar{\gamma}_1}{2}} \leq \frac{2\Delta_*}{3}$. Thus, for $t \geq C_2(\Delta_*/3)$, we have $\Delta_*/3 \leq \tilde{\Delta}_t$. Note that the BUW follows UCB1 only when $t < C_2(\tilde{\Delta}_t)$. From the above, we know that $C_2(\tilde{\Delta}_t) \leq C_2(\Delta_*/3)$ when $t \geq C_2(\Delta_*/3)$ with probability at least $1 - 2Kt^{-2}$. This implies that the BUW follows the WAGP with probability at least $1 - 2Kt^{-2}$ when $t \geq C_2(\Delta_*/3)$.

It is known from Theorem 4 that the WAGP selects an optimal action with probability at least $1 - 1/t^2$ when $t > C_2(\Delta_*)$. Since $C_2(\Delta_*/3) > C_2(\Delta_*)$, when the BUW follows the WAGP, it will select an optimal action with probability at least $1 - 1/t^2$ when $t > C_2(\Delta_*/3)$.

Let I_t^g denote the action that selected by algorithm $g \in \{\text{BUW}, \text{WAGP}, \text{UCB1}\}$, $r_t^g(\theta_*) = \mu^*(\theta_*) - \mu_{I_t^g}(\theta_*)$ denote the one-step regret, and $R_{\theta_*}^g(T_1, T_2)$ denote the cumulative regret incurred by algorithm g from T_1 to T_2. Then, when $T < C_2(\Delta_*/3)$, the regret of the BUW can be written as

$$\begin{aligned} R_{\theta^*}^{BUW}(1,T) &\leq \sum_{t=1}^{T} r_t^{UCB1}(\theta_*) + 2Kt^{-2} \\ &\leq R_{\theta^*}^{UCB1}(1,T) + \frac{2K\pi^2}{3}. \end{aligned}$$

Moreover, when $T \geq C_2(\Delta_*/3)$,

$$\begin{aligned} & R_{\theta^*}^{BUW}(C_2(\Delta_*/3), T) \\ & \leq \sum_{t=C_2(\Delta_*/3)}^{T} r_t^{WAGP}(\theta_*) + 2Kt^{-2} \\ & \leq R_{\theta^*}^{WAGP}(C_2(\Delta_*/3), T) + \frac{2K\pi^2}{3} \end{aligned}$$

This concludes the parameter-dependent regret bound.

The worst-case bound can be proven by replacing $\delta_k = \mu^* - \mu_k = 1/\sqrt{TK\log T}$ for all $k \notin \mathcal{K}^*(\theta_*)$ for the regret bound given above.

2.8.11 Auxiliary Lemma

Lemma 4. *For $\gamma < 0$, $\delta > 0$, the following bound holds for any w_k with $0 \leq w_k \leq 1$ and $\sum_{k=1}^{K} w_k = 1$:*

$$\sum_{k=1}^{K} \exp(-\delta w_k^{\gamma}) \leq K \exp(-\delta)$$

Proof. Let $k_{\max} = \arg\max_k w_k$. Then,

$$\begin{aligned}
& \max_{w_k:\sum_{k=1}^K w_k=1,\, 0\le w_k\le 1} \sum_{k=1}^K \exp\left(-\delta w_k^\gamma\right) \\
&= \max_{w_k:\sum_{k=1}^K w_k=1,\, 0\le w_k\le 1} \exp\left(\log\left(\sum_{k=1}^K \exp\left(-\delta w_k^\gamma\right)\right)\right) \\
\le\ & \max_{w_k:\sum_{k=1}^K w_k=1,\, 0\le w_k\le 1} \exp\left(\max_{k\in\mathcal{K}}\left(-\delta w_k^\gamma\right) + \log K\right) \\
\le\ & K \max_{w_k:\sum_{k=1}^K w_k=1,\, 0\le w_k\le 1} \exp\left(-\delta w_{k_{\max}}^\gamma\right) \\
\le\ & K\exp(-\delta).
\end{aligned}$$

□

CHAPTER 3

Constructing Effective Policies from Observational Datasets with Many Features

3.1 Introduction

The "best" treatment for the current patient must be learned from the treatment(s) of previous patients. However, no two patients are ever *exactly* alike, so the learning process must involve learning the ways in which the current patient *is* alike to previous patients – i.e., has the same or similar features – and which of those features are *relevant* to the treatment(s) under consideration. This already complicated learning process is further complicated because the history of previous patients records only outcomes actually experienced from treatments actually received – not the outcomes that would have been experienced from alternative treatments – the *counterfactuals*. And this learning process is complicated still further because the treatments received by previous patients were (typically) chosen according to some protocol that might or might not be known but was almost surely not random – so the observed data is *biased*.

This chapter proposes a novel approach to addressing such problems. An algorithm that learns a nonlinear policy to recommend an action for each (new) instance is proposed. During the training phase, our algorithm learns the action-dependent relevant features and then uses a feedforward neural network to optimize a nonlinear stochastic policy the output of which is a probability distribution over the actions given the relevant features. When the trained algorithm is applied to a new instance, it chooses the action which has the highest probability. Our algorithm is evaluated in actual data in which our algorithm is significantly superior to existing state-of-the-art algorithms. The proposed methods and the algorithms are widely

Table 3.1: Success rates of two treatments for kidney stones [BPC13]

	Overall	Small stones	Large stones
Open Surgery	78%(273/350)	**93%(81/87)**	**73%(192/263)**
Percutaneous Nephrolithotomy	**83%(289/350)**	87%(234/270)	69%(55/80)

applicable to an enormous range of settings in many medical informatics. In the medical context, typical features are items available in the electronic health record (laboratory tests, previous diagnoses, demographic information, etc.), typical actions are choices of treatments (perhaps including no treatment at all), and typical rewards are recovery rates or 5-year survival rates.

For a simple but striking example from the medical context, consider the problem of choosing the best treatment for a patient with kidney stones. Such patients are usually classified by the size of the stones: small or large; the most common treatments are Open Surgery and Percutaneous Nephrolithotomy. Table 1 summarizes the results. Note that Open Surgery performs better than Percutaneous Nephrolithotomy for patients with small stones *and* for patients with large stones but Percutaneous Nephrolithotomy performs better overall.[1] Of course this would be impossible if the subpopulations that received the two treatments were identical – but they were not. And in fact the policy that created these subpopulations by assigning patients to treatments is not known a priori. The patients are distinguished by a vast array of features in addition to the size of stones – age, gender, weight, kidney function tests, etc. – but relevant features are not known. And of course result of the treatment actually received by each patient is known – but what the result of the alternative treatment would have been (the counterfactual) is not known.

Three more points should be emphasized. Although Table 3.1 shows only two actions, in fact there are a number of other possible actions for kidney stones: they could be treated using any of a number of different medications, they could be treated by ultrasound, or

[1]This is a particular instance of Simpson's Paradox.

they could not be treated at all. This is important for several reasons. The first is that a number of existing methods assume that there are only two actions (corresponding to treat or not-treat); but as this example illustrates, in many contexts (and in the medical context in particular), it is *typically* the case that there are *many* actions, not just two – and, as the papers themselves note, these methods simply do not work when there are more than two actions; see [JSS16]. The second is that the features that are relevant for predicting the success of a particular action typically depend on the action: different features will be found to be relevant for different actions. (The treatment of breast cancer, as discussed in [YDS16], illustrates this point well. The issue is not simply whether or not to apply a regime of chemotherapy, but *which* regime of chemotherapy to apply. Indeed, there are at least six widely used regimes of chemotherapy to treat breast cancer, and the features that are relevant for predicting success of a given regime are different for different regimes.) The third is that this chapter goes much further than the existing literature by allowing for *nonlinear* policies. To do this, the algorithm uses a feedforward neural network, rather than relying on familiar algorithms such as POEM [SJ15a]. To determine the best treatment, the bias in creating the populations, the features that are relevant *for each action* and the policy must all be *learned*. The proposed methods are adequate to this task.

3.2 Related Work

From a conceptual point of view, the paper most closely related to this chapter – at least among recent papers – is perhaps [JSS16] which treats a similar problem: learning relevance in an environment in which the counterfactuals are missing, data is biased and each instance may have many features. The approach taken there is somewhat different from ours in that, rather than identifying the relevant features, they transfer the features to a new representation space. (This process is referred as *domain adaptation* [JSS16].) A more important difference from this chapter is that it assumes that there are only two actions: treat and don't treat. As discussed in the Introduction, the assumption of two actions is unrealistic; in most situations there will be *many* (possible) actions. It states explicitly that the approach

taken there does not work when there are more than two actions and offers the multi-action setting as an obvious but difficult challenge. One might think of this chapter as "solving" this challenge – but the "solution" is not at all a routine extension. Moreover, in addition to this obvious challenge, there is a more subtle – but equally difficult – challenge: when there are more than two actions, it will typically be the case that some features will be relevant for some actions and not for others, and – as discussed in the Introduction – it will be crucial to learn which features are relevant for which actions.

From a technical point of view, this chapter is perhaps most closely related to [SJ15a] in that we use similar methods (IPS-estimates and empirical Bernstein inequalities) to learn counterfactuals. However, it does not treat observational data in which the bias is unknown and does not learn/identify relevant features. Another similar work on policy optimization from observational data is [SLL10].

The work in [WA15] treats the related (but somewhat different) problem of estimating individual treatment effects. The approach there is through causal forests as developed by [AI15], which are variations on the more familiar random forests. However, the emphasis in this work is on asymptotic estimates, and in the many situations for which the number of (possibly) relevant features is large the datasets will typically not be large enough that asymptotic estimates will be of more than limited interest. There are many other works focusing on estimating treatment effects; some include [TAG12, AS17, **?**].

More broadly, our work is related to methods for feature selection and counterfactual inference. The literature on feature selection can be roughly divided into categories according to the extent of supervision: supervised feature selection [SSG12, WES03], unsupervised feature selection [DB04, HCN05] and semi-supervised feature selection [XKL10]. However, this chapter does not fall into any of these categories; instead we need to select features that are informative in determining the rewards of each action. This problem was addressed in [TS14] but in an *online* Contextual Multi-Armed Bandit (CMAB) setting in which experimentation is used to learn relevant features. This chapter treats the *logged* CMAB setting in which experimentation is impossible and relevant features must be learned from the existing logged data. As already noted, there are many circumstances in which experimentation is impossi-

ble. The difference between the settings is important – and the logged setting is much more difficult – because in the online setting it is typically possible to *observe* counterfactuals, while in the current logged setting it is typically *not* possible to observe counterfactuals, and because in the online setting the decision-maker controls the observations so whatever bias there is in the data is known.

With respect to learning, feature selection methods can be divided into three categories – filter models, wrapper models, and embedded models [TAL14]. This chapter is most similar to filter techniques in which features are ranked according to a selected criterion such as a Fisher score [DHS12], correlation based scores [SSG12], mutual information based scores [KS96, YL03, PLD05], Hilbert-Schmidt Independence Criterion (HSIC) [SSG12] and Relief and its variants [KR92, RK03]) etc., and the features having the highest ranks are labeled as relevant. However, these existing methods are developed for classification problems and they cannot easily handle datasets in which the rewards of actions not taken are missing.

The literature on counterfactual inference can be categorized into three groups: direct, inverse propensity re-weighting and doubly robust methods. The direct methods compute counterfactuals by learning a function mapping from feature-action pair to rewards [Pre76, WA15]. The inverse propensity re-weighting methods compute unbiased estimates by weighting the instances by their inverse propensity scores [SJ15a, JS16]. The doubly robust methods compute the counterfactuals by combining direct and inverse propensity score reweighing methods to compute more robust estimates [DLL11, JL16]. With respect to this categorization, the techniques developed here might be view as falling into doubly robust methods.

This chapter can be seen as building on and extending the work of [SJ15a, SJ15c], which learn *linear* stochastic policies. This chapter goes much further by learning a *non-linear stochastic policy.*

3.3 Data

Logged contextual bandit data is the data for which the features of each instance, the action taken and the reward realized in that instance are known – but not the reward that would have been realized had a different action been taken. The assumption in this chapter is that the data has been logged according to some policy and so the data is *biased*. Each data point consists of a feature, an action and a reward. A *feature* is a vector $(x_1, \ldots, x_d)$ where each $x_i \in \mathcal{X}_i$ is a *feature type*. The space of all feature types is $\mathcal{F} = \{1, \ldots, d\}$, the space of all features is $\mathcal{X} = \Pi_{i=1}^{d} \mathcal{X}_i$ and the set of *actions* is $\mathcal{A}$. Another assumption in this chapter is that that the sets of feature types and actions are finite; write $b_i = |\mathcal{X}_i|$ for the cardinality of $\mathcal{X}_i$ and $\mathcal{A} = \{1, 2, \ldots, k\}$ for the set of actions. For $\boldsymbol{x} \in \mathcal{X}$ and $\mathcal{S} \subset \mathcal{F}$, write $\boldsymbol{x}_{\mathcal{S}}$ for the restriction of $\boldsymbol{x}$ to $\mathcal{S}$; i.e. for the vector of feature types whose indices lie in $\mathcal{S}$. It will be convenient to abuse notation and view $\boldsymbol{x}_{\mathcal{S}}$ both as a vector of length $|\mathcal{S}|$ or as a vector of length $d = |\mathcal{F}|$ which is 0 for feature types not in $\mathcal{S}$. A *reward* is a real number; rewards lie in the interval $[0, 1]$. In some cases, the reward will be either 1 or 0 (success or failure; good or bad outcome); in other cases the reward may be interpreted as the probability of a success or failure (good or bad outcome).

We are given a data set

$$\mathcal{D}^n = \{(\boldsymbol{X}_1, A_1, R_1^{\text{obs}}), \ldots, (\boldsymbol{X}_n, A_n, R_n^{\text{obs}})\}$$

The j-th instance/data point $(\boldsymbol{X}_j, A_j, R_j^{\text{obs}})$ is generated according to the following process:

1. The instance is described by a feature vector $\boldsymbol{X}_j$ that arrives according to the fixed but unknown distribution $\Pr(\mathcal{X})$; $\boldsymbol{X}_j \sim \Pr(\mathcal{X})$.

2. The action taken was determined by a policy that draws actions at random according to a (possibly unknown) probability distribution $p_0(\mathcal{A}|\boldsymbol{X}_j)$ on the action space $\mathcal{A}$. (Note that the distribution of actions taken depends on the vector of features).

3. Only the reward of the action actually performed is recorded into the dataset, i.e., $R_j^{\text{obs}} \equiv R_j(A_j)$.

4. For every action a, either taken or not taken, the reward $R_j(a) \sim \Phi_a(\cdot|\boldsymbol{X}_j)$ that would have been realized had a actually been taken is generated by a random draw from an unknown family $\{\Phi_a(\cdot|\boldsymbol{x})\}_{\boldsymbol{x}\in\mathcal{X},a\in\mathcal{A}}$ of reward distributions with support $[0,1]$.

The logging policy corresponds to the choices made by the existing decision-making procedure and so will typically create a biased distribution on the space of feature-action pairs.

This chapter makes two natural assumptions about the rewards and the logging policy; taken together they enable us to generate unbiased estimates of the variables of the interest. The first assumption guarantees that there is enough information in the data-generating process so that counterfactual information can be inferred from what is actually observed.

Assumption 2. *(Common support)* $p_0(a|\boldsymbol{x}) > 0$ *for all action-feature pairs* $(a, \boldsymbol{x})$.

The second assumption is that the logging policy depends only on the observed features – and not on the observed rewards.

Assumption 3. *(Unconfoundness) For each feature vector* $\boldsymbol{X}$*, the rewards of actions* $\{R(a)\}_{a\in\mathcal{A}}$ *are statistically independent of the action actually taken;* $\{R(a)\} \perp\!\!\!\perp A \big| \boldsymbol{X}$.

These assumptions are universal in the counterfactual inference literature – see [JSS16, AI15] for instance – although they can be criticized on the grounds that their validity cannot be determined on the basis of what is actually observed.

3.4 The Algorithm

It seems useful to begin with a brief overview; more details and formalities follow below. The proposed algorithm consists of a training phase and an execution phase; the training phase consists of three steps.

A. In the first step of the training phase, the algorithm either inputs the true propensity scores (if they are known) or uses the logged data to estimate propensity scores (when the true propensity scores are not known); this (partly) corrects the bias in the logged data.

B. In the second step of the training phase, the algorithm uses the known or estimated propensity scores to compute, for each action and each feature, an estimate of relevance for that feature with respect to that action. The algorithm then retains the more relevant features – those for which the estimate is above a threshold – and discards the less relevant features – those for which the estimate is below the threshold. (For reasons that will be discussed below, the threshold used depends on both the action and the feature type.)

C. In the third step of the training phase, the algorithm uses the known or estimated propensity scores and the features identified as relevant, and trains a feedforward neural network model to learn a non-linear stochastic policy that minimizes the "corrected" cross entropy loss.

In the execution phase, the algorithm is presented with a new instance and uses the policy derived in the training phase to recommend an action for this new instance on the basis of the relevant features of that instance.

3.4.1 True Propensities

This chapter begins with the setting in which propensities of the randomized algorithm are actually tracked and available in the dataset. This is often the case in the advertising context, for example. In this case, for each j, set $p_{0,j} = p_0(A_j|X_j)$, and write $\boldsymbol{P}_0 = [p_{0,j}]_{j=1}^n$; this is the vector of *true propensities*.

3.4.2 Relevance

It might seem natural to define the set $\mathcal{S}$ of feature types to be *irrelevant* (for a particular action) if the distribution of rewards (for that action) is independent of the features in $\mathcal{S}$, and to define the set $\mathcal{S}$ to be *relevant* otherwise. In theoretical terms, this definition has much to recommend it. In operational terms, however, this definition is not of much use. That is because finding irrelevant sets of feature types would require many observations (to determine the entire distribution of rewards) and intractable calculations (to examine all sets

of feature types). Moreover, this notion of irrelevance will often be too strong because our interest will often be only in maximizing expected rewards (or more generally some statistical function of rewards), as it would be in the medical context if the reward is five-year survival rate, or in the advertising or financial settings, if the reward is expected revenue or profit and the advertiser or firm is risk-neutral.

Given these objections, this chapter takes an alternative approach. This chapter first defines a measure of how relevant a particular feature type is for the expected reward of a particular action, learn/estimate this measure from observed data, retain features for which this measure is above some endogenously derived threshold (the most relevant features) and discard other features (the least relevant features). Of course, this approach has drawbacks. Most obviously, it might happen that two feature types are individually not very relevant but are jointly quite relevant. However, as shown empirically, this approach has the virtue that it works: the proposed algorithm on the basis of this approach is demonstrably superior to existing algorithms.

3.4.2.1 True Relevance

To begin formalizing the measure of relevance, fix an action a, a feature vector x and a feature type i. Define expected rewards and marginal expected rewards as follows:

$$\begin{aligned} \bar{r}(a, \boldsymbol{x}) &= \mathbb{E}\left[R(a) | \boldsymbol{X} = \boldsymbol{x}\right] \\ \bar{r}(a, \boldsymbol{x}_i) &= \mathbb{E}_{\boldsymbol{X}_{-i}}[\bar{r}(a, \boldsymbol{X}) \Big| \boldsymbol{X}_i = \boldsymbol{x}_i] \\ \bar{r}(a) &= \mathbb{E}_{\boldsymbol{X}}\left[\bar{r}(a, \boldsymbol{X})\right] \end{aligned} \tag{3.1}$$

Define the *true relevance of feature type i for action a* by

$$g(a, i) = \mathbb{E}\left[\ell\left(\bar{r}(a, X_i) - \bar{r}(a)\right)\right], \tag{3.2}$$

where the expectation is taken with respect to the arrival probability distribution of X_i and $\ell(\cdot)$ denotes the loss metric. (Keep in mind that the true arrival probability distribution of X_j is unknown and must be estimated from the data.) The results hold for an arbitrary loss

function, assuming only that it is strictly monotonic and Lipschitz; i.e. there is a constant B such that $|\ell(r) - \ell(r')| \leq B|r - r'|$. These conditions are satisfied by a large class of loss functions including l_1 and l_2 losses. The relevance measure g expresses the weighted difference between the expected reward *of a given action* conditioned on the feature type i and the unconditioned expected reward; $g(a, i) = 0$ exactly when feature type i does not affect the expected reward of action a.[2]

g is referred to as *true* relevance because it is computed using the *true* arrival distribution – but the true arrival distribution is unknown. Hence, even when the true propensities are known, relevance must be *estimated* from observed data. This is the next task.

3.4.2.2 Estimated Relevance

In this subsection, *estimates* of relevance based on observed data (continuing to assume that true propensities are known) are derived. To do so, this chapter first shows how to estimate $\bar{r}(a)$ and $\bar{r}(a, x_i)$ for $x_i \in \mathcal{X}_i$, $i \in \mathcal{F}$ and $a \in \mathcal{A}$ from available observational data. An obvious way to do this is through classical supervised learning based estimators; most obviously, the sample mean estimators for $\bar{r}(a)$ and $\bar{r}(a, x_i)$. However using straightforward sample mean estimation would be wrong because the logging policy introduces a bias into observations. Following the idea of Inverse Propensity Scores [RR83], this bias is corrected by using Importance Sampling.

Write $N(a)$, $N(x_i)$, $N(a, x_i)$ for the number of observations (in the given data set) with action a, with feature x_i, and with the pair consisting of action a and feature x_i, respectively. Rewrite our previous definitions as:

$$\begin{aligned} \bar{r}(a, x_i) &= \mathbb{E}_{(\boldsymbol{X}, A, R^{\text{obs}}) \sim p_0} \left[\frac{\mathbb{I}(A = a) R^{\text{obs}}}{p_0(A|\boldsymbol{X})} \middle| X_i = x_i \right] \\ \bar{r}(a) &= \mathbb{E}_{(\boldsymbol{X}, A, R^{\text{obs}}) \sim p_0} \left[\frac{\mathbb{I}(A = a) R^{\text{obs}}}{p_0(A|\boldsymbol{X})} \right] \end{aligned} \tag{3.3}$$

where $\mathbb{I}(\cdot)$ is the indicator function.

[2]Other measures of relevance have been used in the feature selection literature (e.g., especially Pearson correlation [Hal99] and mutual information [YL03]) – but not for relevance of actions.

Let $\mathcal{J}(x_i)$ denote the time indices in which feature type-i is x_i, i.e., $\mathcal{J}(x_i) = \{j \subseteq \{1, 2, \ldots, n\} : X_{i,j} = x_i\}$. The Importance Sampling approach provides unbiased estimates of $\bar{r}(a)$ and $\bar{r}(a, x_i)$ as

$$\begin{aligned} \widehat{R}(a, x_i; \boldsymbol{P}_0) &= \frac{1}{N(x_i)} \sum_{j \in \mathcal{J}(x_i)} \frac{\mathbb{I}(A_j = a) R_j^{\text{obs}}}{p_{0,j}}, \\ \widehat{R}(a; \boldsymbol{P}_0) &= \frac{1}{n} \sum_{j=1}^{n} \frac{\mathbb{I}(A_j = a) R_j^{\text{obs}}}{p_{0,j}}, \end{aligned} \tag{3.4}$$

(The propensities $\boldsymbol{P}_0$ are included in the notation as a reminder that these estimators are using the *true* propensity scores.)

Define the *estimated relevance of feature type i for action a* as

$$\widehat{G}(a, i; \boldsymbol{P}_0) = \frac{1}{n} \sum_{x_i \in \mathcal{X}_i} N(x_i) \ell \left(\widehat{R}(a, x_i; \boldsymbol{P}_0) - \widehat{R}(a; \boldsymbol{P}_0) \right). \tag{3.5}$$

3.4.2.3 Thresholds

By definition, $\widehat{G}$ is an estimate of relevance so the obvious way to select relevant features is to set a threshold τ, identify a feature i as relevant for action a exactly when $\widehat{G}(a, i; \boldsymbol{P}_0) > \tau$, retain the features that are relevant according to this criterion and discard other features.

However, this approach is a bit too naive for (at least) two reasons. The first is that the proposed empirical estimate of relevance $\widehat{G}$ may in fact be far from the true relevance g. The second is that some features may be highly (positively or negatively) correlated with the remaining features, and hence convey less information. To deal with these objections, thresholds $\tau(a, i)$ are constructed as a weighted sum of an empirical estimate of the error in using $\widehat{G}$ instead of g and the (average absolute) correlation of feature type i with other feature types.

To define the first term an empirical (data-dependent bound) on $|\widehat{G} - g|$ is needed. To derive such a bound, the empirical Bernstein inequality [MP09, AMS09] can be used. (Our bound depends on the *empirical variance* of the estimates.)

To simplify notation, define random variables $U(a; \boldsymbol{P}_0) \equiv \frac{\mathbb{I}(A=a) R^{\text{obs}}}{p_0(A|\boldsymbol{X})}$ and $U_j(a; \boldsymbol{P}_0) \equiv$

$\frac{\mathbb{I}(A_j=a)R_j}{p_{0,j}}$. The sample means and variances are:

$$\begin{aligned}
\mathbb{E}_{(\boldsymbol{X},A,R^{\text{obs}})\sim p_0}[U(a;\boldsymbol{P}_0)] &= \bar{r}(a), \\
\mathbb{E}_{(\boldsymbol{X},A,R^{\text{obs}})\sim p_0}[U(a;\boldsymbol{P}_0)\big|X_i=x_i] &= \bar{r}(a,x_i) \\
\widehat{U}(a;\boldsymbol{P}_0) &= \widehat{R}(a;\boldsymbol{P}_0) \\
&= \frac{1}{n}\sum_{j=1}^{n} U_j(a;\boldsymbol{P}_0), \\
\widehat{U}(a,x_i;\boldsymbol{P}_0) &= \widehat{R}(a,x_i;\boldsymbol{P}_0) \\
&= \frac{1}{N(x_i)}\sum_{j\in\mathcal{J}(x_i)} U_j(a;\boldsymbol{P}_0), \\
V_n(a;\boldsymbol{P}_0) &= \frac{1}{n-1}\sum_{j=1}^{n}\Big(U_j(a;\boldsymbol{P}_0)-\widehat{U}(a;\boldsymbol{P}_0)\Big)^2, \\
V_n(a,x_i;\boldsymbol{P}_0) &= \frac{1}{N(x_i)-1}\sum_{j\in\mathcal{J}(x_i)}\Big(U_j(a;\boldsymbol{P}_0)-\widehat{U}(a,x_i;\boldsymbol{P}_0)\Big)^2.
\end{aligned}$$

The weighted average sample variance is:

$$\bar{V}_n(a,i;\boldsymbol{P}_0) = \sum_{x_i\in\mathcal{X}_i}\frac{N(x_i)V_n(a,x_i;\boldsymbol{P}_0)}{n} \tag{3.6}$$

The proposed empirical (data-dependent) bound is given in Theorem 8.

Theorem 8. *For every $n > 0$, every $\delta \in \left[0,\frac{1}{3}\right]$, and every pair, $(a,i) \in (\mathcal{A},\mathcal{D})$, with probability at least $1-3\delta$ we have:*

$$\begin{aligned}
|\widehat{G}(a,i;\boldsymbol{P}_0) - g(a,i)| \;\leq\; B\Bigg(& \sqrt{\frac{2b_i\ln(3/\delta)\bar{V}_n(a,i;\boldsymbol{P}_0)}{n}} \\
& + \sqrt{\frac{2\ln(3/\delta)V_n(a;\boldsymbol{P}_0)}{n}} \\
& + \frac{M\,(b_i+1)\ln 3/\delta}{n}\Bigg) \\
& + \sqrt{\frac{2\,(\ln 1/\delta + b_i\ln 2)}{n}},
\end{aligned}$$

where $M = \max_{a\in\mathcal{A}}\max_{\boldsymbol{x}\in\mathcal{X}} 1/p_0(a|\boldsymbol{x})$.

The error bound given by Theorem 8 consists of four terms: The first term arises from estimation error of $\widehat{R}(a,x_i)$. The second term arises from estimation error of $\widehat{R}(a)$. The third

term arises from estimation error of feature arrival probabilities. The fourth term arises from randomness of the logging policy.

Now write $\rho_{i,j}$ for the Pearson correlation coefficient between two feature types i and j. (Recall that $\rho_{i,j} = +1$ if i, j are perfectly positively correlated, $\rho_{i,j} = -1$ if i, j are perfectly negatively correlated, and $\rho_{i,j} = 0$ if i, j are uncorrelated.) Then the average absolute correlation of feature type i with other features is

$$\left(\frac{1}{d-1}\right)\left(\sum_{j\in\mathcal{F}\setminus\{i\}} |\rho_{i,j}|\right)$$

Now define the thresholds as

$$\tau(a,i) = \lambda_1\sqrt{\frac{b_i\bar{V}_n(a,i;\boldsymbol{P}_0)}{n}} \;+\; \lambda_2\left(\frac{1}{d-1}\right)\left(\sum_{j\in\mathcal{F}\setminus\{i\}} |\rho_{i,j}|\right)$$

where λ_1, λ_2 are weights (hyper-parameters) to be chosen. Notice that the first term is the dominant term in the error bound given in Theorem 8, and is used to set a higher bar for the feature types that are creating the logging policy bias. The statistical distributions of those features within the the action population and the whole population will be different. By setting the threshold as above, the proposed approach trades-off between three objective: (1) selecting the features that are relevant for the rewards of the actions, (2) eliminating the features which create the logging policy bias, (3) minimizing the redundancy in the feature space.

3.4.2.4 Relevant Feature Types

Finally, the set of feature types that are identified as relevant for an action a is given by

$$\widehat{\mathcal{R}}(a) = \left\{i \;:\; \widehat{G}(a,i;\mathbf{P}_0) > \tau(a,i)\right\} \tag{3.7}$$

Set $\widehat{\boldsymbol{\mathcal{R}}} = \left[\widehat{\mathcal{R}}(a)\right]_{a\in\mathcal{A}}$. Let $\boldsymbol{f}_a$ denote a d dimensional vector whose j^{th} element is 1 if j is contained in the set $\mathcal{R}(a)$ and 0 otherwise.

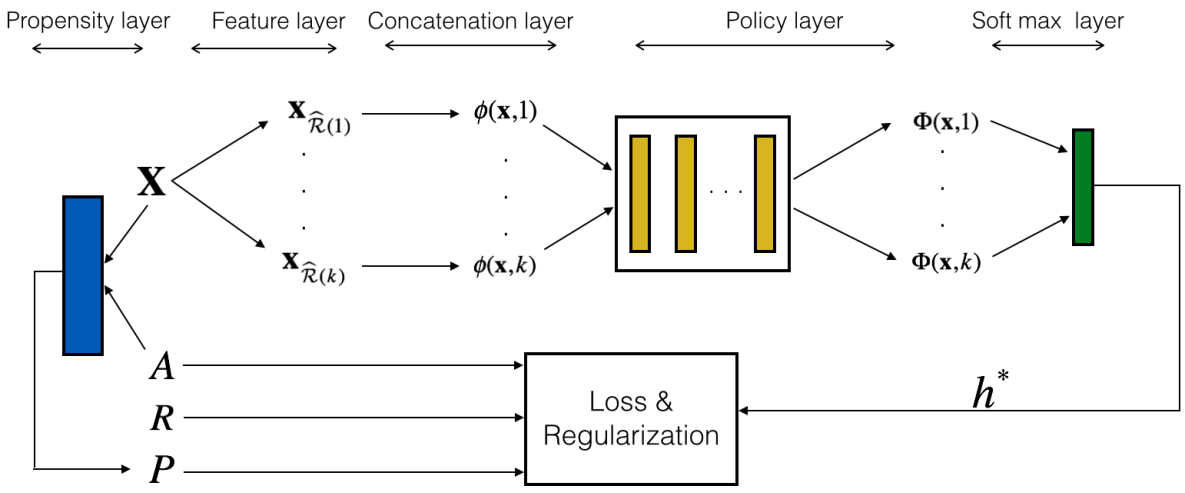


Figure 3.1: Neural network architecture

3.4.3 Policy Optimization

We now build on the identified family of relevant features to construct a policy. By definition, a (stochastic) policy is a map $h : \mathcal{X} \to \triangle(A)$ which assigns to each vector of features a probability distribution $h(\cdot|\boldsymbol{x})$ over actions.

A familiar approach to the construction of stochastic policies is to use the POEM algorithm [SJ15a]. POEM considers only linear stochastic policies; among these, POEM learns one that minimizes risk, adjusted by a variance term. Our approach is substantially more general because we consider arbitrary non-linear stochastic policies. This chapter uses a novel approach that uses a feedforward neural network to find a non-linear policy that minimizes the loss, adjusted by a regularization term. Note that this chapter allows for very general loss and regularization terms so that our approach includes many policy optimizers. If one restricted to a neural network with no hidden layers and a specific regularization term, one would recover POEM.

This chapter proposes a feedforward neural network for learning a policy $h^*(\cdot|\boldsymbol{x})$; the architecture of our neural network is depicted in Fig. 3.1. The proposed feedforward neural network consists of policy layers (L_p hidden layers with $h_p^{(l)}$ units in the l^{th} layer) that use the output of the concatenation layer to generate a policy vector $\Phi(\boldsymbol{x}, a)$, and a softmax

layer that turns the policy vector into a stochastic policy.

For each action a, the concatenation layer takes the feature vector $\boldsymbol{x}$ as an input and generates a action-specific representations $\phi(\boldsymbol{x}, a)$ according to:

$$\begin{aligned} \boldsymbol{x}_{\widehat{\mathcal{R}}(a)} &= \boldsymbol{x} \odot \boldsymbol{f}_a \\ \phi(\boldsymbol{x}, a) &= [\boldsymbol{x}_{\widehat{\mathcal{R}}(\tilde{a})} \mathbb{I}(\tilde{a} = a)]_{\tilde{a} \in \mathcal{A}} \end{aligned}$$

Note that the action-specific representation $\phi(\boldsymbol{x}, a)$ is a $d \times k$ dimensional vector where only the parts corresponding to action a is non-zero and equals to $\boldsymbol{x}_{\widehat{\mathcal{R}}(\tilde{a})}$. For each action a, the policy layers uses the action-specific representation $\phi(\boldsymbol{x}, a)$ generated by the concatenation layers and generates the output vector $\Phi(\boldsymbol{x}, a)$ according to:

$$\Phi(\boldsymbol{x}, a) = \rho \left(\dots \rho \left(\boldsymbol{W}_1^{(p)} \phi(\boldsymbol{x}, a) + \boldsymbol{b}_1^{(p)} \right) \dots + \boldsymbol{b}_{L_p}^{(p)} \right)$$

where $\boldsymbol{W}_l^{(p)}$ and $\boldsymbol{b}_l^{(p)}$ are the weights and bias vectors of the l^{th} layer accordingly. The outputs of the policy layers are used to generate a policy by a softmax layer:

$$h(a|\boldsymbol{x}) = \frac{\exp(\boldsymbol{w}^T \Phi(\boldsymbol{x}, a))}{\sum_{a' \in \mathcal{A}} \exp(\boldsymbol{w}^T \Phi(\boldsymbol{x}, a'))}.$$

Then, the parameters of the policy are chosen in order to minimize an objective of the following form: $\text{Loss}(h^*; \mathcal{D}) + \lambda_3 \mathcal{R}(h^*; \mathcal{D})$; where $\text{Loss}(h^*; \mathcal{D})$ is the loss term, $\mathcal{R}(h^*; \mathcal{D})$ is a regularization term and $\lambda_3 > 0$ represents the trade-off between loss and regularization. The loss function can be either the negative IPS estimate or the corrected cross entropy loss introduced in the next section. Depending on the choice of the loss function and regularizer, our policy optimizer can include a wide-range of objectives including the POEM objective [SJ15a].

In the next subsection, a new objective is proposed, which we refer to as the Policy Neural Network (PONN) objective.

3.4.4 Policy Neural Network (PONN) objective

The proposed PONN objective is motivated by the cross-entropy loss used in the standard multi-class classification setting. In the usual classification setting, usual loss function used

to train the neural network is the standard cross entropy:

$$\widehat{\text{Loss}}_c(h) = \frac{1}{n}\sum_{j=1}^{n}\sum_{a\in\mathcal{A}} -R_j(a)\log h(a|\boldsymbol{X}_j).$$

However, this loss function is not applicable in our setting, for two reasons. The first is that only the rewards of the action taken by the logging policy are recorded in the dataset, not the counterfactuals. The second is that the bias is corrected in the dataset by weighting the instances by their inverse propensities. Hence, the following modified cross entropy loss function is used:

$$\begin{aligned}\widehat{\text{Loss}}_b(h;\boldsymbol{P}_0) &= \frac{1}{n}\sum_{j=1}^{n}\sum_{a\in\mathcal{A}}\frac{-R_j(a)\log h(a|\boldsymbol{X}_j)\mathbb{I}(A_j=a)}{p_{0,j}} \\ &= \frac{1}{n}\sum_{j=1}^{n}\frac{-R_j^{\text{obs}}\log h(A_j|\boldsymbol{X}_j)}{p_{0,j}}. \end{aligned} \tag{3.8}$$

Note that this loss function is an unbiased estimate of the expected cross entropy loss, that is $\mathbb{E}_{(\boldsymbol{X},A,R)\sim p_0}\left[\widehat{\text{Loss}}_b(h^*;\boldsymbol{P}_0)\right] = \mathbb{E}\left[\widehat{\text{Loss}}_c(h^*)\right]$. Our neural network is trained to minimize the regularized loss by Adam optimizer:

$$h^* = \arg\min_{h\in\mathcal{H}} \widehat{\text{Loss}}_b(h;\widehat{\boldsymbol{P}}_0) + \lambda_3\mathcal{R}(h),$$

where $\mathcal{R}(h)$ is the regularization term to avoid overfitting and λ_3 is the hyperparameter to trade-off between the loss and regularization.

3.4.5 Unknown Propensities

As noted, in most settings the logging policy is unknown and hence the actual propensities are also unknown so propensities must be estimated from the dataset and *estimated* propensities are used to correct the bias. In general, this can be accomplished by any supervised learning technique.

In this chapter propensities are estimated by fitting the multinomial logistic regression model:

$$\ln(\Pr(A=a)) = \boldsymbol{\beta}_{0,a}^T\boldsymbol{X} - \ln Z \tag{3.9}$$

where $Z = \sum_{a \in \mathcal{A}} \exp\left(\boldsymbol{\beta}_{0,a}^T \boldsymbol{X}\right)$. The estimated propensities are

$$\widehat{p}_{0,j} \equiv \frac{\exp(\boldsymbol{\beta}_{0,A_j}^T \boldsymbol{X}_j)}{Z_j}$$

where we have written $Z_j = \sum_{a \in \mathcal{A}} \exp(\boldsymbol{\beta}_{0,a}^T \boldsymbol{X}_j)$. Write $\widehat{\boldsymbol{P}}_0 = [\widehat{p}_{0,j}]_{j=1}^n$ for the vector of estimated propensities

In principle, these estimated propensities could be used in place of known propensities and proceed exactly as we have done above. However, there are two problems with doing this. The first is that if the estimated propensities are very small (which might happen because the data was not completely representative of the true propensities), the variance of the estimate $\widehat{G}$ will be too large. The second is that the thresholds we have constructed when propensities are known may no longer be appropriate when propensities must be estimated.

To avoid the first problem, this chapter follows [Ion08] and modifies the estimated rewards by truncating the importance sampling weights. This leads to "truncated" estimated rewards as follows:

$$\begin{aligned}
\widehat{R}_m(a, x_i; \widehat{\boldsymbol{P}}_0) &= \frac{1}{N(x_i)} \sum_{j \in \mathcal{J}(x_i)} \min\left(\frac{\mathbb{I}(A_j = a)}{\widehat{p}_{0,j}}, m\right) R_j^{\text{obs}}, \\
\widehat{R}_m(a; \widehat{\boldsymbol{P}}_0) &= \frac{1}{n} \sum_{j=1}^{n} \min\left(\frac{\mathbb{I}(A_j = a)}{\widehat{p}_{0,j}}, m\right) R_j^{\text{obs}}.
\end{aligned}$$

Given these "truncated" estimated rewards, define a "truncated" estimator of relevance by

$$\widehat{G}_m(a, i; \widehat{\boldsymbol{P}}_0) = \sum_{x_i \in \mathcal{X}_i} \frac{N(x_i)}{n} l\left(\widehat{R}_m(a, x_i; \widehat{\boldsymbol{P}}_0) - \widehat{R}_m(a; \widehat{\boldsymbol{P}}_0)\right)$$

From this point on, our algorithm proceeds exactly as before, using the "truncated" estimator $\widehat{G}_m$ instead of $\widehat{G}$.

Note that $\widehat{R}_m(a, x_i; \widehat{\boldsymbol{P}}_0)$ and $\widehat{R}_m(a; \widehat{\boldsymbol{P}}_0)$ are not unbiased estimators of $\bar{r}(a, x_i)$ and $\bar{r}(a)$. The bias is due to using estimated truncated propensity scores which may deviate from true propensities. Let $\text{bias}(\widehat{R}_m(a; \widehat{\boldsymbol{P}}_0))$ denote the bias of $\widehat{R}_m(a; \widehat{\boldsymbol{P}}_0)$, which is given by

$$\text{bias}(\widehat{R}_m(a; \widehat{\boldsymbol{P}}_0)) = \bar{r}(a) - \mathbb{E}\left[\widehat{R}_m(a; \widehat{\boldsymbol{P}}_0)\right].$$

Algorithm 3 Training Phase of the Algorithm PONN-B

Input: $\lambda_1, \lambda_2, \lambda_3$, L_r, L_p, h_i^r, h_j^a

Step A: Estimate propensities using a logistic regression

Compute $\boldsymbol{\beta}_{0,a}$ for each a by training Logistic regression model from (3.9).

Set $\widehat{p}_{0,j} = \exp(\boldsymbol{\beta}_{0,A_j}^T \boldsymbol{X}_j)/Z_j$ with $Z_j = \sum_{a \in \mathcal{A}} \exp(\boldsymbol{\beta}_{0,a}^T \boldsymbol{X}_j)$.

Step B: Identify the relevant features

Compute $\widehat{R}(a, x_i; \widehat{\boldsymbol{P}}_0)$, $\widehat{R}(a; \widehat{\boldsymbol{P}}_0)$, $\bar{V}_n(a, i; \widehat{\boldsymbol{P}}_0)$, $\rho_{i,l}$ for each a, x_i, i, l.

Compute $\widehat{G}(a, i; \widehat{\boldsymbol{P}}_0)$ for each action-feature type pair.

Solve $\widehat{\mathcal{R}}(a)$ from (3.7).

Step C: Policy Optimization

while until convergence **do**

$\left(\boldsymbol{w}, \boldsymbol{W}_p^{(l)}\right) \leftarrow \text{Adam}\left(\mathcal{D}^{(n)}, \boldsymbol{w}, \boldsymbol{W}_p^{(l)}\right)$

end while

Output of Training Phase: Policy h^*, Features $\widehat{\boldsymbol{\mathcal{R}}}$

Algorithm 4 Execution Phase of the Algorithm PONN-B

Input: Instance with feature $\boldsymbol{X}$

Set $\hat{a}(\boldsymbol{X}) = \arg\max_{a \in \mathcal{A}} h^*(a|\boldsymbol{X})$

Output of Execution phase: Recommended action $\hat{a}(\boldsymbol{X})$

3.5 Pseudo-code for the Algorithm PONN-B

Below, the pseudo-code for the proposed algorithm which we call PONN-B (because it uses the PONN objective and Step B) is exactly as discussed above. The first three steps constitute the offline training phase; the fourth step is the online execution phase. Within the training phase the steps are: Step A: Input propensities (if they are known) or estimate them using a logistic regression (if they are not known). Step B: Construct estimates of relevance (truncated if propensities are estimated), construct thresholds (using given hyper-parameters) and identify the relevant features as those for which the estimated relevance is above the constructed thresholds. Step C: Use the Adam optimizer to train neural network parameters. In the execution phase: Input the features of the new instance, apply the opti-

mal policy to find a probability distribution over actions, and draw a random sample action from this distribution.

3.6 Extension: Relevant Feature Selection with Fine Gradations

The proposed algorithm might be inefficient when there are many features of a particular type – in particular, if one or more feature types are continuous. In that setting, the proposed algorithm can be modified to create bins that consist of *similar* feature values and treat all the values in a single bin identically. In order to conveniently formalize this problem, each feature space is assumed to be bounded, that is, $\mathcal{X}_i = [0, 1]$. In this case, the feature space is partitioned into subintervals (bins), view features in each bin as identical, and apply the proposed algorithm to the finite set of bins.[3] To offer a theoretical justification for this procedure,similar features is assumed to yield similar expected rewards. this is formalized as a Lipschitz condition.

Assumption 4. *There exists $L > 0$ such that for all $a \in \mathcal{A}$, all $i \in \mathcal{F}$ and all $x_i \in \mathcal{X}_i$, we have $|\bar{r}(a, x_i) - \bar{r}(a, \tilde{x}_i)| \leq L|x_i - \tilde{x}_i|$.*

(In the Multi-Armed Bandit literature [Sli14b, TS14] this assumption is commonly made and sometimes referred to as *similarity*.)

For convenience, each feature type X_i is partitioned into s equal subintervals (bins) of length $1/s$. If s is small, the number of bins is small so, given a finite data set, the number of instances that lie in each bin is relatively large; this is useful for estimation. However, when s is small the size $1/s$ of each bin is relatively large so the (true) variation of expected rewards in each bin is relatively large. Because we are free to choose the parameter s, this can balanced to trade-off implicit between choosing few large bins or choosing many small bins; a useful trade-off is achieved by taking $s = \lceil n^{1/3} \rceil$. So begin by fixing $s = \lceil n^{1/3} \rceil$ and partition each $\mathcal{X}_i = [0, 1]$ into s intervals of length $1/s$. Write $\mathcal{C}_i$ for the sets in the

[3]The binning procedure loses the ordering in the interval $[0, 1]$. If this ordering is in fact relevant to the feature, then the binning procedure loses some information that a different procedure might preserve. We leave this for future work.

partition of X_i and write c_i for a typical element of $\mathcal{C}_i$. For each sample j, let $c_{i,j}$ denote the set in which the feature $x_{i,j}$ belongs. Let $\mathcal{J}(c_i)$ be the set of indices for which $x_{i,j} \in c_i$; $\mathcal{J}(c_i) = \{j \in \{1, 2, \ldots, n\} : X_{i,j} \in c_i\}$. We define truncated IPS estimate as

$$\begin{aligned}
\bar{r}_m(a, c_i; \widehat{\boldsymbol{P}}_0) &= \mathbb{E}\left[U(a; \widehat{\boldsymbol{P}}_0) | X_i \in c_i\right] \\
&= \mathbb{E}\left[\min\left(\frac{\mathbb{I}(A=a)}{\widehat{p}_0(A|\boldsymbol{X})}, m\right) R^{\text{obs}} \middle| X_i \in c_i\right], \\
\widehat{R}_m(a, c_i; \widehat{\boldsymbol{P}}_0) &= \frac{1}{N(c_i)} \sum_{j \in \mathcal{J}(c_i)} \min\left(\frac{\mathbb{I}(A_j=a)}{\widehat{p}_{0,j}}, m\right) R_j^{\text{obs}},
\end{aligned}$$

where $N(c_i) = |\mathcal{J}(c_i)|$. In this case, define estimated information gain as

$$\widehat{G}_m(a, i) = \sum_{c_i \in \mathcal{C}_i} \frac{N(c_i)}{n} l\left(\widehat{R}_m(a, c_i; \widehat{\boldsymbol{P}}_0) - \widehat{R}_m(a; \widehat{\boldsymbol{P}}_0)\right).$$

Define the following sample mean and variance :

$$\widehat{U}(a, c_i; \widehat{\boldsymbol{P}}_0) = \widehat{R}_m(a, c_i; \widehat{\boldsymbol{P}}_0) = \frac{1}{N(x_i)} \sum_{j \in \mathcal{J}(c_i)} U_j(a; \widehat{\boldsymbol{P}}_0),$$

$$V_n(a, c_i; \widehat{\boldsymbol{P}}_0) = \frac{1}{n-1} \sum_{j \in \mathcal{J}(c_i)} (U_j(a, c_i; \widehat{\boldsymbol{P}}_0) - \widehat{U}(a, c_i; \widehat{\boldsymbol{P}}_0))^2.$$

Let $\bar{V}_n(a, i; \widehat{\boldsymbol{P}}_0) = \sum_{c_i \in \mathcal{C}_i} \frac{N(c_i) V_n(a, c_i; \widehat{\boldsymbol{P}}_0)}{n}$ denote the weighted average sample variance.

Theorem 9. *For every $n \geq 1$ and $\delta \in \left[0, \frac{1}{3}\right]$, if $s = \lceil n^{1/3} \rceil$, then with probability at least $1 - 3\delta$ we have, for all pairs $(a, i) \in (\mathcal{A}, \mathcal{D})$,*

$$\begin{aligned}
|\widehat{G}_m(a, i; \widehat{\boldsymbol{P}}_0) - g(a, i)| \leq\ & B\Bigg(\frac{\sqrt{4 \ln 3/\delta}}{n^{1/3}} \left(\sqrt{\bar{V}_n(a, i; \widehat{\boldsymbol{P}}_0)} + \sqrt{V_n(a; \widehat{\boldsymbol{P}}_0)}\right) + \frac{L}{n^{1/3}} \\
&+ \left|\text{bias}(\widehat{R}_m(a; \widehat{\boldsymbol{P}_0}))\right| + \mathbb{E}\left|\text{bias}(\widehat{R}_m(a, X_i; \widehat{\boldsymbol{P}_0}))\right|\Bigg) \\
&+ \frac{4mB \ln 3/\delta + \sqrt{2 \ln 1/\delta + \ln 2}}{n^{2/3}}.
\end{aligned}$$

There are two main differences between Theorem 8 and Theorem 9. The first is that the estimation error is decreasing as $n^{1/3}$ (Theorem 9) rather than as $n^{1/2}$ (Theorem 8). The second is that there is an additional error in Theorem 9 arising from the Lipschitz bound.

Theorem 9 suggests a different choice of thresholds, namely:

$$\tau(a, i) = \lambda_1 n^{-1/3} \sqrt{V_n(a, i; \widehat{\boldsymbol{P}}_0)} + \lambda_2 \left(\frac{1}{d-1}\right) \left(\sum_{l \in \mathcal{F} \setminus \{i\}} |\rho_{i,l}|\right).$$

3.7 Numerical Results

This section describes the performance of our algorithm on some real datasets. Note that it is difficult (perhaps impossible) to validate and test the algorithm on the basis of actual logged CMAB data unless the counterfactual action rewards for each instance are available – which would (almost) never be the case. One way to validate and test our algorithm is to use an alternative accepted procedure to infer counterfactuals and to test the prediction of our algorithm against this alternative accepted procedure. This is the route we follow in the experiments below.

3.7.1 Dataset

In this subsection, our algorithm is applied to the choice of recommendations of chemotherapy regimen for breast cancer patients. Our algorithm is evaluated on a dataset of 10,000 records of breast cancer patients participating in the National Surgical Adjuvant Breast and Bowel Project (NSABP) by [YDS16]. Each instance consists of the following information about the patient: age, menopausal, race, estrogen receptor, progesterone receptor, human epidermal growth factor receptor 2 (HER2NEU), tumor stage, tumor grade, Positive Axillary Lymph Node Count(PLNC), WHO score, surgery type, Prior Chemotherapy, prior radiotherapy and histology. The treatment is a choice among six chemotherapy regimes AC, ACT, AT, CAF, CEF, CMF. The outcomes for these regimens were derived based on 32 references from PubMed Clinical Queries. The rewards for these regimens were derived based on 32 references from PubMed Clinical Queries; this is a medically accepted procedure. The details are given in [YDS16].

3.7.2 Comparisons

We compare the performance of our algorithm (PONN-B) with

- **PONN** is PONN-B but *without* Step B (feature selection).

- **POEM** is the standard POEM algorithm [SJ15a].
- **POEM-B** applies Step B of our algorithm, followed by the POEM algorithm.
- **POEM-L1** is the POEM algorithm with the addition of L_1 regularization.
- **Multilayer Perceptron with L_1 regularization (MLP-L1)** is the MLP algorithm on concatenated input $(\boldsymbol{X}, A)$ with L_1 regularization.
- **Logistic Regression with L_1 regularization (LR-L1)** is the separate LR algorithm on input $\boldsymbol{X}$ on each action a with L_1 regularization.
- **Logging** is the logging policy performance.

(In all cases, the objective is optimized with the Adam Optimizer.)

3.7.2.1 Simulation Setup

Artificially biased dataset is generated by the following logistic model. First weights for each label are drawn from an multivariate Gaussian distribution, that is $\theta_{0,y} \sim \mathcal{N}(0, \kappa I)$. Then, the logistic model is used to generate an artificially biased logged off-policy dataset $\mathcal{D}^n = \left(\boldsymbol{X}_j, A_j, R_j^{\text{obs}}\right)_{j=1}^n$ by first drawing an action $A_j \sim p_0(\cdot|\boldsymbol{X}_j)$, then setting the observed reward as $R_j^{\text{obs}} \equiv R_j(A_j)$. (We use $\kappa = 0.25$.) This bandit generation process makes the learning very challenging as the generated off-policy dataset has less number of observed labels.

The dataset is randomly divided into 70% training and 30% testing sets. All algorithms are trained for various parameter sets on the training set, the hyper parameters are selected based on the validation set and performance is tested on the testing set. Our algorithm are trained with $L_r = 2$ representation layers, and $L_p = 2$ policy layers with 50 hidden units for representation layers and 100 hidden units (sigmoid activation) with policy layers. All algorithms are implemented/trained in a Tensorflow environment using Adam Optimizer.

For j-th instance in testing data, let h_g^* denote the optimized policy of algorithm g. Let $\mathcal{J}_{test}$ denote the instances in testing set and $N_{test} = |\mathcal{J}_{test}|$ denote number of instances in

Metric	Accuracy	Improvement
PONN-B	$\mathbf{74.12\% \pm 1.25\%}$	-
PONN	$62.81\% \pm 1.85\%$	30.41%
POEM-B	$55.39\% \pm 0.36\%$	41.98%
POEM	$52.78\% \pm 0.50\%$	45.19%
POEM-L1	$52.72\% \pm 0.55\%$	45.26%
MLP-L1	$61.47\% \pm 0.50\%$	55.05%
LR-L1	$51.96\% \pm 0.43\%$	46.12%
Logging	$18.20\% + 1.30\%$	68.36%

Table 3.2: Performance in the Breast Cancer Experiment

testing dataset. Define (absolute) accuracy of an algorithm g as

$$\mathrm{Acc}(g) = \frac{1}{N_{test}} \sum_{j \in \mathcal{J}_{test}} \sum_{a \in \mathcal{A}} h_g^*(a|\boldsymbol{X}_j) R_j(a).$$

The parameters are selected from the sets $\lambda_1^* \in [0.005, 0.1], \lambda_2^* \in [0, 0.01]$ and $\lambda_3^* \in [0.0001, 0.1]$ so that loss given in (3.8) estimated from the samples in the validation set is minimized. In the testing dataset, the full dataset is used to compute the accuracy of each algorithm.

In each case, the average of the iterations are reported with 95% confidence intervals over 25 iterations.

3.7.2.2 Results

Table 3.2 describes absolute accuracy and the Improvement Scores of the our algorithm. The proposed algorithm achieves significant Improvement Scores with respect to all benchmarks. There are two main reasons for these improvements. The first is that using Step B (feature selection) reduces over-fitting; this can be seen by the improvement of PONN-B over PONN and by the fact that PONN-B improves more over POEM (which does not use Step B) than over POEM-B (which does use feature selection). The second is that PONN-B allows for

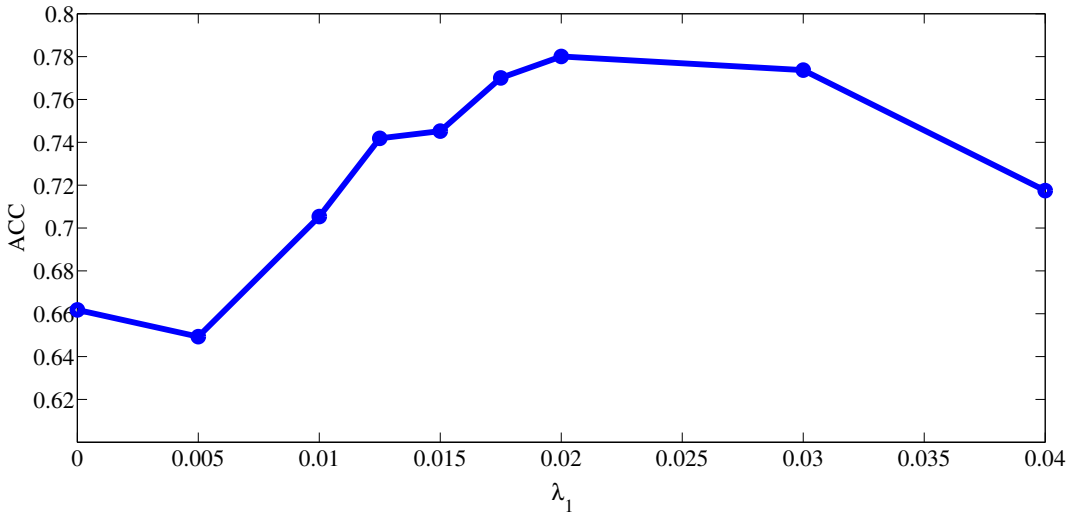


Figure 3.2: Effect of the hyperparameter on the accuracy of our algorithm

non-linear policies, which reduces model misspecification.

Note that proposed action-dependent relevance discovery is also important for interpretability. The selected relevant features given by our algorithm with $\lambda_1 = 0.03$ is as follows: age, tumor stage, tumor grade for AC treatment action, age, tumor grade, lymph node status for ACT treatment action, menopausal status and surgery type for CAF treatment action, age and estrogen receptor for CEF treatment action and estrogen receptor and progesterone receptor for CMF treatment action.

Figure 3.2 shows the accuracy of our algorithm for different choices of the hyper parameter λ_1. As expected – and seen in Figure 3.2 – if λ_1 is too small then there is overfitting; if it is too large then a lot of relevant features are discarded.

3.8 Appendix

Here the proofs of Theorems 8 and 9 are provided. It is convenient to begin by recording some technical lemmas; the first two are in the literature; proofs for the other two are given here.

Lemma 5 (Theorem 1, [AMS09]). *Let $X_1, X_2, \ldots, X_n$ be i.i.d. random variables taking*

their values in $[0, b]$. *Let* $\mu = \mathbb{E}[X_1]$ *be their common expected value. Consider the empirical sample mean* $\bar{X}_n$ *and variance* V_n *defined respectively by*

$$\bar{X}_n = \frac{\sum_{i=1}^n X_i}{n} \text{ and } V_n = \frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{n}. \tag{3.10}$$

Then, for any $n \in \mathbb{N}$ *and* $\delta \in (0, 1)$, *with probability at least* $1 - \delta$,

$$|\bar{X}_n - \mu| \leq \sqrt{\frac{2V_n \log 3/\delta}{n}} + \frac{3b \log 3/\delta}{n}. \tag{3.11}$$

For two probability distributions $\boldsymbol{P}$ and $\boldsymbol{Q}$ on a finite set $\mathcal{A} = \{1, 2, \ldots, a\}$, let

$$\|\boldsymbol{P} - \boldsymbol{Q}\|_1 = \sum_{i=1}^{a} |\boldsymbol{P}(i) - \boldsymbol{Q}(i)| \tag{3.12}$$

denote the L_1 distance between $\boldsymbol{P}$ and $\boldsymbol{Q}$.

Lemma 6. *[WOS03] Let* $\mathcal{A} = \{1, 2, \ldots, a\}$. *Fix a probability distribution* $\boldsymbol{P}$ *on* $\mathcal{A}$ *and draw* n *independent samples* $\boldsymbol{X}^n = X_1, X_2, \ldots, X_n$ *from* $\mathcal{A}$ *according to the distribution* $\boldsymbol{P}$. *Let* $\widehat{\boldsymbol{P}}$ *be the empirical distribution of* $\boldsymbol{X}^n$. *Then, for all* $\epsilon > 0$,

$$\Pr(\|\boldsymbol{P} - \widehat{\boldsymbol{P}}\|_1 \geq \epsilon) \leq (2^a - 2)e^{-\epsilon^2 n/2}. \tag{3.13}$$

The next two lemmas are auxiliary results used in the proof of Theorem 9.

Lemma 7. *Let* $\boldsymbol{P}_0 = [p_0(a|\boldsymbol{x})]$ *be the actual propensities and* $\widehat{\boldsymbol{P}}_0 = [\widehat{p}_0(a|\boldsymbol{x})]$ *be the estimated propensities. Assume that* $\widehat{p}_0(a|\boldsymbol{x}) > 0$ *for all* $a, \boldsymbol{x}$. *The bias of the truncated IS estimator with propensities* $\widehat{\boldsymbol{P}}_0$ *is:*

$$\begin{aligned} \text{bias}(\widehat{R}_m(a; \widehat{\boldsymbol{P}}_0)) \;\; = \;\; & \sum_{j=1}^{n} \mathbb{E}\Bigg[\frac{\bar{r}(a, \boldsymbol{X}_j)}{n} \Bigg(\left(1 - \frac{p_{0,j}}{\widehat{p}_{0,j}}\right) \mathbb{I}\left(\widehat{p}_{0,j} \geq m^{-1}\right) \\ & + \; (1 - p_{0,j} m)\, \mathbb{I}\left(\widehat{p}_{0,j} \leq m^{-1}\right) \Bigg) \Bigg]. \end{aligned}$$

Proof of Lemma 7 The proof is similar to [JS16]. Then,

$$\begin{aligned}
\bar{r}(a) &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{\boldsymbol{X}_j\sim\Pr(\mathcal{X})}\bar{r}(a,\boldsymbol{X}_j),\\
\mathbb{E}(\widehat{R}_m(a;\widehat{\boldsymbol{P}}_0)) &= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{(\boldsymbol{X}_j,A_j,R_j)\sim p_0}\left[\min\left(\frac{\mathbb{I}(A_j=a)}{\widehat{p}_0(A_j|\boldsymbol{X}_j)},m\right)R_j\right]\\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}_{(\boldsymbol{X}_j,A_j)\sim p_0}\left[\min\left(\frac{\mathbb{I}(A_j=a)}{\widehat{p}_0(a|\boldsymbol{X}_j)},m\right)\bar{r}(a,\boldsymbol{X}_j)\right]\\
&= \sum_{j=1}^{n}\mathbb{E}_{\boldsymbol{X}_j\sim\Pr(\mathcal{X})}\left[\frac{\bar{r}(a,\boldsymbol{X}_j)}{n}\min\left(\frac{1}{\widehat{p}_0(a|\boldsymbol{X}_j)},m\right)p_0(a|\boldsymbol{X}_j)\right].
\end{aligned}$$

It follows that

$$\mathrm{bias}(\widehat{R}_m(a;\boldsymbol{P})) = \sum_{j=1}^{n}\mathbb{E}_{\boldsymbol{X}_j\sim\Pr(\mathcal{X})}\left[\frac{\bar{r}(a,\boldsymbol{X}_j)}{n}\left(1-\min\left(\frac{1}{\widehat{p}_0(a|\boldsymbol{X}_j)},m\right)p_0(a|\boldsymbol{X}_j)\right)\right]. \quad (3.14)$$

Dividing (3.14) into the case for which $\widehat{p}_0(a|\boldsymbol{X}_j)\geq m^{-1}$ and the case for which $\widehat{p}_0(a|\boldsymbol{X}_j)< m^{-1}$ and then combining the results yields the desired conclusion.

To state Lemma 8, first define the expected relevance gain with truncated IPS reward using propensities $\widehat{\boldsymbol{P}}_0$ to be

$$g_m(a,i;\widehat{\boldsymbol{P}}_0) = \mathbb{E}\left[\left|\bar{r}_m(a,X_i;\widehat{\boldsymbol{P}}_0)-\bar{r}_m(a;\widehat{\boldsymbol{P}}_0)\right|\right]$$

where

$$\begin{aligned}
\bar{r}_m(a;\widehat{\boldsymbol{P}}_0) &= \mathbb{E}(\widehat{R}_m(a;\widehat{\boldsymbol{P}}_0))\\
&= \mathbb{E}_{(\boldsymbol{X},A,R)\sim p_0}\left[\min\left(\frac{\mathbb{I}(A=a)}{p_0(A|\boldsymbol{X})},m\right)R\right],\\
\bar{r}_m(a,x_i;\widehat{\boldsymbol{P}}_0) &= \mathbb{E}(\widehat{R}_m(a,x_i;\widehat{\boldsymbol{P}}_0))\\
&= \mathbb{E}_{(\boldsymbol{X},A,R)\sim p_0}\left[\min\left(\frac{\mathbb{I}(A=a)}{p_0(A|\boldsymbol{X})},m\right)R\middle|X_i=x_i\right].
\end{aligned}$$

Lemma 8. *The following holds:*

$$|g_m(a,i;\widehat{\boldsymbol{P}}_0)-g(a,i)| \leq B\left(\mathbb{E}\left[\left|\mathrm{bias}(\widehat{R}_m(a,X_i;\widehat{\boldsymbol{P}}_0))\right|\right]+\left|\mathrm{bias}(\widehat{R}_m(a;\widehat{\boldsymbol{P}}_0))\right|\right).$$

Proof of Lemma 8 This follows immediately by iterated expectations:

$$\begin{aligned}
&\left|\mathbb{E}\Big(\ell\left(\mathbb{E}(\widehat{R}_m(a,X_i;\widehat{\boldsymbol{P}}_0))-\mathbb{E}(\widehat{R}_m(a;\widehat{\boldsymbol{P}}_0))\right)-\ell\left(\bar{r}(a,x_i)-\bar{r}(a)\right)\Big)\right|\\
&\quad\leq\ B\mathbb{E}\Big(\left|\mathbb{E}(\widehat{R}_m(a,X_i;\widehat{\boldsymbol{P}}_0))-\bar{r}(a,X_i)\right|\Big)+\ B|\mathbb{E}(\widehat{R}_m(a;\widehat{\boldsymbol{P}}_0))-\bar{r}(a)|. \quad (3.15)
\end{aligned}$$

Proof of Theorem 8 Recall that the true relevance metric is $g(a,i) = \mathbb{E}\left[|\bar{r}(a,x_i) - \bar{r}(a)|\right] = \sum_{x_i \in \mathcal{X}_i} \Pr(X_i = x_i) l(\bar{r}(a,x_i) - \bar{r}(a))$. For any action $a \in \mathcal{A}$ and $x_i \in \mathcal{X}_i$, the error between the estimated relevance metric and the relevance metric can be bounded as

$$\begin{aligned}
|\widehat{G}(a,i;\boldsymbol{P}_0) - g(a,i)| &= \Bigg| \sum_{x_i \in \mathcal{X}_i} \frac{N(x_i)}{n} \ell\left(\widehat{R}(a,x_i;\boldsymbol{P}_0) - \widehat{R}(a;\boldsymbol{P}_0)\right) \\
&\quad - \sum_{x_i \in \mathcal{X}_i} \frac{N(x_i)}{n} \ell\left(\bar{r}(a,x_i) - \bar{r}(a)\right) \\
&\quad + \sum_{x_i \in \mathcal{X}_i} \frac{N(x_i)}{n} \ell\left(\bar{r}(a,x_i) - \bar{r}(a)\right) \\
&\quad - \sum_{x_i \in \mathcal{X}_i} \Pr(X_i = x_i) \ell\left(\bar{r}(a,x_i) - \bar{r}(a)\right) \Bigg| \\
&\leq \sum_{x_i \in \mathcal{X}_i} \frac{N(x_i)}{n} \left(\ell\left(\widehat{R}(a,x_i;\boldsymbol{P}_0) - \widehat{R}(a;\boldsymbol{P}_0)\right) - \ell\left(\bar{r}(a,x_i) - \bar{r}(a)\right)\right) \\
&\quad + \sum_{x_i \in \mathcal{X}_i} \left(\frac{N(x_i)}{n} - \Pr(X_i = x_i)\right) \ell\left(\bar{r}(a,x_i) - \bar{r}(a)\right) \\
&\leq B \sum_{x_i \in \mathcal{X}_i} \frac{N(x_i)}{n} \left|\widehat{R}(a,x_i;\boldsymbol{P}_0) - \bar{r}(a,x_i)\right| + B \left|\widehat{R}(a;\boldsymbol{P}_0) - \bar{r}(a)\right| \\
&\quad + \sum_{x_i \in \mathcal{X}_i} \left|\frac{N(x_i)}{n} - \Pr(X_i = x_i)\right|.
\end{aligned}$$

Each term is bounded separately. Applying Lemma 6, with probability at least $1 - \delta$,

$$\begin{aligned}
\sum_{x_i \in \mathcal{X}_i} \left|\Pr(X_i = x_i) - \frac{N(x_i)}{n}\right| &\leq \sqrt{\frac{2 \ln 2^{b_i}/\delta}{n}} \\
&= \sqrt{\frac{2\left(b_i \ln 2 + \ln 1/\delta\right)}{n}}. \qquad (3.16)
\end{aligned}$$

Using Lemma 5 see that, with probability at least $1 - \delta$,

$$\begin{aligned}
&\sum_{x_i \in \mathcal{X}_i} \frac{N(a,x_i)}{n} \left|\widehat{R}(a,x_i;\boldsymbol{P}_0) - \bar{r}(a,x_i)\right| \\
&\quad \leq \sum_{x_i \in \mathcal{X}_i} \frac{N(a,x_i)}{n} \left(\sqrt{\frac{2V_n(a,x_i;\boldsymbol{P}_0) \ln 3/\delta}{N(a,x_i)}} + \frac{3M \ln 3/\delta}{N(a,x_i)}\right) \\
&\quad \leq \sqrt{\frac{2b_i V_n(a,x_i;\boldsymbol{P}_0) \ln 3/\delta}{n}} + \frac{3Mb_i \ln 3/\delta}{n}, \qquad (3.17)
\end{aligned}$$

where the the second inequality follows from an application of Jensen's inequality. Similarly, using Lemma 5, see that with probability at least $1-\delta$,

$$\left|\widehat{R}(a;\boldsymbol{P}_0) - \bar{r}(a)\right| \leq \sqrt{\frac{2V_n(a;\boldsymbol{P}_0)\ln 3/\delta}{n}} + \frac{3M\ln 3/\delta}{n}. \tag{3.18}$$

The desired result now follows by combining (3.16, 3.17 and 3.18).

Proof of Theorem 8 Let

$$\tilde{g}_m(a,i) = \sum_{c_i \in \mathcal{C}_{i,n}} \Pr(X_i \in c_i)\ell(\bar{r}_m(a,c_i) - \bar{r}_m(a)).$$

Then, decompose the error into

$$\begin{aligned}
|\widehat{G}_m(a,i;\widehat{\boldsymbol{P}}_0) - g(a,i)| &\leq |\widehat{G}_m(a,i;\widehat{\boldsymbol{P}}_0) - g_m(a,i;\widehat{\boldsymbol{P}}_0)| + |g_m(a,i;\widehat{\boldsymbol{P}}_0) - g(a,i)| \\
&\leq |\widehat{G}_m(a,i;\widehat{\boldsymbol{P}}_0) - \tilde{g}_m(a,i;\widehat{\boldsymbol{P}}_0)| \\
&\quad + |\tilde{g}_m(a,i;\widehat{\boldsymbol{P}}_0) - g_m(a,i;\widehat{\boldsymbol{P}}_0)| \\
&\quad + |g_m(a,i;\widehat{\boldsymbol{P}}_0) - g(a,i)|.
\end{aligned} \tag{3.19}$$

The first term (3.19) can be bounded by Theorem 8 by setting $s_n = \lceil n^{1/3} \rceil \leq n^{1/3}+1$, i.e.,

$$\begin{aligned}
|\widehat{G}_m(a,i;\widehat{\boldsymbol{P}}_0) - \tilde{g}_m(a,i;\widehat{\boldsymbol{P}}_0)| &\leq \frac{\sqrt{4B^2\ln 3/\delta}}{n^{1/3}}\left(\sqrt{\bar{V}_n(a,i;\widehat{\boldsymbol{P}}_0)} + \sqrt{V_n(a;\widehat{\boldsymbol{P}}_0)}\right) \\
&\quad + \frac{4mB\ln 3/\delta + \sqrt{2\ln 1/\delta + \ln 2}}{n^{2/3}}.
\end{aligned}$$

The third term in (3.19) is the bias of the estimation due to estimated propensity scores and truncation, i.e.,

$$|g_m(a,i;\widehat{\boldsymbol{P}}_0) - g(a,i)| \leq B\left(\mathbb{E}\left[\left|\text{bias}(\widehat{R}_m(a,X_i);\widehat{\boldsymbol{P}}_0)\right|\right] + \left|\text{bias}(\widehat{R}_m(a);\widehat{\boldsymbol{P}}_0)\right|\right).$$

The second term in (3.19) can be bounded

$$\begin{aligned}
g_m(a,i;\widehat{\boldsymbol{P}}_0) &= \mathbb{E}\left[\ell(\bar{r}_m(a,X_i;\widehat{\boldsymbol{P}}_0) - \bar{r}_m(a;\widehat{\boldsymbol{P}}_0))\right] \\
&= \mathbb{E}\left[\ell(\bar{r}_m(a,X_i;\widehat{\boldsymbol{P}}_0) - \bar{r}_m(a,c_i;\widehat{\boldsymbol{P}}_0) + \bar{r}_m(a,c_i;\widehat{\boldsymbol{P}}_0) - \bar{r}_m(a;\widehat{\boldsymbol{P}}_0))\right] \\
&\leq \mathbb{E}\left[\ell\left(\frac{L}{n^{1/3}} + \bar{r}_m(a,c_i;\widehat{\boldsymbol{P}}_0) - \bar{r}_m(a;\widehat{\boldsymbol{P}}_0)\right)\right] \\
&\leq \frac{LB}{n^{1/3}} + \mathbb{E}\left[\ell(\bar{r}_m(a,c_i;\widehat{\boldsymbol{P}}_0) - \bar{r}_m(a;\widehat{\boldsymbol{P}}_0))\right].
\end{aligned}$$

where the first inequality follows from Assumption 4 and the second inequality follows from smoothness assumption on the loss function $l(\cdot)$, i.e.,

$$l\left(\frac{L}{n^{1/3}}+\bar{r}_m(a,c_i;\widehat{\boldsymbol{P}}_0)-\bar{r}_m(a;\widehat{\boldsymbol{P}}_0)\right)-l\left(\bar{r}_m(a,c_i;\widehat{\boldsymbol{P}}_0)-\bar{r}_m(a;\widehat{\boldsymbol{P}}_0)\right)\leq\frac{LB}{n^{1/3}}.$$

CHAPTER 4

Counterfactual Policy Optimization Using Domain-Adversarial Neural Networks

4.1 Introduction

The choice of a particular policy or plan of action involves consideration of the costs and benefits of the policy/plan under consideration and also of alternative policies/plans that might be undertaken. Examples abound; to mention just a few: Which course of treatment will lead to the most rapid recovery? Which mode of advertisement will lead to the most orders? Which investment strategy will lead to the greatest returns? Obtaining information about the costs and benefits of alternative plans that might have been undertaken is a *counterfactual exercise*. One possible way to estimate the counterfactual information is by conducting controlled experiments. However, controlled experiments are expensive, involve small samples, and are frequently not available. It is therefore important to make decisions entirely on the basis of observational data in which the actions/decisions taken in the data have been selected by an existing *logging* policy. Because the existing logging policy creates a selection bias, learning from observational studies is a challenging problem. This chapter presents theoretical bounds on estimation errors for the evaluation of a new policy from observational data and a principled algorithm to learn the optimal policy. The proposed methods and algorithms are widely applicable (perhaps with some modifications) to an enormous range of settings in healthcare applications. In the medical context, features are the information included in electronic health records, actions are choices of different treatments, and outcomes are the success of treatment.

Theoretical results developed in this chapter show that *true policy outcome* is at least

Literature	Propensities known	Objective	Actions	Solution
[SJS17]	no	ITE estimation	2	Balancing representations
[AS17]	no	ITE estimation	2	Risk based empirical Bayes
[BL09]	yes	policy optimization	> 2	Rejection sampling
[SJ15b, SJ15c]	yes	policy optimization	> 2	IPS reweighing
Ours	no	policy optimization	> 2	Balancing representations

Table 4.1: Comparison with the related literature

as good as the *policy outcome estimated from the observational data* minus the product of the number of actions with the $\mathcal{H}$-divergence between the observational and randomized data. These theoretical bounds are different than ones derived in [SJ15b] because ours do not require the propensity scores to be known. These theoretical results are used to develop algorithm to learn balanced representations for each instance such that they are indistinguishable between the randomized and observational distribution and also predictive of the decision problem at hand. The developed algorithm is evaluated on the breast cancer dataset.

4.2 Related Work

Roughly speaking, work on counterfactual learning from observational data falls into two categories: estimation of Individualized Treatment Effect (ITE) [JSS16, SJS17, AS17] and Policy Optimization [SJ15b, SJ15c]. The work on ITE's aims to estimate the expected difference between outcomes for *treated* and *control* patients, given the feature vector; this work focuses on settings with only two actions (treat/don't treat) - and notes that the approaches derived do not generalize well to settings with more than two actions. The work on policy optimization aims to find a policy that maximizes the expected outcome (minimizes the risk). The policy optimization objective is somewhat easier than ITE objective in the sense that one can turn the ITE to action recommendations but not the other way around. In many healthcare applications, there are much more than 2 actions; one is more interested in learning a good action rather than learning outcomes of each action for each instance.

The work on ITE estimation that is most closely related to this chapter focuses on

learning balanced representations [JSS16, SJS17]. These papers develop neural network algorithms to minimize the mean squared error between predictions and actual outcomes in the observational data and also the discrepancy between the representations of the factual and counterfactual data. As these papers note, there is no principled approach to extend them to more than two treatments. Other recent works in ITE estimation include tree-based methods [Hil11, AI15, WA15] and Gaussian processes [AS17]. The last is perhaps the most successful, but the computational complexity is $O(n^3)$ (where n is the number of instances) so it is not easy to apply to large observational studies.

In the policy optimization literature, the work most closely related to this chapter is [SJ15b, SJ15c] where they develop the Counterfactual Risk Minimization (CRM) principle. The objective of the CRM principle is to minimize both the estimated mean and variance of the Inverse Propensity Score (IPS) instances; to do so the authors propose the POEM algorithm. The algorithm developed in this chapter differs from POEM in several ways: (i) POEM minimizes an objective over the class of linear policies; ours allow for arbitrary policies, (ii) POEM requires the propensity scores to be available in the data; our algorithm addresses the selection bias without using propensity scores, (iii) POEM addresses selection bias by re-weighting each instance with the inverse propensities; our algorithm addresses the selection bias by learning representations. Another related paper on policy optimization is [BL09] which requires the propensity scores to be known and addresses the selection bias via rejection sampling. (For a more detailed comparison see Table 4.1.)

The off-policy evaluation methods include IPS estimator [RR83, SLL10], self normalizing estimator [SJ15c], direct estimation, doubly robust estimator [DLL11, JL16] and matching based methods [HR06a]. The IPS and self-normalizing estimators address the selection bias by re-weighting each instance by their inverse propensities. The doubly robust estimation techniques combine the direct and IPS methods and generate more robust counterfactual estimates. Propensity Score Matching (PSM) replaces the missing counterfactual outcomes of the instance by the outcome of an instance with the closest propensity score.

Theoretical bounds developed in this chapter have strong connection with the domain adaptation bounds given in [BBC07, BCK08]. In particular, this chapter shows that the

expected policy outcome is bounded below by the estimate of the policy outcome from the observational data minus the product of the number of actions with the $\mathcal{H}$-divergence between the observational and randomized data. The algorithm developed in this chapter is based on domain adaptation as in [GUA16]. Other domain adaptation techniques include [ZSM13, Dau09].

4.3 Problem Setup

4.3.1 Observational Data

Denote by $\mathcal{A}$ the set of k actions, by $\mathcal{X}$ the s-dimensional space of features and by $\mathcal{Y} \subseteq R$ the space of outcomes. We assume that an outcome can be identified with a real number and normalize so that outcomes lie in the interval $[0, 1]$. In some cases, the outcome will be either 1 or 0 (success or failure); in other cases the outcome may be interpreted as the probability of success or failure. The potential outcome model described in the Rubin-Neyman causal model [Rub05] is followed; that is, for each instance $x \in \mathcal{X}$, there are k-potential outcomes: $Y^{(0)}, Y^{(1)}, \ldots, Y^{(k-1)} \in \mathcal{Y}$, corresponding to the k different actions. The fundamental problem in this setting is that only the outcome of the action *actually performed* is recorded in the data: $Y = Y^T$. (This is called *bandit feedback* in the machine learning literature [SJ15b].) In this chapter, we focus on the setting in which the action assignment is *not* independent of the feature vector, i.e., $A \not\perp\!\!\!\perp X$; that is, action assignments are *not random*. This dependence is modeled by the conditional distribution $\gamma(a, x) = P(A = a|X = x)$, also known as the *propensity score*.

In this chapter, the following common assumptions are made:

- **Unconfoundedness:** Potential outcomes $(Y^{(0)}, Y^{(1)}, \ldots, Y^{(k-1)})$ are independent of the action assignment given the features, that is $(Y^{(0)}, Y^{(1)}, \ldots, Y^{(k-1)}) \perp\!\!\!\perp A|X$.
- **Overlap:** For each instance $x \in \mathcal{X}$ and each action $a \in \mathcal{A}$, there is a non-zero probability that a patient with feature x received the action a: $0 < \gamma(a, x) < 1$ for all a, x.

These assumptions are sufficient to identify the optimal policy from the data [IW09, Pea17].

The dataset is

$$\mathcal{D}^n = \{(x_i, a_i, y_i)\}_{i=1}^n$$

where each instance i is generated by the following stochastic process:

- Each feature-action pair is drawn according to a fixed but unknown distribution $\mathcal{D}_S$, i.e, $(x_i, a_i) \sim \mathcal{D}_S$.
- Potential outcomes conditional on features are drawn with respect to a distribution $\mathcal{P}$; that is, $(Y_i^{(0)}, Y_i^{(1)}, \ldots, Y_i^{(k-1)}) \sim \mathcal{P}(\cdot|X = x_i, A = a_i)$.
- Only the outcome of the action actually performed is recorded in the data, that is, $y_i = Y_i^{(a_i)}$.

Denote the marginal distribution on the features by $\mathcal{D}$; i.e., $\mathcal{D}(x) = \sum_{a \in \mathcal{A}} \mathcal{D}_S(x, a)$.

4.3.2 Definition of Policy Outcome

A *policy* is a mapping h from features to actions. In this chapter, the goal is learning a policy h that maximizes the policy outcome, defined as:

$$V(h) = \mathbb{E}_{x \sim \mathcal{D}} \left[\mathbb{E} \left[Y^{(h(X))} | X = x \right] \right].$$

We denote by $m_a(x) = \mathbb{E}\left[Y^{(a)}|X = x\right]$ the expected outcome of action a on an instance with feature x. Based on these definitions, the policy outcome of h can be re-written as $V(h) = \mathbb{E}_{x \sim \mathcal{D}} \left[m_{h(x)}(x) \right]$. Estimating $V(h)$ from the data is a challenging task because the counterfactuals are missing and there is a selection bias.

4.4 Counterfactual Estimation Bounds

In this section, a criterion that we will use to learn a policy h^* the maximizes the outcome is provided. The selection bias in our dataset is handled by mapping the features to representations are relevant to policy outcomes and are less biased. Let $\Phi : \mathcal{X} \to \mathcal{Z}$ denote

a representation function which maps the features to representations. The representation function induces a distribution over representations $\mathcal{Z}$ (denoted by $\mathcal{D}^\Phi$) and m_a as follows:

$$\begin{aligned}\mathbb{P}_{\mathcal{D}^\Phi}(\mathcal{B}) &= \mathbb{P}_{\mathcal{D}}(\Phi^{-1}(\mathcal{B})),\\ m_a^\Phi(z) &= \mathbb{E}_{x\sim\mathcal{D}}[m_a(x)|\Phi(x)=z],\end{aligned}$$

for any $\mathcal{B}\subset\mathcal{Z}$ such that $\Phi^{-1}(\mathcal{B})$ is $\mathcal{D}$-measurable. That is, the probability of of an event $\mathcal{B}$ according to $\mathcal{D}^\Phi$ is the probability of the inverse image of the event $\mathcal{B}$ according to $\mathcal{D}$. The learning setting is defined by our choice of the representation function and hypothesis class $\mathcal{H}=\{h:\mathcal{Z}\to\mathcal{A}\}$ of (deterministic) policies.

Recall that $\mathcal{D}_S$ is the source distribution that generated feature-action samples in our observational data. Define the target distribution $\mathcal{D}_T$ by $\mathcal{D}_T(x,a)=(1/k)\mathcal{D}(x)$. Note that $\mathcal{D}_S$ represents an observational study in which the actions are not randomized, while $\mathcal{D}_T$ represents a clinical study in which actions *are* randomized. Let $\mathcal{D}_S^\Phi$ and $\mathcal{D}_T^\Phi$ denote the source and target distributions induced by the representation function Φ over the space $\mathcal{Z}\times\mathcal{A}$, respectively. Let $\mathcal{D}^\Phi$ denote the marginal distribution over the representations and write $V^\Phi(h)$ for the induced policy outcome of h, that is, $V^\Phi(h)=\mathbb{E}_{z\sim\mathcal{D}^\Phi}\left[m^\Phi_{h(z)}(z)\right]$.

For the remainder of the theoretical analysis, suppose that the representation function Φ is fixed. The missing counterfactual outcomes can be addressed by importance sampling. Let $V_S^\Phi(h)$ and $V_T^\Phi(h)$ denote the expected policy outcome with respect to distributions $\mathcal{D}_S$ and $\mathcal{D}_T$, respectively. They are given by

$$\begin{aligned}V_S^\Phi(h) &= \mathbb{E}_{(z,a)\sim\mathcal{D}_S^\Phi}\left[\frac{m_a^\Phi(z)1(h(z)=a)}{1/k}\right],\\ V_T^\Phi(h) &= \mathbb{E}_{(z,a)\sim\mathcal{D}_T^\Phi}\left[\frac{m_a^\Phi(z)1(h(z)=a)}{1/k}\right].\end{aligned}$$

where $1(\cdot)$ is an indicator function if the statement is true and 0 otherwise. The next proposition shows that value function of a policy h evaluated on the target distribution is unbiased. We note that though only the value function evaluated on observational data, $V_S(h)$ can be estimated.

Proposition 2. *Let Φ be a fixed representation function. Then: $V_T^\Phi(h)=V^\Phi(h)$.*

Proof. It follows that

$$\begin{aligned} V_T^{\Phi}(h) &= \mathbb{E}_{z\sim\mathcal{D}^{\Phi}}\left[\sum_{a\in\mathcal{A}} 1/k \frac{m_a^{\Phi}(z)1(h(z)=a)}{1/k}\right] \\ &= \mathbb{E}_{z\sim\mathcal{D}^{\Phi}}\left[m_{h(z)}^{\Phi}(z)\right] = V^{\Phi}(h). \end{aligned}$$

□

As noted above a Monte-Carlo estimator for $V_T^{\Phi}(h)$ is not possible since the observational dataset does not have samples from the target distribution - only has samples from the source distribution. However, domain adaptation theory can be used to bound the difference between $V_S^{\Phi}(h)$ and $V_T^{\Phi}(h)$ in terms of $\mathcal{H}$-divergence. In order to do that, let us introduce a distance metric between distributions. For any policy $h \in \mathcal{H}$, let $\mathcal{I}_h$ denote the characteristic set that contains all representation-action pairs that is mapped to label a under function h, i.e., $\mathcal{I}_h = \{(z, a) : h(z) = a\}$.

Definition 4. *Suppose $\mathcal{D}$, $\mathcal{D}'$ be probability distributions over $\mathcal{Z} \times \mathcal{A}$ such that every characteristic set $\mathcal{I}_h$ of $h \in \mathcal{H}$ is measurable with respect to both distributions. Then, the $\mathcal{H}$-divergence between distributions $\mathcal{D}$ and $\mathcal{D}'$ is*

$$d_{\mathcal{H}}(\mathcal{D}, \mathcal{D}') = \sup_{h\in\mathcal{H}} \left| \mathbb{P}_{(z,a)\sim\mathcal{D}}(\mathcal{I}_h) - \mathbb{P}_{(z,a)\sim\mathcal{D}'}(\mathcal{I}_h) \right|.$$

The $\mathcal{H}$-divergence measures the difference between the behavior of policies in $\mathcal{H}$ when examples are drawn from $\mathcal{D}$, $\mathcal{D}'$; this plays an important role in theoretical bounds. In the next lemma, a bound on the difference between $V_S^{\Phi}(h)$ and $V_T^{\Phi}(h)$ is established based on the $\mathcal{H}$-divergence between source and target.

Lemma 9. *Let $h \in \mathcal{H}$ and let Φ be a representation function. Then*

$$V^{\Phi}(h) \geq V_S^{\Phi}(h) - k d_{\mathcal{H}}(\mathcal{D}_T^{\Phi}, \mathcal{D}_S^{\Phi})$$

Proof. The proof is similar to [BBC07, BCK08]. The following inequality holds:

$$\begin{aligned} V_S^{\Phi}(h) &= \mathbb{E}_{(z,a)\sim\mathcal{D}_S^{\Phi}}\left[\frac{m_a(z)}{1/k}1(h(z)=a)\right] \\ &\leq \mathbb{E}_{(z,a)\sim\mathcal{D}_T^{\Phi}}\left[\frac{m_a(z)}{1/k}1(h(z)=a)\right] \\ &+ k\left|\mathbb{P}_{(z,a)\sim\mathcal{D}_T^{\Phi}}(\mathcal{I}_h) - \mathbb{P}_{(z,a)\sim\mathcal{D}_S^{\Phi}}(\mathcal{I}_h)\right| \\ &\leq V^{\Phi}(h) + kd_{\mathcal{H}}(\mathcal{D}_S^{\Phi}, \mathcal{D}_T^{\Phi}) \end{aligned}$$

where the first inequality holds because $\frac{m_a(z)}{1/k} \leq k$ for all pairs (z, a) and outcomes lie in the interval $[0, 1]$. □

Lemma 9 shows that the true policy outcome is at least as good as the policy outcome in the observational data minus the product of the number of actions times the $\mathcal{H}$-divergence between the observational and randomized data. (So, if the divergence is small, a policy that is found to be good with respect to the observational data is guaranteed to be a good policy with respect to the true distribution.)

Definition 5. *Let* Φ *be a representation function such that* $\Phi(x_i) = z_i$*. The* Monte-Carlo estimator *for the policy outcome in source data is given by:*

$$\widehat{V}_S^{\Phi}(h) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i 1(h(z_i)=a_i)}{1/K}.$$

In order to provide uniform bounds on the Monte-Carlo estimator for an infinitely large class of recommendation functions, we need to first define a complexity term for a class $\mathcal{H}$. For $\epsilon > 0$, a policy class $\mathcal{H}$ and integer n, the growth function is defined as

$$\mathcal{N}_{\infty}(\epsilon, \mathcal{H}, n) = \sup_{\boldsymbol{z}\in\boldsymbol{Z}^n} \mathcal{N}(\epsilon, \mathcal{H}(\boldsymbol{z}), \|\cdot\|_{\infty}),$$

where $\mathcal{H}(\boldsymbol{z}) = \{(h(z_1), \ldots, h(z_n)) : h \in \mathcal{H}\} \subset \mathbb{R}^n$, $\boldsymbol{Z}^n$ is the set of all possible n representations and for $\mathcal{A} \subset \mathbb{R}^n$ the number $\mathcal{N}(\epsilon, A, \|\cdot\|_{\infty})$ is the cardinality $|\mathcal{A}_0|$ of the smallest set $\mathcal{A}_0 \subseteq \mathcal{A}$ such that A is contained in the union of ϵ-balls centered at points in $\mathcal{A}_0$ in the metric induced by $\|\cdot\|_{\infty}$. (This is often called the covering number.) Set $\mathcal{M}(n) = 10\mathcal{N}_{\infty}(1/n, \mathcal{H}, 2n)$. The following result provides an inequality between the estimated and true $V_S^{\Phi}(h)$ for all $h \in \mathcal{H}$.

Lemma 10. *[MP09] Fix $\delta \in (0,1)$, $n \geq 16$. Then, with probability $1-\delta$, we have for all $h \in \mathcal{H}$:*

$$V_S^{\Phi}(h) \geq \widehat{V}_S^{\Phi}(h) - \sqrt{\frac{18\ln(\mathcal{M}(n)/\delta)}{n}} - \frac{15\ln(\mathcal{M}(n)/\delta)}{n}$$

In order to provide a data dependent bound on the estimation error between $V(h)$ and $\widehat{V}_S(h)$, we need to provide data-dependent bounds on the $\mathcal{H}$-divergence between source and target distributions. However, we aren't given samples from the target data so we need to generate (random) target data. Let $\widehat{\mathcal{D}}_S^{\Phi} = \{(Z_i, A_i)\}_{i=1}^n$ denote the empirical distribution of the source data. From the empirical source distribution, target data can be generated by simply sampling the actions uniformly, that is, $\widehat{\mathcal{D}}_T^{\Phi} = \{(Z_i, \widetilde{A}_i)\}_{i=1}^n$ where $\widetilde{A}_i \sim \text{Multinomial}([1/K, \ldots, 1/K])$. Then, we have $\widehat{\mathcal{D}}_S^{\Phi} \sim \mathcal{D}_S^{\Phi}$ and $\widehat{\mathcal{D}}_T^{\Phi} \sim \mathcal{D}_T^{\Phi}$. Then, define the empirical probability estimates of the characteristic functions as

$$\begin{aligned}
\mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_S^{\Phi}}(\mathcal{I}_h) &= \frac{1}{n}\sum_{i=1}^n 1(h(Z_i) = A_i), \\
\mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_T^{\Phi}}(\mathcal{I}_h) &= \frac{1}{n}\sum_{i=1}^n 1(h(Z_i) = \tilde{A}_i).
\end{aligned}$$

Then, one can compute empirical $\mathcal{H}$-divergence between two samples $\widehat{\mathcal{D}}_S^{\Phi}$ and $\widehat{\mathcal{D}}_T^{\Phi}$ by

$$d_{\mathcal{H}}(\widehat{\mathcal{D}}_T^{\Phi}, \widehat{\mathcal{D}}_S^{\Phi}) = \sup_{h\in\mathcal{H}} \left| \mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_T^{\Phi}}(\mathcal{I}_h) - \mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_S^{\Phi}}(\mathcal{I}_h) \right|. \tag{4.1}$$

In the next lemma, we provide estimation bounds between the empirical $\mathcal{H}$-divergence and true $\mathcal{H}$-divergence.

Lemma 11. *Fix $\delta \in (0,1)$, $n \geq 16$. Then, with probability $1-2\delta$, we have for all $h \in \mathcal{H}$:*

$$\begin{aligned}
d_{\mathcal{H}}(\mathcal{D}_T^{\Phi}, \mathcal{D}_S^{\Phi}) &\geq d_{\mathcal{H}}(\widehat{\mathcal{D}}_T^{\Phi}, \widehat{\mathcal{D}}_S^{\Phi}) \\
&- 2\left[\sqrt{\frac{18\ln(\mathcal{M}(n)/\delta)}{n}} - \frac{15\ln(\mathcal{M}(n)/\delta)}{n}\right]
\end{aligned}$$

Proof. Define $\beta(\delta, n) = \sqrt{\frac{18\ln(\mathcal{M}(n)/\delta)}{n}} - \frac{15\ln(\mathcal{M}(n)/\delta)}{n}$. By [MP09], with probability $1-\delta$, we have for each hypothesis $h \in \mathcal{H}$,

$$\begin{aligned}
\mathbb{P}_{(z,a)\sim\mathcal{D}_T^{\Phi}}(\mathcal{I}_h) &\geq \mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_T^{\Phi}}(\mathcal{I}_h) - \beta(\delta, n) \\
\mathbb{P}_{(z,a)\sim\mathcal{D}_S^{\Phi}}(\mathcal{I}_h) &\leq \mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_S^{\Phi}}(\mathcal{I}_h) + \beta(\delta, n)
\end{aligned}$$

Hence, by union bound, the following equation holds for all $h \in \mathcal{H}$ with probability $1 - 2\delta$:

$$\left|\mathbb{P}_{(z,a)\sim\mathcal{D}_T^\Phi}(\mathcal{I}_h) - \mathbb{P}_{(z,a)\sim\mathcal{D}_S^\Phi}(\mathcal{I}_h)\right| \geq \left|\mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_T^\Phi}(\mathcal{I}_h) - \mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_S^\Phi}(\mathcal{I}_h) - 2\beta(\delta,n)\right|$$

The inequality still holds by taking supremum over $\mathcal{H}$ with $1 - 2\delta$, that is,

$$\begin{aligned} & d_{\mathcal{H}}(\mathcal{D}_T^\Phi, \mathcal{D}_S^\Phi) \\ \geq & \sup_{h\in\mathcal{H}} \left|\mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_T^\Phi}(\mathcal{I}_h) - \mathbb{P}_{(z,a)\sim\widehat{\mathcal{D}}_S^\Phi}(\mathcal{I}_h) - 2\beta(\delta,n)\right| \\ \geq & d_{\mathcal{H}}(\widehat{\mathcal{D}}_T^\Phi, \widehat{\mathcal{D}}_S^\Phi) - 2\beta(\delta,n). \end{aligned}$$

where the last inequality follows from the triangle inequality. □

Finally, by combining Lemmas 9,10 and 11,a data-dependent bound on the counterfactual estimation error is obtained.

Theorem 10. *Fix $\delta \in (0,1)$, $n \geq 16$. Let Φ be the representation function and let $\mathcal{H}$ be the set of policies. Then, with probability at least $1 - 3\delta$, we have for all $h \in \mathcal{H}$:*

$$\begin{aligned} V^\Phi(h) \geq & \widehat{V}_S^\Phi(h) - k d_{\mathcal{H}}(\widehat{\mathcal{D}}_S^\Phi, \widehat{\mathcal{D}}_T^\Phi) \\ & - 3k\left[\sqrt{\frac{18\ln(\mathcal{M}(n)/\delta)}{n}} - \frac{15\ln(\mathcal{M}(n)/\delta)}{n}\right] \end{aligned}$$

This result extends the result provided in [SJS17] since their theoretical bounds are restricted to two-action problems and extends the result in [SJ15b] since they require the propensity scores to be known. The result provided in Theorem 10 is constructive and motivates our optimization criteria.

4.5 Counterfactual Policy Optimization (CPO)

Theorem 10 motivates a general framework for designing policy learning from observational data with bandit feedback. A learning algorithm following this criterion solves:

$$\widehat{\Phi}, \widehat{h} = \arg\max_{\Phi,h} \ \widehat{V}_S^\Phi(h) - \lambda d_{\mathcal{H}}(\widehat{\mathcal{D}}_S^\Phi, \widehat{\mathcal{D}}_T^\Phi),$$

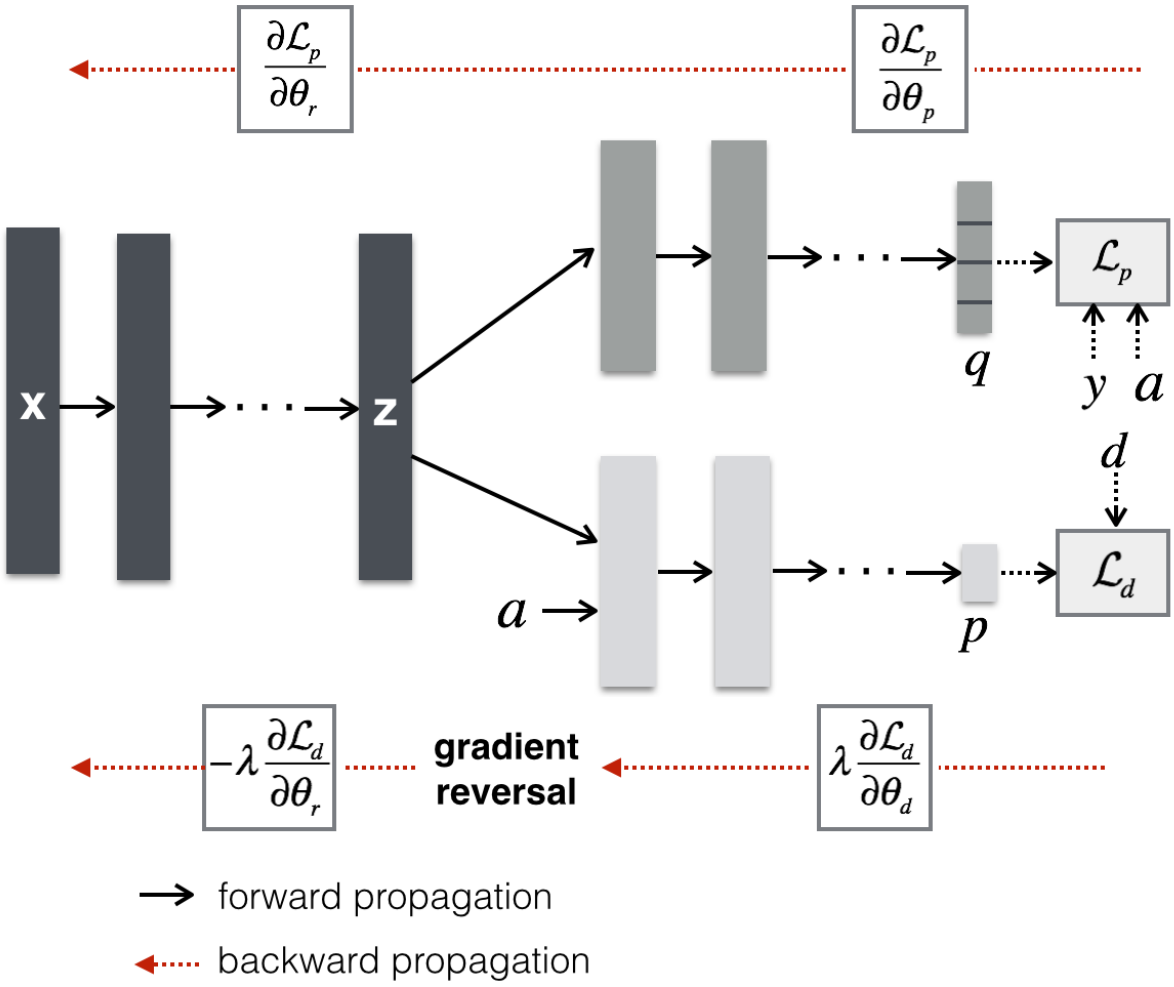


Figure 4.1: Neural network model based on [GUA16]

where $\lambda > 0$ is the trade-off parameter between the empirical policy outcome in the source data and the empirical $\mathcal{H}$-divergence between the source and target distributions. This optimization criterion seeks to find a representation function where the source and the target domain are indistinguishable. Computing the empirical $\mathcal{H}$-divergence between the source and target distributions is known to be NP-hard [GUA16], but we can use recent developments in domain adversarial neural networks to find a good approximation.

Algorithm 5 Procedure: Generate − Batch

Input: Data: $\mathcal{D}_n$, Batch size: m

Sample $\mathcal{U} = \{u_1, \ldots, u_m\} \subset \mathcal{N} = \{1, \ldots, n\}$.

Set source set $\mathcal{S} = \{(x_{u_i}, a_{u_i}, y_{u_i}, d_i = 0)\}_{i=1}^{m}$.

Sample $\mathcal{V} = \{v_1, \ldots, v_m\} \subset \mathcal{N} \setminus \mathcal{U}$.

Set $\mathcal{T} = \emptyset$

for i = 1, ..., m: **do**

    Sample $\widetilde{a}_i \sim$ Multinomial($[1/K, \ldots, 1/K]$).

    $\mathcal{T} = \mathcal{T} \cup \{(x_{v_i}, \widetilde{a}_i, d_i = 1)\}$.

end for

Output: $\mathcal{S}$, $\mathcal{T}$.

4.5.1 Domain Adversarial Neural Networks

This chapter follows the recent work in domain adversarial training of neural networks [GUA16]. For this, only samples from observed data is needed - sometimes referred to as source data ($\mathcal{D}_S$) - and unlabeled samples from an ideal dataset - referred to as target data ($\mathcal{D}_T$). As mentioned, there are no samples from ideal dataset. Hence, target dataset needs to be generated from source dataset by batch sampling. Given a batch size of m, instances from the source data $\mathcal{D}$ are sampled uniformly randomly and their domain variable set to 0, $d = 0$, indicating this is the source data. Then, additional m instances are uniformly randomly sampled excluding the instances from the source data, are randomly assigned to one of the treatments by a Multinomial distribution, Multinomial($[1/k, \ldots, 1/k]$); finally, their domain variable are set to 1, $d = 1$, indicating this is the target data. The batch generation procedure is depicted in Algorithm 5.

The algorithm, referred to as Domain Adverse training of Counterfactual POLicy training (DACPOL), consists of three blocks: representation, domain and policy blocks. In the representation block, DACPOL seeks to find a map $\Phi : \mathcal{X} \to \mathcal{Z}$ combining two objectives: (i) high predictive power on the outcomes, (ii) low predictive power on the domain prediction. Let F_r denote a parametric function that maps the patient features to representations,

that is, $z_i = F_r(x_i; \theta_r)$ where θ_r is the parameter vector of the representation block. The representations are input to both survival and policy blocks. Let F_p denote the mappings from representation-action pair (z_i, a_i) to probabilities over the actions $\hat{q}_i = [\hat{q}_{i,0}, \ldots, \hat{q}_{i,K-1}]$, i.e., $\hat{q}_i = F_p(z_i, a_i; \theta_p)$ where θ_p is the parameter vector of the policy block. For an instance with features x_i and action a_i, an element in output of policy block $\hat{q}_{i,a}$ is the probability of recommending action a for subject i. The estimated policy outcome in source data is then given by

$$\widehat{V}_S^{\Phi}(h) = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i q_{a_i}}{1/k}.$$

Although theory developed above applies only to deterministic policies, DACPOL allows for stochastic policies in order to make the optimization problem tractable. This is not optimal; however, as shown later in our numerical results, this approach is still able to achieve significant gains with respect to benchmark algorithms. Let G_d be a mapping from representation-action pair (z_i, a_i) to probability of the instances generated from target, i.e., $\hat{p}_i = G_d(z_i, a_i; \theta_d)$ where θ_d is the parameters of the domain block.

Note that the last layer of the policy block is a softmax operation, which has exponential terms. Instead of directly maximizing $\widehat{V}_S(h)$, we use a modified cross-entropy loss to make the optimization criteria more robust. The policy loss is then

$$\mathcal{L}_p^i(\theta_r, \theta_s) = \frac{-y_i \log(q_{i,a_i})}{1/k}$$

At the testing stage, we can then convert these probabilities to action recommendations simply by recommending the action with highest probability $q_{i,a}$. We set the domain loss to be the standard cross entropy loss between the estimated domain probability p_i and the actual domain probability d_i; this is the standard classification loss used in the literature and is given by

$$\mathcal{L}_d^i(\theta_r, \theta_s) = d_i \log(p_i) + (1 - d_i) \log p_i.$$

The DACPOL aims to find the saddle point that optimizes the weighted sum of the

Algorithm 6 Training Procedure: DACPOL

Input: Data: $\mathcal{D}$, Batch size: m, Learning rate: μ

$(\mathcal{S}, \mathcal{T}) =$ Generate-Batch$(\mathcal{D}, m)$.

for until convergence **do**

Compute $\mathcal{L}_p^{\mathcal{S}}(\theta_r, \theta_s) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathcal{L}_p^i(\theta_r, \theta_s)$

Compute $\mathcal{L}_d^{\mathcal{S}}(\theta_r, \theta_d) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \mathcal{L}_d^i(\theta_r, \theta_d)$

Compute $\mathcal{L}_d^{\mathcal{T}}(\theta_r, \theta_d) = \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{T}} \mathcal{L}_d^i(\theta_r, \theta_d)$

Compute $\mathcal{L}_d(\theta_r, \theta_d) = \mathcal{L}_d^{\mathcal{S}}(\theta_r, \theta_d) + \mathcal{L}_d^{\mathcal{T}}(\theta_r, \theta_d)$

$\theta_r \rightarrow \theta_r - \mu \left(\frac{\partial \mathcal{L}_s^{\mathcal{S}}(\theta_r, \theta_s)}{\partial \theta_r} - \lambda \frac{\partial \mathcal{L}_d(\theta_r, \theta_d)}{\partial \theta_r} \right)$

$\theta_p \leftarrow \theta_p - \mu \frac{\partial \mathcal{L}_p^{\mathcal{S}}(\theta_r, \theta_s)}{\theta_s}$

$\theta_d \leftarrow \theta_d - \mu \frac{\partial \mathcal{L}_d(\theta_r, \theta_d)}{\theta_d}$

end for

survival and domain loss. The total loss is given by

$$\begin{aligned} \mathcal{E}(\theta_r, \theta_s, \theta_d) &= \sum_{i \in \mathcal{S}} \mathcal{L}_s^i(\theta_r, \theta_s) \\ &\quad -\lambda \left(\sum_{i \in \mathcal{S}} \mathcal{L}_d^i(\theta_r, \theta_d) + \sum_{i \in \mathcal{T}} \mathcal{L}_d^i(\theta_r, \theta_d) \right) \end{aligned}$$

where $\lambda > 0$ is the trade-off between survival and domain loss. The saddle point is

$$\begin{aligned} \left(\hat{\theta}_r, \hat{\theta}_s \right) &= \arg\min_{\theta_r, \theta_p} \mathcal{E}(\theta_r, \theta_p, \hat{\theta}_d), \\ \hat{\theta}_d &= \arg\max_{\theta_d} \mathcal{E}(\hat{\theta}_r, \hat{\theta}_p, \theta_d). \end{aligned}$$

The training procedure of the DACPOL is depicted in Algorithm 6. The neural network architecture is depicted in Figure 4.1.

For a test instance with covariates x^*, DACPOL computes the action recommendations with the following procedure: it first computes the representations by $z^* = G_r(x^*; \theta_r)$, then computes the action probabilities $q^* = F_p(z^*, \theta_p)$. It finally recommends the action with $\hat{A}(x^*) = \arg\max_{a \in \mathcal{A}} q_a^*$.

4.6 Numerical Results

This section describes the performance of our algorithm. Note that it is difficult (almost impossible) to test and validate the algorithm on real data with missing counterfactual survival outcomes. In this chapter, we provide results both on a semi-synthetic breast cancer (description can be found in 3.7.1).

4.6.1 Experimental Setup

Artificially biased dataset $\mathcal{D}^n = \{(X_i, A_i, Y_i)\}$ is generated by the following procedure: (i) first random weights $W \in \mathbb{R}^{s\times k}$ with $w_{j,a} \sim \mathcal{N}(0, \sigma I)$ are drawn where $\sigma > 0$ is a parameter used to generate datasets with different selection bias levels, (ii) actions are generated in the data according to the logistic distribution $A \sim \exp(x^T w_a)/(\sum_{a\in\mathcal{A}} \exp(x^T w_a))$.

For the breast cancer data set, we generate a 56/24/20 split of the data to train, validate and test our DACPOL. The hyperparameter list in our validation set is $10^\gamma/2$ with $\gamma \in [-4, -3, -2, -1, 0, 0.5, 0.75, 1, 1.5, 2, 3]$. 100 different datasets are generated by following the procedure described above. Average of the metrics together with 95% confidence levels are reported.

The performance metric used to evaluate our algorithm in this paper is loss, which is defined to be $1 -$ accuracy; accuracy is defined as the fraction of test instances in which the recommended and best action match. Note that the accuracy metric can be evaluated since we have the ground truth outcomes in the testing set, but of course the ground truth outcomes are not used by any algorithm in the training and validation test. In the experiments, we use 1-1-2 representation/domain/outcome fully-connected layers. The neural network is trained by back propagation via Adam Optimizer [KB14] with an initial learning rate of .01. We begin with an initial learning rate μ and tradeoff parameter λ and use iterative adaptive parameters to get our result; along the way we decrease the learning rate μ and increases the tradeoff parameter. This is standard procedure in training domain adversarial neural networks [GUA16]. The DACPOL is implemented in the Tensorflow environment.

4.6.2 Benchmarks

We compare performance of DACPOL with two benchmarks

- **POEM** [SJ15b] is a linear policy optimization algorithm which minimizes the empirical risk of IPS estimator and variance.
- **IPS** is POEM without variance regularization.

Both IPS and POEM deal with the selection bias in the data by using the propensity scores. Note that DACPOL does not require the propensity scores to be known in order to address the selection bias. Hence, in order to make fair comparisons, we estimate the propensity scores from the data, and use these estimates in IPS and POEM.

4.6.3 Results

4.6.3.1 Comparisons with the benchmarks

Table 4.2 shows the discriminative performance of DACPOL (in which we optimize λ) and DACPOL(0) (in which we set $\lambda = 0$) with the benchmark algorithms. The experiments are conducted in breast cancer data in which actions generated by logistic model with $\sigma = 0.3$. As seen from the table, DACPOL outperforms the benchmark algorithms in terms of the loss metric defined above. The empirical gain with respect to POEM algorithm has three sources: (i) DACPOL does not need propensity scores, (ii) DACPOL optimizes over all policies not just linear policies, (iii) DACPOL trades off between the predictive power and bias introduced by the features. (the last source of gain is illustrated in the next subsection with a toy example.)

4.6.3.2 Domain Loss and Policy Loss

The hyperparameter λ controls the domain loss in the training procedure. As λ increases, the domain loss in training DACPOL increases; eventually source and target become indistinguishable, the representations become balanced, and the loss of DACPOL reaches a

Algorithm	Breast Cancer
DACPOL	$.292 \pm .006$
DACPOL(0)	$.321 \pm .006$
POEM	$.394 \pm .004$
IPS	$.397 \pm .004$

Table 4.2: Loss Comparisons for Breast Cancer Dataset; Means and 95% Confidence Intervals

minimum. If λ is increased beyond that point, the DACPOL classifies the source as the target and the target as the source, representations become unbalanced, and the the loss of DACPOL increases again. Figure 4.2 illustrates this effect for the breast cancer dataset.

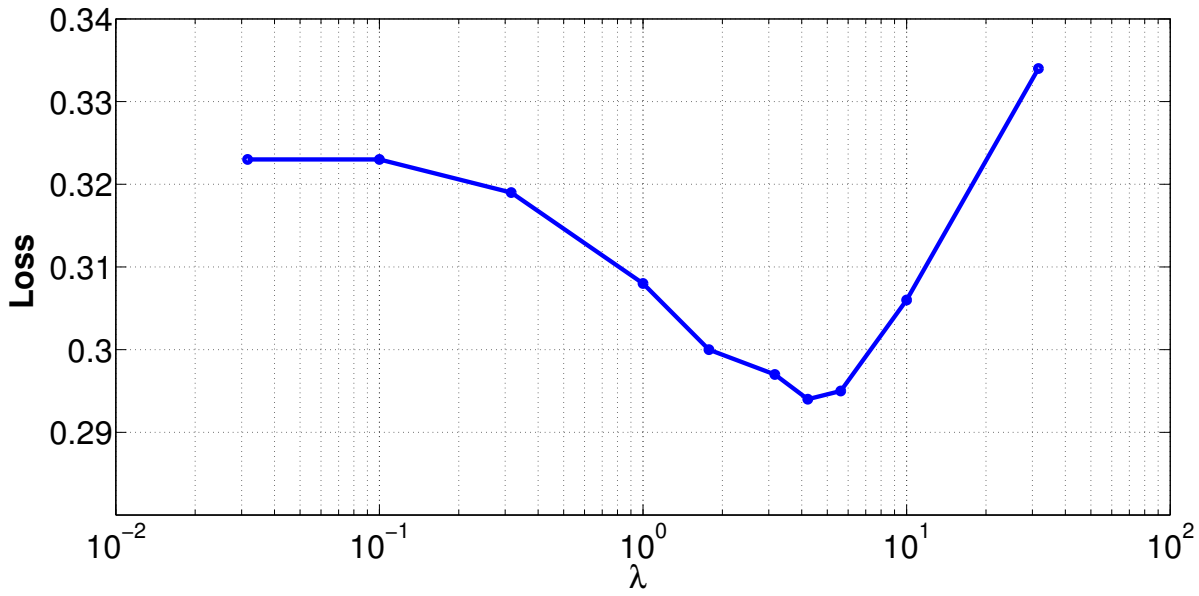


Figure 4.2: The effect of domain loss in DACPOL performance

4.6.3.3 The effect of selection bias in DACPOL

In this subsection, we show the effect of the selection bias in the performance of our algorithm by varying the parameter σ in our data generation process: a larger value of σ creates more biased data. Figure 4.3 shows two important points: (i) as the selection bias increases, the loss of DACPOL increases, (ii) as the selection bias increases, domain adversarial training becomes more efficient, and hence the improvement of DACPOL over DACPOL(0) increases.

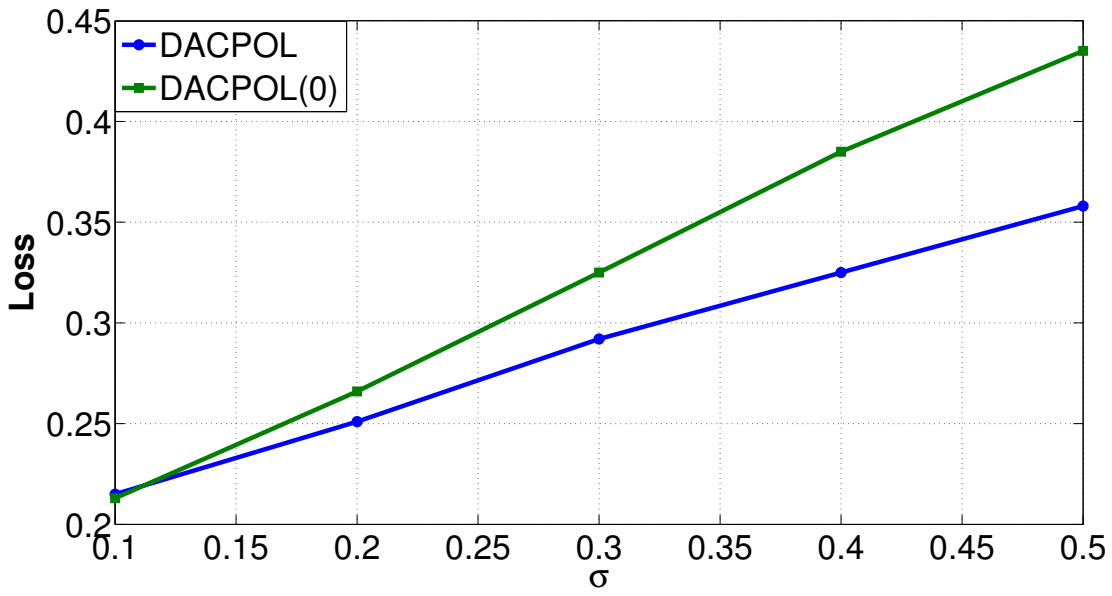


Figure 4.3: The effect of selection bias in DACPOL performance

4.6.3.4 Empirical Gains with respect to CRM

In this subsection, we show an advantage of our CPO principle over the CRM principle: selection bias from irrelevant features is less important. This happens because our representation optimization is able to remove the effect of the irrelevant features in the outputted representations and then uses only the relevant features to directly estimate the policy outcome as if it had access to randomized data. However, the performance of the CRM principle (whose objective is to maximize the IPS estimator minus the variance of the policy outcome) decreases with additional irrelevant features, because the inverse propensities due to irrelevant features become large, and hence the variance of the IPS estimator will also become large. To see this, we use a toy example. We begin with 15 relevant features x. We then generate d additional irrelevant features $z \sim \mathcal{N}(0, I)$. We create a logging policy that depends only on the irrelevant features using the logistic distribution. As d increases, the selection bias also increases. As Figure 4.4 shows, POEM is more sensitive than DACPOL to this increase in the selection bias.

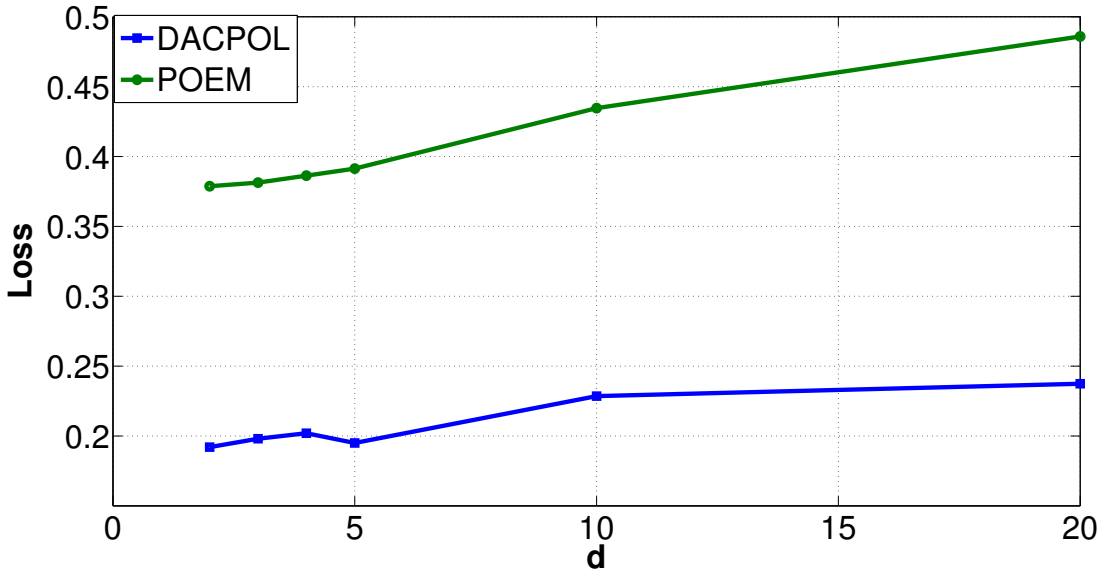


Figure 4.4: The effect of irrelevant features in DACPOL vs POEM

CHAPTER 5

Online Decision Making with Costly Observations

5.1 Introduction

In numerous real-world settings, acquiring useful information is often costly. In health-care applications, the decision-maker performs costly research/experimentation to learn valuable information. However, classical contextual Multi-Armed Bandit (MAB) formulations ([CLR11, Sli11, LPP10, DHK11, LZ07]) have not previously considered these important informational costs and are thus unable to provide satisfactory performance in such settings. This chapter presents new and powerful methods and algorithms for Contextual MAB with Costly Observations (CMAB-CO). The experiments on breast cancer dataset show that algorithms achieve significant performance gains with respect to benchmarks.

A major challenge in these settings is the learning of both optimal observations and actions. Current MAB methods could potentially be modified to address this issue by combining the choice of the context to observe and the action to be taken as a single meta-action and folding the costs of observations in the rewards. However, the regret of such an approach can be shown to be exponential in the number of actions and the number of possible context states; therefore, it is so inefficient as to be impractical for any realistic problem. Therefore there is a strong need for the development of new algorithms that achieve better performance.

To overcome the limitations and challenges discussed above, we propose an alternative approach. We formalize the CMAB-CO problem and show that this problem can be reduced to a two stage Markov Decision Process (MDP) problem with a canonical start state. We propose two different algorithms for this learning problem: Sim-OOS and Seq-OOS where observations are made simultaneously and sequentially, respectively. These algorithms build

upon the UCRL2 algorithm of ([JOA10]) to efficiently learn optimal observations and actions. We show that both Sim-OOS and Seq-OOS algorithms achieve a regret that is sublinear in time. These algorithm thus perform well when the number of observations is small, and it represents a significant improvement over existing algorithms, which would be exponential in the number of observations as well as actions.

The main contributions in this chapter can be summarized as follows:

- We formalize the CMAB-CO problem as a fixed stage MDP.
- We propose two algorithms under two assumptions: simultaneous and sequential observation selection. We show sublinear in time regret bounds for both algorithms.
- We use a breast cancer dataset and show that we can achieve up to significant improvement in performance with respect to benchmarks.

In the medical context, the observations might consist of different types of (costly) medical tests (e.g., blood tests, MRI, etc.), actions might consist of choices of treatment, and rewards might consist of 5 year survival rates. Hence, an important aspect of the decision-making is which medical tests to conduct and which treatment option to recommend.

5.2 Related Work

This chapter contributes to multiple strands of literature, including MAB, MDP and budgeted learning.

5.2.1 MAB Literature

This chapter relates to various strands of research in the MAB literature ([TS15c, TS15b, TYS17, STS16, CLR11, Sli11, LPP10, DHK11, LZ07]). For example, [TS15c] focuses on learning the optimal actions by discovering relevant information. However, this work does not consider the costs associated with gathering information and is thus unable to provide satisfactory performance in the considered setting. The CMAB-CO problem is similar to combi-

natorial semi-bandits since multiple actions (observations and real actions) are selected and the rewards of all selected actions (observation cost and real action rewards) are selected in our setting. However, combinatorial semi-bandits do not utilize the observed states when taking the action.

This chapter is also very related to online probing ([ZBG13]). However, the goal in ([ZBG13]) is to learn the optimal observations and a single best function that maps observed features to labels in order to minimize the loss and the observation cost jointly. Unlike in the considered CMAB-CO setting, an adversarial setup is assumed and a complete loss feedback (the loss associated with all the various actions) is obtained at each stage.

5.2.2 MDP literature

The CMAB-CO problem which we consider can be formalized as a two-stage MDP ([JOA10, OA07,OVW16]) with a canonical start state. The action set available in the start state is the set of observations. Following an observation action in the start state, the decision-maker moves to a new state (which consists of the realized states of the selected observations) from which the decision-maker selects a real action and moves back to the start state. The reward in the first step is the observation cost (negative) and the second step is the random reward obtained by taking the real action. Stemming from this and building upon the UCRL2 algorithm of ([OA07, JOA10]), efficient algorithms are constructed by exploiting the structure of the CMAB-CO problem: sparse observation probabilities, known costs.

There are also large strands of work in signal processing community studying Partially Observable Markov Decision Process (POMDP) [ZLM14, ZM17]. The goal in these works is to estimate joint probability distribution over the variables. The goal in this chapter on the other hand is to use the estimated variables to make observation and action decisions.

5.2.3 Budgeted Learning

The CMAB-CO problem is also similar to budgeted learning as the decision-maker's goal there is to adaptively choose which features to observe in order to minimize the loss. For

example, ([CSS11, HK12]) adaptively choose the features of the next training example in order to train a linear regression model while having restricted access to only a subset of the features. However, these problems do not consider information costs and are restricted to batch learning.

Another related work is adaptive submodularity ([GK10]) which aims to maximize rewards by selecting at most m observations/actions. However, their approach assumes that observation states are statistically independent and rewards have a sub-modular structure in observations.

5.3 Contextual Multi-armed Bandits with Costly Observations

5.3.1 Problem Formulation

In this subsection, we present our problem formulation and illustrate it with a specific example from in the medical context. Let $\mathcal{D} = \{1, 2, \ldots, D\}$ be a finite set of observations (types of medical tests such as MRI, mamogram, ultrasound etc.). Each observation $i \in \mathcal{D}$ is in a (initially unknown) particular state from a finite set of $\mathcal{X}_i$ of possible values (describing the outcomes of the medical tests such as the BIRADS score associated with a mamogram). Let $\mathcal{X} = \cup_{i \in \mathcal{D}} \mathcal{X}_i$ represent the set of all possible state vectors. The state vector is $\phi = (\phi[1], \phi[2], \ldots, \phi[D])$, where $\phi[i]$ is the state of observation i, which represents the context in the CMAB formulation. We assume that the state vector is drawn according to a fixed but unknown distribution. Write ϕ to denote a random state vector and $p(\phi) = \Pr(\phi = \phi)$ to denote the probability of state vector ϕ being drawn. In the medical context, $p(\cdot)$ models a joint probability over the results of the medical tests.

We assume that only the states of the observations that are selected by the decision-maker are revealed in each time instance. Let ψ denote a partial state vector, which only contains the state of a subset of the selected observations. For example, for selected observations

$\mathcal{I} \subseteq \mathcal{D}$, the partial state vector is $\psi = (\psi[1], \psi[2], \ldots, \psi[D])$ with

$$\psi[i] = \begin{cases} \phi[i] & \text{if } i \in \mathcal{I} \\ ? & \text{if } i \notin \mathcal{I} \end{cases}$$

where ? denotes our symbol for missing observation states. Denote $\text{dom}(\psi) = \{i \in \mathcal{D} : \psi[i] \neq ?\}$ as the domain of ψ (i.e., the set of the medical test outcomes realized in ψ). Let $\Psi^+(\mathcal{I}) = \{\psi : \text{dom}(\psi) = \mathcal{I}\}$ denote the set of all possible partial state vectors with observations from $\mathcal{I}$ (i.e., the set of all possible medical test outcomes of $\mathcal{I}$). Let $\Psi = \cup_{\mathcal{I} \subseteq \mathcal{D}} \Psi^+(\mathcal{I})$ denote the set of all possible partial state vector states. We say ψ is *consistent* with ϕ if they are equal everywhere in the domain of ψ, i.e., $\psi[i] = \phi[i]$ for all $i \in \text{dom}(\psi)$. In this case, we write $\phi \sim \psi$. If ψ and ψ' are both consistent with some ϕ, and $\text{dom}(\psi) \subseteq \text{dom}(\psi')$, we say ψ is a *substate* of ψ'. In this case, we write $\psi' \succeq \psi$.

We illustrate these definitions on a simple example. Let $\phi = (-1, 1, 1)$ be a state vector, and $\psi_1 = (-1, ?, -1)$ and $\psi_2 = (-1, ?, ?)$ be partial state vectors. Then, all of the following claims are true:

$$\phi \sim \psi_2,\ \psi_1 \succeq \psi_2,\ \text{dom}(\psi_1) = \{1, 3\}.$$

A MAB setting is considered with costly observations where the following sequence of the events is taking place at each time t:

- The environment draws a state vector ϕ_t according to unknown distribution $p(\cdot)$. The state vector is initially unknown to the decision-maker.

- The decision-maker is allowed to select at most m observation at time t, denoted as $\mathcal{I}_t$, with paying a known cost of $c_i \in [0, 1]$ for each observations i in the set $\mathcal{I}_t$. We assume that the decision-maker has an upper bound m on the maximum number of observations that can be made at each time t. Let $\mathcal{P}_{\leq m}(\mathcal{D})$ denote the subset of the observations with cardinality less than m, i.e., $\mathcal{P}_{\leq m}(\mathcal{D}) = \{\mathcal{I} \subseteq \mathcal{D} : |\mathcal{I}| \leq m\}$. The partial state vector ψ_t from the observations $\mathcal{I}_t$ is revealed to the decision-maker, while the remainder of the states remain unknown to the decision-maker.

- Based on its available information ψ_t, the decision-maker takes an action a_t from a finite set of actions $\mathcal{A} = \{1, 2, \ldots, A\}$ and observes a random reward r_t with support $[0, 1]$ and $\mathbb{E}[r_t | A_t = a, \psi_t = \psi] = \bar{r}(a, \psi)$ where $\bar{r} : \mathcal{A} \times \Psi \to [0, 1]$ is an unknown expected reward function.

We write $p(\psi) = \Pr(\phi \sim \psi)$ to denote the marginal probability of ψ being realized. Observe that $\sum_{\psi \in \Psi^+(\mathcal{I})} p(\psi) = 1$ for any $\mathcal{I}$.

The *policy* π for selecting observations and associated actions consists of a set of observations $\mathcal{I}$ and an adaptive action strategy $h : \Psi^+(\mathcal{I}) \to \mathcal{A}$, which maps each possible partial state vectors from $\mathcal{I}$ to actions (e.g., a policy consists of a subset of medical tests $\mathcal{I}$ to be conducted and treatment recommendation for each possible test results). The expected gain of the policy $\pi = \{\mathcal{I}, h\}$ is given by

$$V(\pi) = \beta \sum_{\psi \in \Psi^+(\mathcal{I})} p(\psi) \bar{r}(h(\psi), \psi) - \sum_{i \in \mathcal{I}} c_i, \tag{5.1}$$

where $\beta > 1$ is the gain parameter, which balances the trade-off between the rewards and observation costs. The expected gain of the policy π is the expected reward of π minus the observation cost incurred by π. Without loss of generality, we assume that decision-maker is allowed to make at most m observations. Let Π_m denote the set of all possible policies with at most m observations. The oracle policy is given by $\pi^* = \arg\max_{\pi=(\mathcal{I},h) \in \Pi_m} V(\pi)$.

The expected gain of the oracle policy is given by $V^* = V(\pi^*)$. Note that the oracle is different than the oracle used in the contextual bandit literature. To illustrate the difference, define $\bar{r}^*(\psi) = \bar{r}(a^*(\psi), \psi) = \max_{a \in \mathcal{A}} \bar{r}(a, \psi)$ to be the expected reward of the best action when the partial state vector is ψ. We refer to the policy that selects observations $\mathcal{I}$ and the best actions $a^*(\psi)$ for all $\psi \in \Psi^+(\mathcal{I})$ as the fixed $\mathcal{I}$-oracle policy. The expected reward of the fixed $\mathcal{I}$-oracle policy is given by

$$V^*(\mathcal{I}) = \beta \sum_{\psi \in \Psi^+(\mathcal{I})} p(\psi) \bar{r}^*(\psi) - \sum_{i \in \mathcal{I}} c_i.$$

It can be shown that the oracle policy $\pi^* = (\mathcal{I}^*, h^*)$ is given by $h^*(\psi) = \arg\max_{a \in \mathcal{A}} \bar{r}(a, \psi)$ and $\mathcal{I}^* = \arg\max_{\mathcal{I} \in \mathcal{P}_{\leq m}(\mathcal{D})} V^*(\mathcal{I})$. Note that $V^* = V^*(\mathcal{I}^*)$. Therefore, the oracle defined in our setting achieves the best expected reward among all the fixed $\mathcal{I}$-oracle policies.

Consider an adaptive policy $\pi_{1:T} = [\mathcal{I}_t, h_t]_{t=1}^T$, which takes observation-action $\mathcal{I}_t$, observes ψ_t, uses this observation to take an action $a_t = h_t(\psi_t)$ and receives the reward of r_t. The cumulative reward of $\pi_{1:T}$ is $\sum_{t=1}^T \left(\beta r_t - \sum_{i \in \mathcal{I}_t} c_i\right)$. The T-time regret of the policy $\pi_{1:T} = [\mathcal{I}_t, h_t]_{t=1}^T$ is given by

$$\mathrm{Reg}_T^{\pi_{1:T}} = TV^* - \sum_{t=1}^{T} \left(\beta r_t - \sum_{i \in \mathcal{I}_t} c_i \right).$$

The goal here is to compute the policy $\pi_{1:T}$ to minimize this regret by selecting at most m observations.

Current online learning methods could be modified to address the CMAB-CO problem by defining a set of *meta-actions* that comprises all the combinations of observation subsets and actions taken based on these observations, and then applying a standard MAB algorithm (such as the UCB algorithm [ACF02b]) by considering these meta-actions to be the action space. While this algorithm is straightforward to implement, it scales linearly with the total number of policies $|\Pi| = \sum_{\mathcal{I} \in \mathcal{P}_{\leq m}(\mathcal{D})} A^{|\Psi^+(\mathcal{I})|}$. This is exponential in the number of state vectors. This makes such algorithms computationally infeasible and suboptimal (compared to the lower bound) even when the numbers of actions and partial states is small. This poor scaling performance is due to the fact that the algorithm does not take into account that selecting an action yields information for many policies.

5.3.2 Simultaneous Optimistic Observation Selection (Sim-OOS) Algorithm

To address the above mentioned limitations of such MAB algorithms, a new algorithm, which we refer to as Simultaneous Optimistic Observation Selection (Sim-OOS) is developed. Sim-OOS operates in rounds $k = 1, 2, \ldots$. Let t_k denote time at the beginning of round k. The decision-maker keeps track of the estimates of the mean rewards and the observation probabilities. Note that when the partial state vector ψ_t from observation set $\mathcal{I}_t$ is revealed, the decision-maker can use this information to not only update the observation probability estimate of ψ_t but also update the observation probability estimate of all substates of ψ_t. However, the decision-maker cannot update the mean reward estimate of pairs of a_t and substates of ψ_t since this would result in a bias on the mean reward estimates. Therefore,

Algorithm 7 Simultaneous Optimistic Observation Selection (Sim-OOS)

Input: $m, [c_i]_{i\in\mathcal{D}}$, $\text{conf}_1(n,t)$, $\text{conf}_2(n,t), \beta$.

Initialize: $\mathcal{E}(\text{dom}(\psi), \psi) \leftarrow \emptyset$ for all $\psi \in \Psi$.

Initialize: $\mathcal{E}(\mathcal{I}) \leftarrow \emptyset$ for all $\mathcal{I} \in \mathcal{P}_{\leq m}(\mathcal{D})$.

Initialize: $\mathcal{E}(a, \psi) \leftarrow \emptyset$ for all $a \in \mathcal{A}$ and $\psi \in \Psi$.

for rounds $k = 1, 2, \ldots$ **do**

$\text{conf}_{1,k}(a, \psi) \leftarrow \text{conf}_1(N_k(a, \psi), t_k)$.

$\text{conf}_{2,k}(\mathcal{I}) \leftarrow \text{conf}_{2,k}(N_k(\mathcal{I}), t_k)$.

$\widehat{r}_k(a, \psi) = \frac{1}{N_k(a,\psi)} \sum_{\tau\in\mathcal{E}_k(a,\psi)} r_\tau$ for all $a \in \mathcal{A}$ and $\psi \in \Psi$.

$\widehat{p}_k(\psi) = \frac{N_k(\text{dom}(\psi),\psi)}{N_k(\text{dom}(\psi))}$ for all $\psi \in \Psi$.

$\widehat{h}_k(\psi) \leftarrow \arg\max_{a\in\mathcal{A}} \widehat{r}_k(a, \psi) + \text{conf}_{1,k}(a, \psi)$

Solve the convex optimization problem given in (5.3) for all $\mathcal{I} \in \mathcal{P}_{\leq m}(\mathcal{D})$

Set $\widehat{V}_k(\mathcal{I})$ as the maximizer.

$\widehat{\mathcal{I}}_k \leftarrow \arg\max_{\mathcal{I}\in\mathcal{P}_{\leq m}(\mathcal{D})} \widehat{V}_k(\mathcal{I})$.

$\nu_k(a, \psi) \leftarrow 0$ for all a and $\psi \in \Psi$.

while $\forall(a, \psi) : \nu_k(a, \psi) < \max(1, N_k(a, \psi))$ **do**.

Select observations $\widehat{\mathcal{I}}_k$, observe the partial state vector ψ_t.

Select action $a_t = \widehat{h}_k(\psi_t)$, observe reward r_t.

Update $\nu_k(a_t, \psi_t) \leftarrow \nu_k(a_t, \psi_t) + 1$.

for $\psi : \psi_t \succeq \psi$ **do**

$\mathcal{E}_{k+1}(\text{dom}(\psi), \psi) \leftarrow \mathcal{E}_{k+1}(\psi, \text{dom}(\psi)) \cup t$.

$\mathcal{E}_{k+1}(\text{dom}(\psi)) \leftarrow \mathcal{E}_{k+1}(\text{dom}(\psi)) \cup t$.

end for

$\mathcal{E}_{k+1}(a_t, \psi_t) \leftarrow \mathcal{E}_{k+1}(a, \psi) \cup t$.

$t \leftarrow t + 1$.

end while

end for

at each round k, we define $\mathcal{E}_k(a, \psi) = \{\tau < t_k : a_\tau = a, \psi_\tau = \psi\}$, $\mathcal{E}_k(\mathcal{I}) = \{\tau < t_k : \mathcal{I} \subseteq \mathcal{I}_\tau\}$ and $\mathcal{E}_k(\psi, \mathcal{I}) = \{\tau < t_k : \mathcal{I} \subseteq \mathcal{I}_\tau, \psi_\tau \succeq \psi\}$ if $\psi \in \Psi^+(\mathcal{I})$ and $\mathcal{E}_k(\psi, \mathcal{I}) = \emptyset$ if $\psi \notin \Psi^+(\mathcal{I})$.

We define the following counters: $N_k(\mathcal{I},\psi) = |\mathcal{E}_k(\mathcal{I},\psi)|$, $N_k(\mathcal{I}) = |\mathcal{E}_k(\mathcal{I})|$, $N_k(a,\psi) = |\mathcal{E}_k(a,\psi)|$. In addition to these counters, we also keep counters of partial state-action pair visits in a specific round k. Let $\nu_k(a,\psi)$ denote the number of times action a is taken when partial state ψ is observed in round k. Furthermore, we can express the mean reward estimate and observation probability estimates as follows:

$$\widehat{r}_k(a,\psi) = \frac{1}{N_k(a,\psi)} \sum_{\tau \in \mathcal{E}_k(a,\psi)} r_\tau,$$

$$\widehat{p}_k(\psi) = \frac{N_k(\mathrm{dom}(\psi),\psi)}{N_k(\mathrm{dom}(\psi))}$$

provided that $N_k(a,\psi) > 0$ and $N_k(\mathrm{dom}(\psi)) > 0$. Since these estimates can deviate from their true mean values, we need to add appropriate confidence intervals when optimizing the policy. In the beginning of each round k, the Sim-OOS computes the policy of round k by solving an optimization problem given in (5.2). The optimization problem with the mean reward estimate and observation probability estimates is given by

$$\begin{aligned}
&\underset{\pi=\{\mathcal{I},h\},\tilde{p},\tilde{r}}{\text{maximize}} \ \beta \sum_{\psi\in\Psi^+(\mathcal{I})} \tilde{p}(\psi)\tilde{r}(h(\psi),\psi) - \sum_{i\in\mathcal{I}} c_i \\
&\text{subject to } \ |\tilde{r}(a,\psi) - \widehat{r}_k(a,\psi)| \le \mathrm{conf}_{1,k}(a,\psi), \ \ \forall (a,\psi), \\
&\sum_{\psi\in\Psi^+(\mathcal{I})} |\tilde{p}(\psi) - \widehat{p}_k(\psi)| \le \mathrm{conf}_{2,k}(\mathcal{I}), \\
&\sum_{\psi\in\Psi^+(\mathcal{I})} \tilde{p}(\psi) = 1, \ \ \forall \mathcal{I} \in \mathcal{P}_{\le m}(\mathcal{D}),
\end{aligned} \tag{5.2}$$

where $\mathrm{conf}_{1,k}(a,\psi)$ and $\mathrm{conf}_{2,k}(\mathcal{I})$ are the confidence bounds on the estimators at time t_k. We will set these confidence bounds later in order to achieve provable regret guarantees with high probability. Let $\widehat{\pi}_k = \{\widehat{\mathcal{I}}_k, \widehat{h}_k\}$ denote the policy computed by the Sim-OOS.

The Sim-OOS follows policy $\widehat{\pi}_k$ in round k. At time t in round k ($t_k \le t \le t_{k+1}$), the Sim-OOS selects $\widehat{\mathcal{I}}_k$ and observes the partial state vector ψ_t from observations $\mathcal{I}_k$ and on the basis of this, it takes an action $\widehat{h}_k(\psi_t)$. Round k ends when one of the visits to the partial state vector-action pair in round k is the same as $N_k(a,\psi)$ (the total observations of the partial state-action pair from previous rounds $k' = 1,\ldots,k-1$). This ensures that

the optimization problem given in (5.2) is only solved when the estimates and confidence bounds are improved.

The optimization problem in (5.2) can be reduced to a set of convex optimization problems which can be solved efficiently in polynomial time complexity ([BV04]) (the details of this reduction are discussed in the supplementary material). In round k, let $\widehat{r}_k^*(\psi) = \max_{a\in\mathcal{A}}\ \widehat{r}_k(a,\psi) + \mathrm{conf}_{1,k}(a,\psi)$ be the optimistic reward of value of the partial state vector ψ in round of k. The optimistic gain of a fixed $\mathcal{I}$-oracle in round k, denoted by $\widehat{V}_k(\mathcal{I})$, is defined as the maximizer of the following optimization problem:

$$\begin{aligned}
&\underset{[\tilde{p}(\psi)]_{\psi\in\Psi^+(\mathcal{I})}}{\text{maximize}} \quad \beta \sum_{\psi\in\Psi^+(\mathcal{I})} \tilde{p}(\psi)\widehat{r}_k^*(\psi) - \sum_{i\in\mathcal{I}} c_i \\
&\text{subject to} \quad \sum_{\psi\in\Psi^+(\mathcal{I})} |\tilde{p}(\psi) - \widehat{p}_k(\psi)| \le \mathrm{conf}_{2,k}(\mathcal{I}), \\
&\sum_{\psi\in\Psi^+(\mathcal{I})} \tilde{p}(\psi) = 1. \qquad (5.3)
\end{aligned}$$

At any time t of round k, it can be shown that the optimization in (5.2) can be solved as: $\widehat{h}_k(\psi) = \arg\max_{a\in\mathcal{A}}\ \widehat{r}_k(a,\psi) + \mathrm{conf}_{1,k}(a,\psi)$ and $\widehat{\mathcal{I}}_k = \arg\max_{\mathcal{I}\in\mathcal{P}_{\le m}(\mathcal{D})}\ \widehat{V}_k(\mathcal{I})$. Then, define $\widehat{V}_k^* = \widehat{V}_k(\widehat{\mathcal{I}}_k)$.

The pseudocode for the Sim-OOS is given in Algorithm 7. It can be easily shown that the computational complexity of the Sim-OOS algorithm for T instances is $\mathcal{O}\left(A\,\mathrm{poly}(\Psi_{tot})\log T\right)$.

5.3.3 Regret Bounds for the Sim-OOS algorithm

In this subsection, we provide distribution-independent regret bounds for the Sim-OOS algorithm. Let $\psi_{\mathrm{tot}} = \sum_{\mathcal{I}\in\mathcal{P}_{\le m}(\mathcal{D})} |\Psi^+(\mathcal{I})|$ denote the number of all possible states (all possible results from at most m distinct medical tests).

Theorem 11. *Suppose $\beta = 1$. For any $0 < \delta < 1$, set*

$$\mathrm{conf}_1(n,t) = \min\left(1, \sqrt{\frac{\log\left(20\Psi_{tot}At^5/\delta\right)}{2\max\left(1,n\right)}}\right)$$

and

$$\mathrm{conf}_2(n,t) = \min\left(1, \sqrt{\frac{10\Psi_{tot}\log\left(4t/\delta\right)}{\max\left(1,n\right)}}\right).$$

Then, with probability at least $1-\delta$*, the regret of the Sim-OOS satisfies*

$$\mathrm{Reg}_T^{\textit{Sim-OOS}} = \mathcal{O}\left(\left(\sqrt{A}+\sqrt{|\mathcal{P}_{\leq m}(\mathcal{D})|}\right)\sqrt{\Psi_{tot} T \log\left(T/\delta\right)}\right).$$

The UCRL2 ([JOA10]) is designed for general MDP problems and achieves a regret of $\tilde{O}\left(\sqrt{\Psi_{\text{tot}}^2 AT}\right)$. Hence, these regret results are better than those obtained by UCRL2. This is an important result since it demonstrates that the Sim-OOS can effectively exploit the structure of our CMAB-CO problem to achieve efficient regret bounds which scale better than these that can be obtained for general MDP problems.

We illustrate this bound using the same example above. Suppose $|\mathcal{X}_i| = X$ for all $i \in \mathcal{D}$ and $m = D$. The upper bound given in Theorem 11 is in the order of $\tilde{\mathcal{O}}\left(\sqrt{\sum_{m=1}^{D} X^m 2^D T} + \sqrt{\sum_{m=1}^{D} X^m AT}\right)$

The Sim-OOS algorithm performs well for smaller values of which is the case in the medical setting, as it is for instance the case in breast cancer screening, in which imaging tests are limited to a small set: mammogram, MRI and ultrasound ([SBB07]). In this context, the observations are usually selected sequentially. To address such settings, we next propose the Seq-OOS algorithm that selects observations sequentially.

5.4 Multi-armed Bandits with Sequential Costly Observations

5.4.1 Problem Formulation

Our current setting assumes that decision-maker makes all the observations simultaneously. If the decision-maker is allowed to make observations sequentially, she can use the partial state from already selected observations to inform the selection of future observations. For example, in the medical settings, although a positive result in a medical test is usually followed by additional medical test for validity, a negative result in a medical test is not usually followed by additional medical tests. Since any resulting simultaneous observation policy can be achieved by a sequential observation policy, the oracle defined with sequential observations achieves higher expected reward than that with simultaneous observations. At each time t, the following sequence of events is taking place:

i The decision-maker has initially no observations. In phase 0, we denote the empty partial state as $\psi_{0,t} = \psi_0$ where $\mathrm{dom}(\psi_0) = \emptyset$.

ii At each phase $l \in \mathcal{L} = \{1, \ldots, m\}$, if the partial state is $\psi_{l,t}$ and observation $i_{l,t} \in (\mathcal{D} \setminus \mathrm{dom}(\psi_{l,t})) \cup \emptyset$ is made, the resulting partial state is $\psi_{l+1,t}$ where $\psi_{l+1,t} = \psi_{l,t} \cup (i_{l,t}, \phi_t(i_{l,t}))$ if $i_{l,t} \neq \emptyset$ and $\psi_{l+1,t} = \psi_{l,t}$ otherwise.

iii The decision-maker takes an action a_t when either observation $i_{l,t} = \emptyset$ is made or the final phase m is reached and observes a random reward r_t.

Let $\Psi^+(\psi, i)$ be the set of resulting partial state when observation i is made at previous partial state of ψ, i.e., $\Psi^+(\psi, i) = \{\psi' : \exists x, \psi' = \psi \cup (i, x)\}$. In this section, we define $p(\psi'|\psi, i)$ as the probability of resulting partial state ψ' when the observation i is made at previous partial state of ψ, which is referred to as *partial state transition probability*. For all $\psi' \in \Psi^+(\psi, i)$, the partial state transition probability is defined as $p(\psi'|\psi, i) = \Pr(\Phi(i) = \psi'(i)|\Phi \sim \psi)$ if $i \in \mathcal{D} \setminus \mathrm{dom}(\psi)$ and $p(\psi'|\psi, i) = 0$ otherwise. In the medical example, this is the probability of observing test i's result as $\psi'(i)$ given the previous test results (records) ψ. We define $p(\psi|\psi, \emptyset) = 1$ and $p(\psi'|\psi, \emptyset) = 0$ for all $\psi' \neq \psi$. Let $\boldsymbol{P} = [p(\psi'|\psi, i)]$ denote partial state transition probability matrix.

A sequential policy $\pi = \{g, h\}$ consists of observation function g and action function h where $g : \Psi \to \mathcal{D} \cup \emptyset$ and $h : \Psi \to \mathcal{A}$ (e.g., $g(\psi)$ refers to the next medical test applied on a patient with previous records (test results) ψ and $h(\psi)$ refers to treatment recommendation for a patient with previous records(test results) ψ). A sequential policy $\pi = \{g, h\}$ works as follows. Decision-maker keeps making observations $g(\psi)$ until either m observations are made or an empty observation $g(\psi) = \emptyset$ is picked and takes an action $h(\psi)$ in a terminal state ψ where terminal partial states of policy π is the state with either cardinality m or with $g(\psi) = \emptyset$.

We illustrate these definitions in a medical example. Assume that there are 2 different tests with possible outcomes of positive $(+)$ and negative $(-)$ result and 3 different possible treatments. Suppose that a sequential policy $\pi = (g, h)$ with $g(\emptyset) = \{1\}, g(\{(1, +)\}) = \{2\}, g(\{(1, -)\}) = \emptyset,\ h(\{(1, +), (2, +)\}) = a_1,\ h(\{(1, +), (2, -)\}) = a_2, h(\{(1, -)\}) = a_3.$

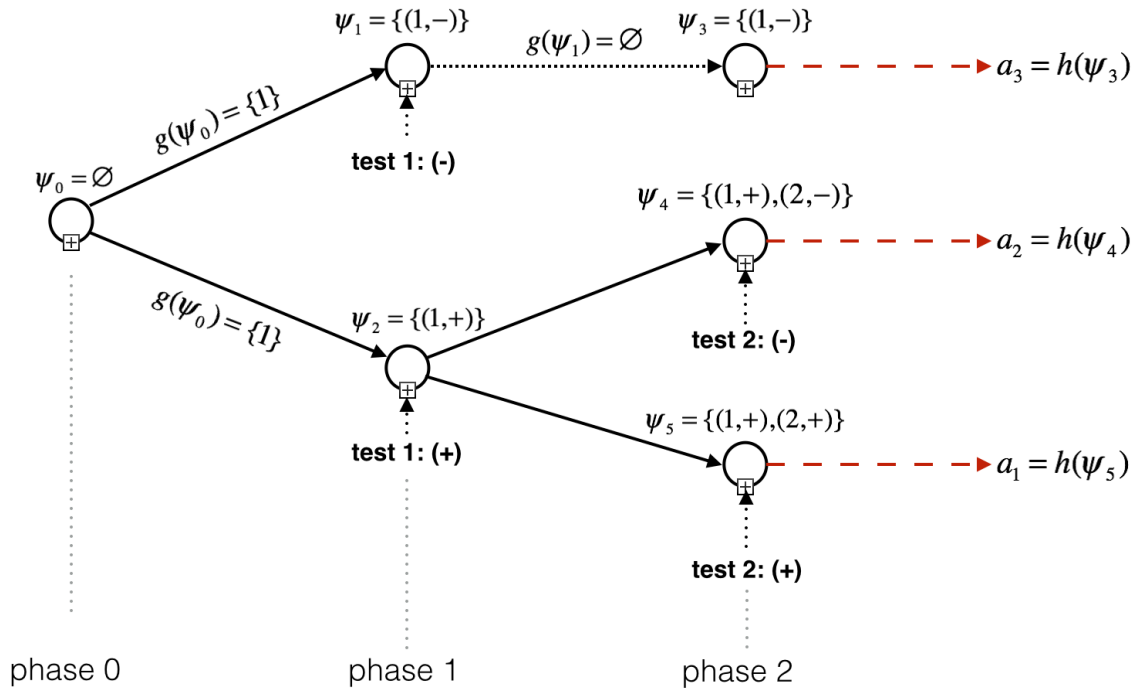


Figure 5.1: Illustration of sequential policy

Basically, this policy initially picks the medical test 1 for all patients ($g(\emptyset) = \{1\}$). If the result of the medical test 1 is positive ($+$), the policy picks medical test 2 ($g(\{(1,+)\}) = \{2\}$). On the other hand, if the result of medical test 1 is negative ($-$), the policy does not make any additional test. In this example, terminal partial states of policy π are ψ_3, ψ_4, ψ_5. This is illustrated in Figure 5.1.

Given a sequential policy π, let ψ_l denote the random partial state in phase l and $c_l = c_{g(\psi_l)}$ denote the random cost in phase l by making observation $g(\psi_l)$. Note that c_l is random since partial state in phase l is random. Similarly, let r_m denote random reward revealed by taking action $a_m = h(\psi_m)$ in terminal partial state. Then, for each sequential policy $\pi = (g, h)$, we define a value function for $l = 0, \ldots, m$:

$$F_l^\pi(\psi) = \mathbb{E}^\pi \left[\beta r_m - \sum_{\tau=l}^{m-1} c_\tau \middle| \psi_l = \psi \right], \tag{5.4}$$

where expectation is taken with respect to randomness of the states and rewards under policy π. In the terminal phase, we define value function as $F_m^\pi(\psi) = \bar{r}(h(\psi), \psi)$. The optimal value function is defined by $F_l^*(\psi) = \sup_\pi F_l^\pi(\psi)$. A policy π^* is said to be optimal

if $F_0^{\pi^*}(\psi) = F_0^*(\psi)$. It is also useful to define partial state-observation optimal value function for $l = 0, \ldots, m-1$:

$$\begin{aligned} Q_l^*(\psi, i) &= \mathbb{E}\left[-c_i + F_{l+1}^*(\psi_{l+1}) | \psi_l = \psi, i_l = i\right] \\ &= -c_i + \sum_{\psi' \in \Psi^+(\psi,i)} p(\psi'|\psi, i) F_{l+1}^*(\psi'). \end{aligned}$$

A sequential policy $\pi^* = (g^*, h^*)$ is optimal if and only if $g^*(\psi) = \arg\max_{i \in (\mathcal{D} \cup \emptyset)} Q_{|\operatorname{dom}(\psi)|}^*(\psi, i)$, $h^*(\psi) = \arg\max_{a \in \mathcal{A}} \bar{r}(a, \psi)$.

Consider a sequential learning algorithm $\pi_{1:T} = (g_t, h_t)_{t=1}^T$. The algorithm makes observation $i_{l,t} = g(\psi_{l,t})$ and realizes a cost $c_{l,t}$ in phase l of time t and then selects action $a_t = h_t(\psi_{m,t})$ and realizes a random reward r_t, which realizes a reward of $r_t - \sum_{l=0}^{m-1} c_{l,t}$. To quantify the performance of sequential learning algorithm, we define cumulative regret of sequential learning algorithm $\pi_{1:T}$ up to time T as

$$\operatorname{Reg}_T^{\pi_{1:T}} = T F_0^* - \sum_{t=1}^{T} \left(r_t - \sum_{l=0}^{m-1} c_{l,t} \right)$$

where $F_0^* = F_0^*(\psi_0)$ and $\psi_0 = \emptyset$ denotes empty state. In the next subsection, we propose a sequential learning algorithm, which aims to minimize regret.

5.4.2 Sequential Optimistic Observation Selection (Seq-OOS)

In addition to observation sets that are tracked by Sim-OOS, Seq-OOS keeps track of the following sets at each round k : $\mathcal{E}_k(\psi, i) = \{\tau < t_k : \exists l \in \mathcal{L},\ \psi_{l,\tau} = \psi, i_{l,t} = i\}$, $\mathcal{E}_k(\psi, i, \psi') = \{\tau < t_k : \exists l \in \mathcal{L},\ \psi_{l,\tau} = \psi, i_{l,\tau} = i, \psi_{l+1,\tau} = \psi'\}$. Let $N_k(\psi, i) = |\mathcal{E}_k(\psi, i)|$ and $N_k(\psi, i, \psi') = |\mathcal{E}_k(\psi, i, \psi')|$. In addition to these counters, we also keep counters of visits in partial state-action pairs and state-observation pairs in a particular round k. Let $\nu_k(\psi, i)$ denote the number of times observation i is made when partial state ψ is realized in round k. We can express the estimated transition probabilities as $\hat{p}_k(\psi'|\psi, i) = \frac{N_k(\psi,i,\psi')}{N_k(\psi,i)}$, provided that $N_k(\psi, i) > 0$.

The Seq-OOS works in rounds $k = 1, \ldots$. In the beginning of round k (t_k denotes time of beginning of round k), the Seq-OOS solves Optimistic Dynamic Programming (ODP), which

takes the estimates $\hat{\boldsymbol{P}}_k = [\hat{p}_k(\psi'|\psi, i)]$ and $\hat{\boldsymbol{R}}_k = [\hat{r}_k(a, \psi)]$ as an input and outputs a policy π_k. The ODP first orders the partial states with respect to size of their domains. Let Ψ_l denote partial states with l observations, which is defined by $\Psi_l = \{\psi : |\operatorname{dom}(\psi)| = l\}$ (e.g., all possible results from l distinct medical tests). Since the decision-maker is not allowed to make any more observations for any state $\psi \in \Psi_m$, estimated value of state ψ is computed by $\hat{F}_{m,k}(\psi) = \max_{a\in\mathcal{A}} \hat{r}_k(a, \psi) + \operatorname{conf}_{1,k}(a, \psi)$ where $\operatorname{conf}_{1,k}(a, \psi)$ is the confidence interval for partial state-action pair in round k. The action and observation functions on partial state $\psi \in \Psi_m$ computed by ODP is given by $\hat{g}_k(\psi) = \emptyset$ and $\hat{h}_k(\psi) = \arg\max_{a\in\mathcal{A}} \hat{r}_k(a, \psi) + \operatorname{conf}_{1,k}(a, \psi)$. After computing value and policy in partial states $\psi \in \Psi_m$, the ODP solves convex optimization problem to compute optimistic value function for each partial state-observation pair $\psi \in \Psi_{m-1}$ and $i \in \mathcal{D} \setminus \operatorname{dom}(\psi)$. Let $\hat{Q}_{m-1,k}(\psi, i)$ denote optimistic value function for making observation i in partial state ψ in round k of phase $m-1$, which is the solution of the following convex optimization problem :

$$\begin{aligned}
&\underset{[\tilde{p}(\cdot|\psi,i)]}{\text{maximize}} \; -c_i + \sum_{\psi'\in\Psi^+(\psi,i)} \tilde{p}(\psi'|\psi, i)\hat{F}_{m,k}(\psi') \\
&\text{subject to} \sum_{\psi'\in\Psi^+(\psi,i)} |\tilde{p}(\psi'|\psi, i) - \hat{p}_k(\psi'|\psi, i)| \le \operatorname{conf}_{2,k}(\psi, i), \\
&\qquad\qquad \sum_{\psi'\in\Psi^+(\psi,i)} \tilde{p}(\psi'|\psi, i) = 1. \qquad (5.5)
\end{aligned}$$

Note that the variables $(\hat{F}_{m,k}(\psi'))$ used in the convex optimization problem given in (5.5) is computed in the previous step by the ODP. The optimistic value of the empty observation $\emptyset$ in partial state ψ in round k is computed by $\hat{Q}_{m-1,k}(\psi, \emptyset) = \max_{a\in\mathcal{A}} \beta\hat{r}_k(a, \psi) + \operatorname{conf}_{1,k}(a, \psi)$. Based on the optimistic value of partial state-observation pairs $\left[\hat{Q}_{m,k}(\psi, i)\right]$, the ODP computes the optimistic value of partial state ψ and action and observation function of partial state $\psi \in \Psi_{m-1}$ as $\hat{F}_{m-1,k}(\psi) = \max_{i\in(\mathcal{D}\setminus\operatorname{dom}(\psi))\cup\emptyset} \hat{Q}_{m-1,k}(\psi, i)$, $\hat{h}_k(\psi) = \arg\max_{a\in\mathcal{A}} \beta\hat{r}_k(a, \psi) + \operatorname{conf}_{1,k}(a, \psi)$, $\hat{g}_k(\psi) = \arg\max_{i\in(\mathcal{D}\setminus\operatorname{dom}(\psi))\cup\emptyset} \hat{Q}_{m-1,k}(\psi, i)$. These computations are repeated for $l = m-2, \ldots, 0$ to find the complete policy $\hat{\pi}_k$.

Given $\hat{\pi}_k = (\hat{g}_k, \hat{h}_k)$, at each time t of round k ($t_{k-1} \le t \le t_k$), the Seq-OOS follows the policy $\hat{\pi}_k$. Basically, if the state at phase l is $\psi_{l,t}$, the Seq-OOS decides to make the observation $i_{l,t} = \hat{g}_k(\psi_{l,t})$ and observes the state $\psi_{l+1,t}$. If the state is $\psi_{l,t}$ at phase $l < m$

and observation $i_{l,t} = \hat{g}_k(\psi_{l,t})$ computed by the ODP is empty set, i.e., $\hat{g}_k(\psi_{l,t}) = \emptyset$, then Seq-OOS takes action $\hat{h}_k(\psi_{l,t})$. If it is a terminal phase, i.e., $l = m$, Seq-OOS takes an action $\hat{h}_k(\psi_{m,t})$.

5.4.3 Regret Bounds of the Seq-OOS

The analysis of the regret of the Seq-OOS exhibits similarities to the analysis of the regret of the Sim-OOS. The Seq-OOS has at most $m + 1$ phases in which it makes observations sequentially followed by an action while Sim-OOS has 2 phases in which it makes simultaneous observations at once followed by an action. The difference is that we need to decompose the regret of the Seq-OOS into regret due to phases with suboptimal observations and regret due to suboptimal actions. Let $\Psi_{\max} = \max_{\psi} \max_{i \in \mathcal{D}} |\Psi^+(\psi, i)|$. The next theorem bounds the distribution-independent regret.

Theorem 12. *Suppose $\beta = 1$. For $0 < \delta < 1$, set*

$$\text{conf}_1(n, t) = \min\left(1, \sqrt{\frac{\log\left(20\Psi_{tot} A t^5/\delta\right)}{2\max\left(1, n\right)}}\right)$$

and

$$\text{conf}_2(n, t) = \min\left(1, \sqrt{\frac{10\Psi_{\max}\log\left(4D\Psi_{tot} t/\delta\right)}{\max\left(1, n\right)}}\right).$$

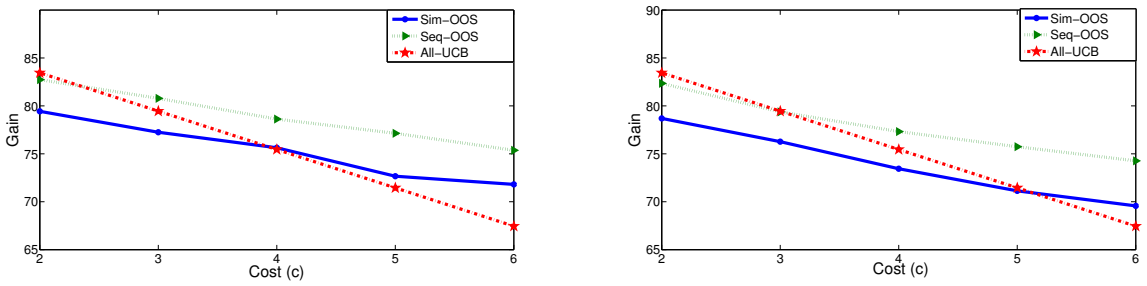
Then, with probability at least $1 - \delta$, regret of the Seq-OOS satisfies

$$\text{Reg}_T^{Seq\text{-}OOS} = \mathcal{O}\left(\left(m\sqrt{\Psi_{max} D} + \sqrt{A}\right)\sqrt{\Psi_{tot} T \log\left(T/\delta\right)}\right)$$

.

The difference in the regret bounds of Sim-OOS and Seq-OOS is because Sim-OOS estimates the observation probabilities $p(\psi)$ for each $\psi \in \Psi$ whereas Seq-OOS estimates observation transition probabilities $p(\cdot|\psi, i)$ for each $\psi \in \Psi$ and $i \in \mathcal{D}$.

Now, we illustrate and compare the regret bounds on our algorithms. Suppose that $|\mathcal{X}_i| = X$ for all $i \in \mathcal{D}$ and $m = D$. In this case, we have the distribution independent regret of $\mathcal{O}\left(2^D\sqrt{AX^D T \log T/\delta}\right)$ for Sim-OOS and $\mathcal{O}\left(D\sqrt{D2^D X^{D+1} AT \log T/\delta}\right)$ for Seq-OOS



(a) $m = 3$ (b) $m = 4$

Figure 5.2: Performance of Sim-OOS and Seq-OOS

with probability at least $1 - \delta$. Our algorithms become computationally feasible when X^D is small.

5.5 Illustrative Results

We evaluate the Sim-OOS and Seq-OOS on a breast cancer dataset (description can be found in 3.7.1). For each instance, the following information about the patient are used: age, estrogen receptor, tumor stage, WHO score. The treatment is a choice among six chemotherapy regimens of which only 4 of them are used: AC, ACT, CAF, CEF. We generate 200000 instances by randomly selecting a sample from the breast cancer dataset. In each instance, we set the observations as $\mathcal{D} = \{\text{age}, \text{estrogen receptor}, \text{tumor stage}, WHOScore\}$, and the rewards as 1 if the treatment with the highest outcome is given to the patient and 0 otherwise. For the experimental results, we set $\beta = 100$ and $m = 3$.

We compare Sim-OOS and Seq-OOS algorithms with a contextual bandit algorithm that observes realization of all observation states ϕ by paying cost of $\sum_{i=1}^{D} c_i$, referred to as Contextual-UCB. We define the following metric of *Gain* of our algorithms ,which make observations $\mathcal{I}_t$ and receives reward of r_t by taking action a_t at each time t, over T time steps by $\text{Gain} = \frac{1}{T}\sum_{t=1}^{T}\left[\beta r_t - \sum_{i\in\mathcal{I}_t} c_i\right]$.

Performance of the Sim-OOS and Seq-OOS with Different Costs: We consider that the cost of each observation $c_i = c$. We illustrate gain of Sim-OOS, Seq-OOS and

Contextual-UCB algorithms for increasing values of cost c. As Figure 1 illustrate, the gain of the Sim-OOS and Seq-OOS algorithm decreases as the observation cost increases. However, it should be noted that these algorithms learn the best simultaneous and sequential policies while simultaneously taking actions irrespective of the costs of observation. These figures show that when the observation cost is increasing, the Sim-OOS and Seq-OOS achieves better gains than Contextual-UCB by observing less information, hence paying less cost. Therefore, the slope of the gain-cost curve of the Sim-OOS and Seq-OOS illustrated in Figure 5.2 decreases as the observation cost increases.

5.6 Proofs

Reduction of (5.2) to convex optimization problem : For notational convenience, we drop the subscript k in this argument. The result of maximization can be found by fixing the observations $\mathcal{I}$ and $\tilde{p}(\cdot)$, and then maximizing with respect to action-function. Let $h^*_{\mathcal{I},\tilde{p}}$ denote the action function which maximizes (5.2). It is easy to see that $h^*_{\mathcal{I},\tilde{p}}(\psi) = \hat{h}^*(\psi) = \arg\max_{a\in\mathcal{A}} \hat{r}(a,\psi) + \text{conf}_{1,k}(a,\psi)$. By fixing h to $\hat{h}^*$ in (5.2), we obtain the following optimization problem:

$$\begin{aligned} &\underset{\mathcal{I}}{\text{maximize}} \quad \sum_{\psi\in\Psi^+(\mathcal{I})} \tilde{p}(\psi)\hat{r}^*(\psi) - \sum_{i\in\mathcal{I}} c_i \\ &\text{subject to} \quad \sum_{\psi\in\Psi^+(\mathcal{I})} |\tilde{p}(\psi) - \hat{p}_k(\psi)| \leq \text{conf}_{2,k}(\mathcal{I}), \\ &\sum_{\psi\in\Psi^+(\mathcal{I})} \tilde{p}(\psi) = 1, \;\; \forall \mathcal{I} \in \mathcal{P}_{\leq m}(\mathcal{D}), \end{aligned} \tag{5.6}$$

We solve (5.6) by first fixing $\mathcal{I}$ and then optimizing the parameters $\tilde{p}$, which results in optimization problem given in (5.3). Denote the result of the optimization problem $\hat{V}(\mathcal{I})$. Then, the optimal observations can be found as $\hat{\mathcal{I}} = \arg\max_{\mathcal{I}\in\mathcal{P}_{\leq m}(\mathcal{D})} \hat{V}^*(\mathcal{I})$.

Proof of Theorem 11

We are going to define some notation before presenting the proofs. Let $N_t(a,\psi) = |\{\tau < t+1 : a_\tau = a, \psi_\tau = \psi\}|$ and $\boldsymbol{N}_t = [N_t(a,\psi)]$ denote matrix whose elements are the number of times of partial state-action pair is observed. Let K denote the total number of rounds

until time T and $k(t)$ denote the round time t belongs to. Let τ_k denote the length of round k, i.e., $\tau_k = t_{k+1} - t_k$. Based on these, we have the following: $N_T(a,\psi) = \sum_{i=1}^{K} \nu_k(a,\psi)$, $N_k(a,\psi) = \sum_{i=1}^{k-1} \nu_k(a,\psi)$.

For round k, let $\mathcal{P}_k$ be set of observation/transition probabilities matrices and $\mathcal{R}_k$ be the set of mean rewards of partial state-action pairs, which satisfy the constraints in optimization problem of (5.2). We define the expected gain of action a of state ψ as the expected reward of action a minus the observation cost of state as $\mu(a,\psi) = \bar{r}(a,\psi) - \sum_{i\in\text{dom}(\psi)} c_i$.

In round of k, let $\tilde{r}_k(a,\psi) = \hat{r}_k(a,\psi) + \text{conf}_{1,k}(a,\psi)$ be the optimistic reward estimate of action-partial state pair (a,ψ). Let $\tilde{p}_k(\psi)$ is the maximizer of the optimization problem given in (5.3) (optimistic observation probability estimate of partial state ψ). Based on this, we define

$$\tilde{\mu}_k(a,\psi) = \tilde{r}_k(a,\psi) - \sum_{i\in\text{dom}(\psi)} c_i$$

$$\tilde{\boldsymbol{P}}_k = [\tilde{p}_k(\psi)]_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)}\,,\ \boldsymbol{P}_k = [p(\psi)]_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)}$$

$$\hat{\boldsymbol{P}}_k = [p(\psi)]_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)}\,,\ \tilde{\boldsymbol{\mu}}_k = \left[\tilde{\mu}_k(\hat{h}_k(\psi),\psi)\right]_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)}$$

$$\boldsymbol{\mu}_k = \left[\mu_k(\hat{h}_k(\psi),\psi)\right]_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)}\,,\ \boldsymbol{\nu}_k = \left[\nu_k(\hat{h}_k(\psi),\psi)\right]_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)}$$

We define optimistic reward in round k as $\tilde{V}_k^*$, which is the result of the optimization problem in (5.2). Based on definitions above, $\tilde{V}_k^* = \langle\tilde{\boldsymbol{P}}_k, \tilde{\boldsymbol{\mu}}_k\rangle$ where $\langle\cdot,\cdot\rangle$ denotes the dot product between two vectors.

Also, let $\boldsymbol{e}_{k,\psi}$ be unit vector with size of $|\Psi^+(\hat{\mathcal{I}}_k)|$ with the index representing state ψ is being equal to 1 and 0 else. Let $\mathcal{B}_k$ denote the event that optimistic rewards and observation probabilities satisfy the constraints of the optimization problem in (5.2), i,e., $\mathcal{B}_k = \left((\widehat{\boldsymbol{P}}_k, \widehat{\boldsymbol{R}}_k) \in (\mathcal{P}_k, \mathcal{R}_k)\right)$ and $\bar{\mathcal{B}}_k$ is the complement of the event $\mathcal{B}_k$. Also, we denote $\mathbb{I}(\mathcal{B}_k)$ as the indicator being equal to 1 if event $\mathcal{B}_k$ happens. Observe that when event $\mathcal{B}_k$ happens, we have $\tilde{V}_k^* \geq V^*$ with probability 1.

The main steps of the proof is to decompose the regret in the regimes where confidence intervals are achieved and violated. We show that regret is small when the confidence inter-

vals are achieved. We also show that confidence bounds are satisfied with high probability. By combining these two results, we prove our regret bounds.

Step 1 (Regret Decomposition) : We have

$$\begin{aligned}
\mathrm{Reg}_T^{\text{Sim-OOS}} &= TV^* - \sum_{t=1}^{T}\left(r_t - \sum_{i\in\mathcal{I}_t} c_i\right) \\
&= \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} N_T(a,\psi)\left(V^* - \mu(a,\psi)\right) \\
&+ \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} N_T(a,\psi)\bar{r}(a,\psi) - \sum_{t=1}^{T} r_t \\
&= \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \nu_k(a,\psi)\left(V^* - \mu(a,\psi)\right)\mathbb{I}(\mathcal{B}_k) \qquad (5.7) \\
&+ \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \nu_k(a,\psi)\left(V^* - \mu(a,\psi)\right)\mathbb{I}(\bar{\mathcal{B}}_k) \\
&+ \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} N_T(a,\psi)\bar{r}(a,\psi) - \sum_{t=1}^{T} r_t
\end{aligned}$$

Given the set of observations $\hat{\mathcal{I}}_k$ and action function $\hat{h}_k(\cdot)$, $\nu_k(\cdot,\cdot)$ is only non-zero for the form of action-partial state $(\hat{h}_k(\psi),\psi)_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)}$. Therefore, equation (5.7) can be further decomposed into

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \nu_k(a,\psi)\left(V^* - \mu(a,\psi)\right)\mathbb{I}(\mathcal{B}_k) \\
&\overset{a}{\leq} \sum_{k=1}^{K}\sum_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)} \nu_k(\hat{h}_k(\psi),\psi)\left(\tilde{V}_k^* - \tilde{\mu}_k(\hat{h}_k(\psi),\psi)\right)\mathbb{I}(\mathcal{B}_k) \\
&+ \sum_{k=1}^{K}\sum_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)} \nu_k(\hat{h}_k(\psi),\psi)\left(\tilde{\mu}_k(\hat{h}_k(\psi),\psi) - \mu(\hat{h}_k(\psi),\psi)\right)\mathbb{I}(\mathcal{B}_k) \\
&\overset{b}{=} \sum_{k=1}^{K} \tau_k\langle(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k), \tilde{\boldsymbol{\mu}}_k\rangle\mathbb{I}(\mathcal{B}_k) \\
&+ \sum_{k=1}^{K}\sum_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)} \nu_k(\hat{h}_k(\psi),\psi)\langle(\boldsymbol{P}_k - \boldsymbol{e}_{k,\psi}), \tilde{\boldsymbol{\mu}}_k\rangle\mathbb{I}(\mathcal{B}_k) \\
&+ \sum_{k=1}^{K}\langle\boldsymbol{v}_k, (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k)\rangle\mathbb{I}(\mathcal{B}_k).
\end{aligned}$$

where (a) follows from $\tilde{V}_k^* \geq V^*$ when event $\mathcal{B}_k$ happens and (b) follows from $\tilde{\rho}_m(k) - \tilde{\mu}_k(\hat{h}_k(\psi), \psi) = \langle \tilde{\boldsymbol{P}}_k, \tilde{\boldsymbol{\mu}}_k \rangle - \langle \boldsymbol{e}_{k,\psi}, \tilde{\boldsymbol{\mu}}_k \rangle = \langle (\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k), \tilde{\boldsymbol{\mu}}_k \rangle + \langle (\boldsymbol{P}_k - \boldsymbol{e}_{k,\psi}), \tilde{\boldsymbol{\mu}}_k \rangle$. Then, regret decomposition is

$$
\begin{aligned}
&\mathrm{Reg}_T^{\text{Sim-OOS}} \\
&\leq \sum_{a \in \mathcal{A}} \sum_{\psi \in \Psi} N_T(a, \psi) \bar{r}(a, \psi) - \sum_{t=1}^{T} r_t && (5.8) \\
&+ \sum_{k=1}^{K} \sum_{a \in \mathcal{A}} \sum_{\psi \in \Psi} \nu_k(a, \psi) \left(V^* - \mu(a, \psi) \right) \mathbb{I}(\bar{\mathcal{B}}_k) && (5.9) \\
&+ \sum_{k=1}^{K} \tau_k \langle (\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k), \tilde{\boldsymbol{\mu}}_k \rangle \mathbb{I}(\mathcal{B}_k) && (5.10) \\
&+ \sum_{k=1}^{K} \sum_{\psi \in \Psi^+(\hat{\mathcal{I}}_k)} \nu_k(\hat{h}_k(\psi), \psi) \langle (\boldsymbol{P}_k - \boldsymbol{e}_{k,\psi}), \tilde{\boldsymbol{\mu}}_k \rangle \mathbb{I}(\mathcal{B}_k) && (5.11) \\
&+ \sum_{k=1}^{K} \langle \boldsymbol{v}_k, (\tilde{\boldsymbol{\mu}}_k - \boldsymbol{\mu}_k) \rangle \mathbb{I}(\mathcal{B}_k) && (5.12)
\end{aligned}
$$

Step 2 (Regret due to randomness of the rewards) : Observe that

$$
\mathbb{E}\left[\sum_{\tau=1}^{T} r_\tau \middle| \boldsymbol{N}_T \right] = \sum_{a \in \mathcal{A}} \sum_{\psi \in \Psi} N_T(a, \psi) \bar{r}(a, \psi). \tag{5.13}
$$

Given counts in observations of state-action pairs $\boldsymbol{N}_T$, the random variables $r_t(a_t)$ are independent over time t. Then,

$$
\begin{aligned}
&\Pr\left(\sum_{\tau=1}^{T} r_\tau \leq \sum_{a \in \mathcal{A}} \sum_{\psi \in \Psi} N_T(a, \psi) \bar{r}(a, \psi) - \sqrt{\frac{T}{2} \log \frac{4T}{\delta}} \middle| \boldsymbol{N}_T \right) \\
&\overset{c}{\leq} \exp\left(-2 \frac{1}{2T} \log\left(\frac{4T}{\delta} \right) T \right) \leq \frac{\delta}{4T},
\end{aligned}
$$

where (c) follows from Hoeffding's inequality. Therefore, the equation (5.8) is bounded with probability $1 - \frac{\delta}{4T}$ by

$$
\sum_{a \in \mathcal{A}} \sum_{\psi \in \Psi} N_T(a, \psi) \bar{r}(a, \psi) - \sum_{t=1}^{T} r_t \leq \sqrt{\frac{T}{2} \log \frac{4T}{\delta}}.
$$

Step 3 (Regret due to Failure of Confidence Intervals): In this step, consider the regret of the episodes in which $\bar{\mathcal{B}}_k$. Define $\mathcal{P}_t$, $\mathcal{R}_t$ as the set of plausible observation/transition probability matrices and set of plausible mean rewards of partial state-action pairs using the estimates of time step t and $\widehat{\boldsymbol{P}}_t$ and $\widehat{\boldsymbol{R}}_t$ as the estimates in time step t. Then,

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}\nu_k(a,\psi)\left(V^*-\mu(a,\psi)\right)\mathbb{I}\left(\bar{\mathcal{B}}_k\right)\\
&\leq 2\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}\nu_k(a,\psi)\mathbb{I}\left(\bar{\mathcal{B}}_k\right)\\
&\leq 2\sum_{k=1}^{K}t_k\mathbb{I}\left(\bar{\mathcal{B}}_k\right)\leq 2\sum_{t=1}^{T}t\mathbb{I}\left(\left(\widehat{\boldsymbol{P}}_t,\widehat{\boldsymbol{R}}_t\right)\notin\left(\mathcal{P}_t,\mathcal{R}_t\right)\right)\\
&\leq 2\sqrt{T}+2\sum_{t=\lfloor T^{1/4}\rfloor+1}^{T}t\mathbb{I}\left(\left(\widehat{\boldsymbol{P}}_t,\widehat{\boldsymbol{R}}_t\right)\notin\left(\mathcal{P}_t,\mathcal{R}_t\right)\right)
\end{aligned}$$

Then,

$$\begin{aligned}
&\Pr\left(\sum_{t=\lfloor T^{1/4}\rfloor+1}^{T}t\mathbb{I}\left(\left(\widehat{\boldsymbol{P}}_t,\widehat{\boldsymbol{R}}_t\right)\notin\left(\mathcal{P}_t,\mathcal{R}_t\right)\right)>0\right)\\
&\leq\Pr\left(\exists t:\lfloor T^{1/4}\rfloor+1\leq t\leq T\text{ s.t. }\left(\widehat{\boldsymbol{P}}_t,\widehat{\boldsymbol{R}}_t\right)\notin\left(\mathcal{P}_t,\mathcal{R}_t\right)\right)\\
&\leq\sum_{t=\lfloor T^{1/4}\rfloor+1}^{T}\Pr\left(\left(\widehat{\boldsymbol{P}}_t,\widehat{\boldsymbol{R}}_t\right)\notin\left(\mathcal{P}_t,\mathcal{R}_t\right)\right)\overset{d}{\leq}\sum_{t=\lfloor T^{1/4}\rfloor+1}^{T}\frac{\delta}{5t^5}\\
&\leq\frac{\delta}{5T^{5/4}}+\int_{\lfloor T^{1/4}\rfloor+1}^{\infty}\frac{\delta}{5t^5}\,dt\leq\frac{\delta}{5T^{5/4}}+\frac{\delta}{20T}\leq\frac{\delta}{4T}
\end{aligned}$$

where (d) follows from Lemma 13.

Therefore, with probability at least $1-\frac{\delta}{4T}$, we have

$$\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}\nu_k(a,\psi)\left(V^*-\mu(a,\psi)\right)\mathbb{I}\left(\bar{\mathcal{B}}_k\right)\leq 2\sqrt{T}$$

Step 4 (Regret due to estimation error of transition probabilities): We start

with some more definitions and observations. For any k, we have

$$\begin{aligned}
&\langle(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k), \tilde{\boldsymbol{\mu}}_k\rangle = \langle(\tilde{\boldsymbol{P}}_k - \hat{\boldsymbol{P}}_k), \tilde{\boldsymbol{\mu}}_k\rangle + \langle(\hat{\boldsymbol{P}}_k - \boldsymbol{P}_k), \tilde{\boldsymbol{\mu}}_k\rangle \\
&\overset{e}{\leq} (\|(\tilde{\boldsymbol{P}}_k - \hat{\boldsymbol{P}}_k)\|_1 + \|\hat{\boldsymbol{P}}_k - \hat{\boldsymbol{P}}_k\|_1)\|\tilde{\boldsymbol{\mu}}_k\|_\infty \\
&\leq \sqrt{\frac{160\Psi_{\text{tot}} \log 4T/\delta}{\max(1, N_k(\hat{\mathcal{I}}_k))}}
\end{aligned} \tag{5.14}$$

where (e) holds when event $\mathcal{B}_k$ happens and follows from $\|\tilde{\boldsymbol{\mu}}_k\|_\infty < 2$. Let lengths of the rounds with $\hat{\mathcal{I}}_k = \mathcal{I}$ be denoted as the sequence $(\tau_1(\mathcal{I}), \tau_2(\mathcal{I}), \ldots, \tau_{K(\mathcal{I})}(\mathcal{I}))$ where $K(\mathcal{I})$ is the number of rounds with $\hat{\mathcal{I}}_k = \mathcal{I}$, i.e., $K(\mathcal{I}) = |\{1 \leq k \leq K : \hat{\mathcal{I}}_k = \mathcal{I}\}|$ and $n_k(\mathcal{I}) = \sum_{i=1}^{k-1} \tau_i(\mathcal{I})$ is the number of times observations $\mathcal{I}$ are made in the rounds $(\tau_1(\mathcal{I}), \tau_2(\mathcal{I}), \ldots, \tau_{k-1}(\mathcal{I}))$. We have $N_k(\mathcal{I}) \geq n_k(\mathcal{I})$ since $N_k(\mathcal{I})$ is the number of the observations that contain $\mathcal{I}$ are made. We have also $T = \sum_{\mathcal{I} \in \mathcal{P}_{\leq m}(\mathcal{D})} n_{K(\mathcal{I})}(\mathcal{I})$. Then, equation (5.10) is bounded by

$$\begin{aligned}
&\sum_{k=1}^{K} \tau_k \langle(\tilde{\boldsymbol{P}}_k - \boldsymbol{P}_k), \tilde{\boldsymbol{\mu}}_k\rangle \mathbb{I}(\mathcal{B}_k) \\
&\leq \sqrt{160\Psi_{\text{tot}} \log 4T/\delta} \sum_{k=1}^{K} \frac{\tau_k}{\sqrt{\max\left(1, N_k(\hat{\mathcal{I}}_k)\right)}} \\
&\leq \sqrt{160\Psi_{\text{tot}} \log 4T/\delta} \sum_{\mathcal{I} \in \mathcal{P}_{\leq m}(\mathcal{D})} \sum_{k=1}^{K(\mathcal{I})} \frac{\tau_k(\mathcal{I})}{\sqrt{\max\left(1, n_k(\mathcal{I})\right)}} \\
&\overset{f}{\leq} (1+\sqrt{2})\sqrt{160\Psi_{\text{tot}} \log 4T/\delta} \sum_{\mathcal{I} \in \mathcal{P}_{\leq m}(\mathcal{D})} \sqrt{N_{K(\mathcal{I})}(\mathcal{I})} \\
&\overset{g}{\leq} (1+\sqrt{2})\sqrt{160\Psi_{\text{tot}} |\mathcal{P}_{\leq m}(\mathcal{D})| T \log 4T/\delta}
\end{aligned}$$

where (f) follows from Lemma 15 and (g) from Jensen's inequality.

Step 5(Regret due to randomness due to transition probabilities): Let $X_t = \langle(\boldsymbol{P}_{k(t)} - \boldsymbol{e}_{k(t),\psi_t}), \tilde{\boldsymbol{\mu}}_{k(t)}\rangle \mathbb{I}(\mathcal{B}_{k(t)})$. The term (5.11) can be written as

$$\sum_{k=1}^{K} \sum_{\psi \in \Psi^+(\hat{\mathcal{I}}_k)} \nu_k(\hat{h}_k(\psi), \psi) \langle(\boldsymbol{P}_k - \boldsymbol{e}_{k,\psi})\tilde{\boldsymbol{\mu}}_k\rangle \mathbb{I}(\mathcal{B}_k) = \sum_{t=1}^{T} X_t.$$

Observe that

$$\begin{aligned}&\mathbb{E}\left[X_t|(\mathcal{I}_1,\psi_1,a_1,r_1),\ldots,(\mathcal{I}_{t-1},\psi_{t-1},a_{t-1},r_{t-1}),\mathcal{I}_t\right]\\&=\langle\mathbb{E}\left[\boldsymbol{P}_{k(t)}-\boldsymbol{e}_{k(t),\psi_t}\right],\tilde{\boldsymbol{\mu}}_{k(t)}\rangle\mathbb{I}(\mathcal{B}_{k(t)})=0\end{aligned}$$

and

$$|X_t|\leq\left(\|\boldsymbol{P}_{k(t)}\|_1+\|\boldsymbol{e}_{k(t),\psi_t}\|_1\right)\|\tilde{\boldsymbol{\mu}}_{k(t)}\|_\infty\mathbb{I}(\mathcal{B}_{k(t)})\leq 4.$$

By Azuma-Hoeffding bound, we have

$$\Pr\left(\sum_{t=1}^{T}X_t\geq 4\sqrt{2T\log 4T/\delta}\right)\leq\frac{\delta}{4T}.$$

With probability at least $1-\frac{\delta}{4T}$, we have

$$\begin{aligned}&\sum_{k=1}^{K}\sum_{\psi\in\Psi^+(\hat{\mathcal{I}}_k)}\nu_k(\hat{h}_k(\psi),\psi)\left(\boldsymbol{P}_k-\boldsymbol{e}_{k,\psi}\right)^T\tilde{\boldsymbol{\mu}}_k\mathbb{I}(\mathcal{B}_k)\\&\leq 4\sqrt{2T\log 4T/\delta}.\end{aligned}$$

Step 6(Regret due to estimation error of rewards): Observe that when $\mathcal{B}_k$ happens,

$$\begin{aligned}&\tilde{\mu}_k(\hat{h}_k(\psi),\psi)-\mu(\hat{h}_k(\psi),\psi)=\tilde{r}_k(\hat{h}_k(\psi),\psi)-\bar{r}(\hat{h}_k(\psi),\psi)\\&=\tilde{r}_k(\hat{h}_k(\psi),\psi)-\hat{r}_k(\hat{h}_k(\psi),\psi)\\&\qquad+\hat{r}_k(\hat{h}_k(\psi),\psi)-\bar{r}(\hat{h}_k(\psi),\psi)\\&\leq|\tilde{r}_k(\hat{h}_k(\psi),\psi)-\hat{r}_k(\hat{h}_k(\psi),\psi)|\\&\qquad+|\hat{r}_k(\hat{h}_k(\psi),\psi)-\bar{r}(\hat{h}_k(\psi),\psi)|\\&\leq\sqrt{\frac{2\log\left(20\Psi_{\text{tot}}At^5/\delta\right)}{N_k(\hat{h}_k(\psi),\psi)}}\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\sum_{k=1}^{K}\langle\boldsymbol{v}_k,(\tilde{\boldsymbol{\mu}}_k-\boldsymbol{\mu}_k)\rangle\mathbb{I}(\mathcal{B}_k)\\&\leq\sqrt{2\log\left(20\Psi_{\text{tot}}AT/\delta\right)}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}\sum_{k=1}^{K}\frac{\nu_k(a,\psi)}{\max\left(1,\sqrt{N_k(a,\psi)}\right)}.\end{aligned}$$

Since $N_k(a,\psi) = \sum_{i=1}^{k-1} \nu_k(a,\psi)$, by Lemma 15, we have

$$\sum_{k=1}^{K} \frac{\nu_k(a,\psi)}{\max\left(1, \sqrt{N_k(a,\psi)}\right)} \leq (1+\sqrt{2})\sqrt{N_T(a,\psi)}.$$

By Jensen's inequality, we have

$$\sqrt{2\log\left(20\Psi_{\text{tot}}AT/\delta\right)} \sum_{a\in\mathcal{A}} \sum_{\psi\in\Psi} \sum_{k=1}^{K} \frac{\nu_k(a,\psi)}{\max\left(1, \sqrt{N_k(a,\psi)}\right)}$$
$$\leq (1+\sqrt{2})\sqrt{2\Psi_{\text{tot}}AT\log\left(20\Psi_{\text{tot}}AT/\delta\right)}$$

Therefore, with probability at least $1-\delta$, we have

$$\text{Reg}_T^{\text{Sim-OOS}} \leq \sqrt{\frac{T}{2}\log 4T/\delta} + 4\sqrt{2T\log 4T/\delta} + 2\sqrt{T}$$
$$+ (1+\sqrt{2})\Bigg(\sqrt{160|\mathcal{P}_{\leq m}(\mathcal{D})|T\log 4T/\delta}$$
$$+ \sqrt{2\Psi_{\text{tot}}AT\log 20\Psi_{\text{tot}}AT/\delta}\Bigg)$$

Proof of Theorem 12

Let us define the following notation that is used throughout proof of Theorem 12:

$$\tilde{\boldsymbol{P}}_k(\cdot|\psi,i) = \left[\tilde{p}_k(\psi'|\psi,i)\right]_{\psi'\in\Psi^+(\psi,i)},$$
$$\hat{\boldsymbol{P}}_k(\cdot|\psi,i) = \left[\hat{p}_k(\psi'|\psi,i)\right]_{\psi'\in\Psi^+(\psi,i)},$$
$$\boldsymbol{P}(\cdot|\psi,i) = \left[p(\psi'|\psi,i)\right]_{\psi'\in\Psi^+(\psi,i)},$$
$$\hat{\boldsymbol{F}}_{l,k} = \left[\hat{F}_{l,k}(\psi)\right]_{\psi\in\Psi_l},$$

for any $l = 0,\ldots,m$. We need to define the event $\mathcal{B}_k$ as the event that reward and transition probability estimates achieve the confidence levels. Formally, it is defined as

$$\mathcal{B}_k = \{\psi\in\Psi,\ a\in\mathcal{A} : |\bar{r}(a,\psi) - \hat{r}_k(a,\psi)| \leq \text{conf}_{1,k}(a,\psi)\}$$
$$\cap\{\psi\in\Psi,\ i\in i\in\mathcal{D}\setminus\text{dom}(\psi) :$$
$$\|\boldsymbol{P}(\cdot|\psi,i) - \hat{\boldsymbol{P}}_k(\cdot|\psi,i)\|_1 \leq \text{conf}_{2,k}(\psi,i)\}.$$

Note that when event $\mathcal{B}_k$ happens, we have $\hat{F}_{0,k}(\psi_0) \geq F_0^*(\psi_0)$. Let $\Psi_{\pi,l}^+$ denote the set of realizations in phase l under policy π. Therefore, $\Psi_{\pi,m}^+$ denotes the set of terminal realizations

under policy π. Let $\tilde{p}_k(\psi'|\psi, i)$ denote the optimistic observation transition probabilities from state ψ to ψ' by observation i (the maximizer of the optimization problem given in (5.5)).

Step 1 (Regret Decomposition) : We can show by following the same as the step 1 of the proof of Theorem 11 that regret of the Seq-OOS can be decomposed as

$$\begin{aligned}
\mathrm{Reg}_T^{\text{Seq-OOS}} &= TF_0^*(\psi_0) - \left[\sum_{t=1}^{T}\left(r_t - \sum_{\tau=0}^{m-1} c_\tau\right)\right] \\
&= \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \nu_k(a,\psi)\left[F_0^*(\psi_0) - \mu(a,\psi)\right] \\
&+ \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} N_T(a,\psi)\bar{r}(a,\psi) - \sum_{t=1}^{T} r_t \\
&= \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \nu_k(a,\psi)\left[F_0^*(\psi_0) - \mu(a,\psi)\right]\mathbb{I}(\mathcal{B}_k) \qquad (5.15) \\
&+ \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \nu_k(a,\psi)\left[F_0^*(\psi_0) - \mu(a,\psi)\right]\mathbb{I}(\bar{\mathcal{B}}_k) \\
&+ \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} N_T(a,\psi)\bar{r}(a,\psi) - \sum_{t=1}^{T} r_t
\end{aligned}$$

Equation (5.15) can be bounded and decomposed as

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \nu_k(a,\psi)\left[F_0^*(\psi_0) - \mu(a,\psi)\right]\mathbb{I}(\mathcal{B}_k) \\
&\leq \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \nu_k(a,\psi)\left[\hat{F}_{0,k}(\psi_0) - \tilde{\mu}_k(a,\psi)\right]\mathbb{I}(\mathcal{B}_k) \qquad (5.16) \\
&+ \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \nu_k(a,\psi)\left[\tilde{r}_k(a,\psi) - \bar{r}(a,\psi)\right]\mathbb{I}(\mathcal{B}_k). \qquad (5.17)
\end{aligned}$$

We can further decompose equation (5.16). Observe that $\nu_k(\cdot,\cdot)$ is non zero only for partial state-action pairs of the form $(\hat{h}_k(\psi),\psi)_{\psi\in\Psi^+_{\tilde{\pi}_k,m}}$. Therefore, we can rewrite equation (5.16) as

$$\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}\nu_k(a,\psi)\left[\hat{F}_{0,k}(\psi_0)-\tilde{\mu}(a,\psi)\right]\mathbb{I}(\mathcal{B}_k)$$
$$=\sum_{k=1}^{K}\sum_{\psi\in\Psi^{+}_{\tilde{\pi}_k,m}}\nu_k(\hat{h}_k(\psi),\psi)\left(\hat{F}_{0,k}(\psi_0)-\tilde{\mu}(\hat{h}_k(\psi),\psi)\right)\mathbb{I}(\mathcal{B}_k)$$
$$=\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\left(\hat{F}_{0,k}(\psi_0)-\tilde{\mu}(\hat{h}_k(\psi_{m,t}),\psi_{m,t})\right)\mathbb{I}(\mathcal{B}_k) \tag{5.18}$$

We also have $\hat{F}_{l,k}(\psi)=\hat{Q}_{l,k}(\psi,\hat{g}_k(\psi))$ for all $\psi\in\Psi_l$ and $l=0,\ldots,m-1$ and $\hat{F}_{m,k}(\psi)=\tilde{r}(\hat{h}_k(\psi),\psi)=\tilde{\mu}(\hat{h}_k(\psi),\psi)+\sum_{i\in\mathrm{dom}(\psi)}c_i$ by their definitions. Therefore, we can rewrite equation (5.17) as

$$\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\left(\hat{F}_{0,k}(\psi_0)-\tilde{\mu}(\hat{h}_k(\psi_{m,t}),\psi_{m,t})\right)\mathbb{I}(\mathcal{B}_k)$$
$$=\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\Bigg(\hat{Q}_{0,k}(\psi_0,\hat{g}_k(\psi_0))-\hat{F}_{m,k}(\psi_{m,t})+\sum_{l=0}^{m-1}c_{\hat{g}_k(\psi_{l,t})}\Bigg)\mathbb{I}(\mathcal{B}_k)$$
$$=\sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\sum_{l=0}^{m-1}\Bigg(\hat{Q}_{l,k}(\psi_{l,t},\hat{g}_k(\psi_{l,t}))-\hat{F}_{l,k}(\psi_{l+1,t})+c_{\hat{g}_k(\psi_{l,t})}\Bigg)\mathbb{I}(\mathcal{B}_k).$$

By definition, we have

$$\hat{Q}_{l,k}(\psi,i)=\sum_{\psi'\in\Psi^{+}(\psi,i)}\tilde{p}_k(\psi'|\psi,i)\hat{F}_{l+1,k}(\psi')-c_i \tag{5.19}$$

Then,

$$\begin{aligned}
&\sum_{t=t_k}^{t_{k+1}-1}\sum_{l=0}^{m-1}\Big(\hat{Q}_{l,k}(\psi_{l,t},\hat{g}_k(\psi_{l,t})) \\
&\qquad\qquad - \hat{F}_{l,k}(\psi_{l+1,t}) + c_{\hat{g}_k(\psi_{l,t})}\Big)\mathbb{I}(\mathcal{B}_k) \\
&= \sum_{t=t_k}^{t_{k+1}-1}\sum_{l=0}^{m-1}\langle(\tilde{\boldsymbol{P}}_k(\cdot|\psi_{l,t},\hat{g}_k(\psi_{l,t})) - \boldsymbol{e}_{k,\psi_{l+1,t}}),\hat{\boldsymbol{F}}_{l,k}\rangle\mathbb{I}(\mathcal{B}_k) \\
&= \sum_{l=0}^{m-1}\sum_{\psi\in\Psi_l}\sum_{i\in\mathcal{D}\setminus\mathrm{dom}(\psi)}\nu_k(\psi,i)\langle(\tilde{\boldsymbol{P}}_k(\cdot|\psi,i) \\
&\qquad\qquad - \boldsymbol{P}(\cdot|\psi,i)),\hat{\boldsymbol{F}}_{l,k}\rangle\mathbb{I}(\mathcal{B}_k) \\
&+ \sum_{t=t_k}^{t_{k+1}-1}\sum_{l=0}^{m-1}\langle(\boldsymbol{P}(\cdot|\psi_{l,t},\hat{g}_k(\psi_{l,t})) - \boldsymbol{e}_{\psi_{l+1,t}}),\hat{\boldsymbol{F}}_{l,k}\rangle\mathbb{I}(\mathcal{B}_k).
\end{aligned}$$

Therefore, regret can be decomposed as

$$\begin{aligned}
&\mathrm{Reg}_T^{\text{Seq-OOS}} \\
&\leq \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}N_T(a,\psi)\bar{r}(a,\psi) - \sum_{t=1}^{T}r_t && (5.20)\\
&+ \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}\nu_k(a,\psi)(F_0^*(\psi_0)-\mu(a,\psi))\mathbb{I}(\bar{\mathcal{B}}_k) && (5.21)\\
&+ \sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}\nu_k(a,\psi)(\tilde{\mu}_k(a,\psi)-\mu(a,\psi))\mathbb{I}(\mathcal{B}_k) && (5.22)\\
&+ \sum_{k=1}^{K}\sum_{l=0}^{m-1}\sum_{\psi\in\Psi_l}\sum_{i\in\mathcal{D}\setminus\mathrm{dom}(\psi)}\nu_k(\psi,i)\langle(\tilde{\boldsymbol{P}}_k(\cdot|\psi,i) \\
&\qquad\qquad - \boldsymbol{P}(\cdot|\psi,i)),\hat{\boldsymbol{F}}_{l,k}\rangle\mathbb{I}(\mathcal{B}_k) && (5.23)\\
&+ \sum_{k=1}^{K}\sum_{t=t_k}^{t_{k+1}-1}\sum_{l=0}^{m-1}\langle(\boldsymbol{P}(\cdot|\psi_{l,t},\hat{g}_k(\psi_{l,t})) \\
&\qquad\qquad - \boldsymbol{e}_{\psi_{l+1,t}}),\hat{\boldsymbol{F}}_{l,k}\rangle\mathbb{I}(\mathcal{B}_k). && (5.24)
\end{aligned}$$

Step 2 (Regret due to randomness of the rewards) : We can show by following the same as the step 2 of the proof of Theorem 11 that equation (5.8) with probability $1-\frac{\delta}{4T}$

as

$$\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} N_T(a,\psi)\bar{r}(a,\psi) - \sum_{t=1}^{T} r_t \leq \sqrt{\frac{T}{2}\log(4T/\delta)}.$$

Step 3 (Regret due to Violation of Confidence Intervals) : The equation (5.21) can be bounded by

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}\nu_k(a,\psi)(F_0^*(\psi_0)-\mu(a,\psi))\mathbb{I}(\bar{\mathcal{B}}_k)\\
&\leq 2\sqrt{T} + \sum_{t=\lfloor T^{1/4}\rfloor+1}^{T} t\mathbb{I}\left((\widehat{\boldsymbol{P}}_t,\widehat{\boldsymbol{R}}_t)\notin(\mathcal{P}_t,\mathcal{R}_t)\right).
\end{aligned}$$

where $\Pr\left((\widehat{\boldsymbol{P}}_t,\widehat{\boldsymbol{R}}_t)\notin(\mathcal{P}_t,\mathcal{R}_t)\right)\leq\frac{\delta}{5t^5}$. Therefore, following the same steps as the Step 3 of proof of Theorem 11, with probability at least $1-\frac{\delta}{4T}$

$$\sum_{k=1}^{K}\sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi}\nu_k(a,\psi)(F_0^*(\psi_0)-\mu(a,\psi))\mathbb{I}(\bar{\mathcal{B}}_k)\leq 2\sqrt{T}.$$

Step4 (Regret due to estimation error of transition probabilities): Define a constant $Z=\sqrt{160\Psi_{\max}\log(4D\Psi_{\text{tot}}T/\delta)}$. We can show that

$$\begin{aligned}
&\sum_{k=1}^{K}\sum_{l=0}^{m-1}\sum_{\psi\in\Psi_l}\sum_{i\in\mathcal{P}_1(\psi)}\nu_k(\psi,i)\langle(\tilde{\boldsymbol{P}}_k(\cdot|\psi,i)\\
&\qquad\qquad - \boldsymbol{P}(\cdot|\psi,i)),\hat{\boldsymbol{F}}_{l,k}\rangle\mathbb{I}(\mathcal{B}_k)\\
&\leq Z\sum_{l=0}^{m-1}\sum_{\psi\in\Psi_l}\sum_{i\in\mathcal{D}\setminus\text{dom}(\psi)}\sum_{k=1}^{K}\frac{\nu_k(\psi,i)}{\sqrt{\max(1,N_k(\psi,i))}}\\
&\overset{a}{\leq}(1+\sqrt{2})Z\sum_{l=0}^{m-1}\sum_{\psi\in\Psi_l}\sum_{i\in\mathcal{D}\setminus\text{dom}(\psi)}\sqrt{N_T(\psi,i)}\\
&\overset{b}{\leq}(1+\sqrt{2})\sqrt{160mD\Psi_{\max}\Psi_{\text{tot}}T\log(4D\Psi_{\text{tot}}T/\delta)}
\end{aligned}$$

where (a) follows from Lemma 15 and (b) from Jensen's inequality since $\sum_{\psi\in\Psi}\sum_{i\in\mathcal{D}\cup\emptyset}N_T(\psi,i)=mT$.

Step 5 (Regret due to randomness due to transition probabilities) We define $X_{l,t}=\langle(\boldsymbol{P}(\cdot|\psi_{l,t},\hat{g}_k(\psi_{l,t}))-\boldsymbol{e}_{\psi_{l+1,k(t)}}),\hat{\boldsymbol{F}}_{l,k(t)}\rangle\mathbb{I}(\mathcal{B}_{k(t)})$. Observe that $\|\hat{\boldsymbol{F}}_{l,k(t)}\|_\infty\leq 2$ since

$\hat{r}_t(a,\psi) \leq 1$ for each partial state- action pair and confidence levels are than 1. Therefore, for any $l = 0, 1, \ldots, m-1$,

$$|X_{l,t}| = \Bigg(\|\boldsymbol{p}(\cdot|\psi_{l,t}, \hat{g}_k(\psi_{l,t}))\|_1 + \|\boldsymbol{e}_{\psi_{l+1,k(t)}}\|_1 \Bigg) \|\hat{\boldsymbol{F}}_{l,k(t)}\|_\infty \mathbb{I}(\mathcal{B}_{k(t)}) \leq 4.$$

By Azuma-Hoeffding bound and following the same steps as the Step 5 of proof of Theorem 1, with probability $1 - \frac{\delta}{4T}$

$$\sum_{k=1}^{K} \sum_{t=t_k}^{t_{k+1}-1} \sum_{l=0}^{m-1} \langle (\boldsymbol{P}(\cdot|\psi_{l,t}, \hat{g}_k(\psi_{l,t})) - \boldsymbol{e}_{\psi_{l+1,t}}), \hat{\boldsymbol{F}}_{l,k} \rangle \mathbb{I}(\mathcal{B}_k)$$
$$= \sum_{l=0}^{m-1} \sum_{t=1}^{T} X_{l,t} \leq 4m\sqrt{2T \log(4T/\delta)}.$$

Step 6(Regret due to estimation error of rewards) Following the same steps as the Step 6 of proof of Theorem 11, equation 5.20 can be bounded as

$$(1+\sqrt{2})\sqrt{2\Psi_{\text{tot}} AT \log(20\Psi_{\text{tot}} AT/\delta)}.$$

Therefore,

$$\text{Reg}_T^{\text{Seq-OOS}} \leq \sqrt{\frac{T}{2}\log(4T/\delta)} + 2\sqrt{T} + 4m\sqrt{2T\log(4T/\delta)}$$
$$+ (1+\sqrt{2})\sqrt{160 m D \Psi_{\max} \Psi_{\text{tot}} T \log(4D\Psi_{\text{tot}} T/\delta)}$$
$$+ (1+\sqrt{2})\sqrt{10\Psi_{\text{tot}} AT \log(20\Psi_{\text{tot}} AT/\delta)}.$$

5.7 Appendices

5.7.1 Probability of Confidence Intervals Violation for Sim-OOS

Let $\widehat{\boldsymbol{P}}_t = (\hat{p}_t(\psi))_{\psi\in\Psi}$ and $\widehat{\boldsymbol{R}}_t = (\hat{r}_t(a,\psi))_{\psi\in\Psi}$ denote the observation/transition probability estimates and mean reward estimates at time t. The following lemma bounds the probability of $(\boldsymbol{P}, \boldsymbol{R})$ to be in the plausible set of observation/transition probability and mean rewards using the estimates at time t.

Lemma 12. *For any $t \geq 1$, if we set* $\text{conf}_1(n,t) = \min\left(1, \sqrt{\frac{\log(20\Psi_{tot}At^5/\delta)}{2\max(1,n)}}\right)$ *and* $\text{conf}_2(n,t) = \min\left(1, \sqrt{\frac{\log(10\Psi_{tot}\log(4t/\delta))}{\max(1,n)}}\right)$*, the probability that* $(\widehat{\boldsymbol{P}}_t, \widehat{\boldsymbol{R}}_t)$ *is not contained in the plausible set* $(\mathcal{P}_t, \mathcal{R}_t)$ *at time* t *is*

$$\Pr\left(\left(\widehat{\boldsymbol{P}}_t, \widehat{\boldsymbol{R}}_t\right) \notin (\mathcal{P}_t, \mathcal{R}_t)\right) \leq \frac{\delta}{5t^5}. \tag{5.25}$$

Proof.

$$\begin{aligned}
&\Pr\left(\left(\widehat{\boldsymbol{P}}_t, \widehat{\boldsymbol{R}}_t\right) \notin (\mathcal{P}_t, \mathcal{R}_t)\right) \\
&\leq \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \Pr\left(|\bar{r}(a,\psi) - \hat{r}_t(a,\psi)| \geq \text{conf}_{1,t}(a,\psi)\right) \\
&+ \sum_{\mathcal{I}\in\mathcal{P}_{\leq m}(\mathcal{D})} \Pr\left(\sum_{\psi\in\Psi^+(\mathcal{I})} |p(\psi) - \hat{p}_t(\psi)| \geq \text{conf}_{2,t}(\mathcal{I})\right) \\
&\overset{a}{\leq} \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \frac{2\delta}{20\Psi_{\text{tot}}At^5} + \frac{\delta}{10t^5} \leq \frac{\delta}{5t^5},
\end{aligned}$$

where $\text{conf}_{2,t}(\mathcal{I}) = \sqrt{\frac{10\Psi_{\text{tot}}\log(4t/\delta)}{N_t(\mathcal{I})}} \geq \sqrt{\frac{2\log(10\Psi_{\text{tot}}2^{\Psi_{\text{tot}}}t^5/\delta)}{\max(1,N_t(\mathcal{I}))}}$ and

$$\begin{aligned}
&\sum_{\mathcal{I}\in\mathcal{P}_{\leq m}(\mathcal{D})} \Pr\left(\sum_{\psi\in\Psi^+(\mathcal{I})} |p(\psi) - \hat{p}_t(\psi)| \geq \text{conf}_{2,t}(\mathcal{I})\right) \\
&\leq \sum_{\mathcal{I}\in\mathcal{P}_{\leq m}(\mathcal{D})} \Pr\left(\sum_{\psi\in\Psi^+(\mathcal{I})} |p(\psi) - \hat{p}_t(\psi)| \right. \\
&\qquad\qquad \left. \geq \sqrt{\frac{2\log(10\Psi_{\text{tot}}2^{\Psi_{\text{tot}}}t^5/\delta)}{\max(1,N_t(\mathcal{I}))}}\right) \\
&\overset{b}{\leq} \sum_{\mathcal{I}\in\mathcal{P}_{\leq m}(\mathcal{D})} \frac{\delta}{5\Psi_{\text{tot}}t^5} \leq \frac{\delta}{10t^5}
\end{aligned} \tag{5.26}$$

where (b) follows from Lemma 14. □

5.7.2 Probability of Confidence Intervals Violation for Seq-OOS

Let $\hat{\boldsymbol{P}}_t = \left(\hat{p}_t(\psi'|\psi,i)\right)_{\psi\in\Psi, i\in\mathcal{D}}$ and $\hat{\boldsymbol{R}}_t = \left(\hat{r}_t(a,\psi)\right)_{\psi\in\Psi}$ denote the observation/transition probability estimates and mean reward estimates at time t. The following lemma bounds the probability of $(\boldsymbol{P}, \boldsymbol{R})$ to be in the plausible set of observation/transition probability and mean rewards using the estimates at time t.

Lemma 13. *For any $t \geq 1$, if we set* $\mathrm{conf}_1(n,t) = \min\left(1, \sqrt{\frac{\log(20\Psi_{tot}At^5/\delta)}{2\max(1,n)}}\right)$ *and* $\mathrm{conf}_2(n,t) = \min\left(1, \sqrt{\frac{\log(10\Psi_{\max}\log(4t\Psi_{tot}D/\delta))}{\max(1,n)}}\right)$*, the probability that* $(\widehat{\boldsymbol{P}}_t, \widehat{\boldsymbol{R}}_t)$ *is not contained in the plausible set* $(\mathcal{P}_t, \mathcal{R}_t)$ *at time* t *is*

$$\Pr((\widehat{\boldsymbol{P}}_t, \widehat{\boldsymbol{R}}_t) \notin (\mathcal{P}_t, \mathcal{R}_t)) \leq \frac{\delta}{10t^5}. \tag{5.27}$$

Proof.

$$\begin{aligned}
&\Pr\left(\left(\widehat{\boldsymbol{P}}_t, \widehat{\boldsymbol{R}}_t\right) \notin (\mathcal{P}_t, \mathcal{R}_t)\right) \\
&\leq \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \Pr\left(|\bar{r}(a,\psi) - \hat{r}_t(a,\psi)| \geq \mathrm{conf}_{1,t}(a,\psi)\right) \\
&+ \sum_{\psi\in\Psi}\sum_{i\in\mathcal{D}\setminus\mathrm{dom}(\psi)} \Pr\Bigg(\sum_{\psi'\in\Psi^+(\psi,i)} |p(\psi'|\psi,i) - \hat{p}_t(\psi'|\psi,i)| \\
&\qquad\qquad \geq \mathrm{conf}_{2,t}(\psi,i)\Bigg) \\
&\overset{a}{\leq} \sum_{a\in\mathcal{A}}\sum_{\psi\in\Psi} \frac{2\delta}{20\Psi_{\mathrm{tot}}At^5} + \frac{\delta}{10t^5} \leq \frac{\delta}{5t^5},
\end{aligned}$$

where $\mathrm{conf}_{2,t}(\psi,i) = \sqrt{\frac{10\Psi_{\max}\log(4\Psi_{\mathrm{tot}}Dt/\delta)}{N_t(\mathcal{I})}} \geq \sqrt{\frac{2\log(10\Psi_{\mathrm{tot}}D2^{\Psi_{\max}}t^5/\delta)}{\max(1,N_t(\psi,i))}}$ and

$$\begin{aligned}
&\sum_{\psi\in\Psi}\sum_{i\in\mathcal{D}\setminus\mathrm{dom}(\psi)} \Pr\Bigg(\sum_{\psi'\in\Psi^+(\psi,i)} |p(\psi'|\psi,i) - \hat{p}_t(\psi'|\psi,i)| \\
&\qquad\qquad \geq \mathrm{conf}_{2,t}(\psi,i)\Bigg) \\
&\leq \sum_{\psi\in\Psi}\sum_{i\in\mathcal{D}\setminus\mathrm{dom}(\psi)} \Pr\Bigg(\sum_{\psi'\in\Psi^+(\psi,i)} |p(\psi'|\psi,i) - \hat{p}_t(\psi'|\psi,i)| \\
&\qquad\qquad \geq \sqrt{\frac{2\log(10\Psi_{\max}2^{\Psi_{\mathrm{tot}}}Dt^5/\delta)}{\max(1,N_t(\psi,i))}}\Bigg) \\
&\overset{b}{\leq} \sum_{\psi\in\Psi}\sum_{i\in\mathcal{D}\setminus\mathrm{dom}(\psi)} \frac{\delta}{10\Psi_{\mathrm{tot}}Dt^5} \leq \frac{\delta}{10t^5}
\end{aligned} \tag{5.28}$$

where (b) follows from Lemma 14. □

5.7.3 L_1 deviation of true and empirical distributions

Let $\mathcal{A}$ denote the finite set $\{1, 2, \ldots, a\}$. For two probability distributions $\boldsymbol{P}$ and $\boldsymbol{Q}$ on $\mathcal{A}$, let

$$\|\boldsymbol{P} - \boldsymbol{Q}\|_1 = \sum_{i=1}^{a} |P(i) - Q(i)|$$

denote L_1 distance between $\boldsymbol{P}$ and $\boldsymbol{Q}$. For a sequence $\boldsymbol{x}^n = x_1, \ldots, x_n \in \mathcal{A}^n$, let $\hat{\boldsymbol{P}}$ be the empirical probability distribution on $\mathcal{A}$ defined by

$$\hat{P}(j) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(x_i = j).$$

Lemma 14. *[WOS03] Let $\boldsymbol{P}$ be a probability distribution on the set $\mathcal{A} = \{1, 2, \ldots, a\}$. Let $\boldsymbol{X}^n = X_1, X_2, \ldots, X_n$ be i.i.d. random variables according to $\boldsymbol{P}$. Then, for all $\epsilon > 0$,*

$$\Pr(\|\boldsymbol{P} - \hat{\boldsymbol{P}}\| \geq \epsilon) \leq (2^a - 2)e^{-\epsilon^2 n/2}. \tag{5.29}$$

5.7.4 Summation bound

Lemma 15. *[JOA10] For any sequence of numbers $z_1, \ldots, z_n$ with $0 \leq z_k \leq Z_{k-1} = \max\left(1, \sum_{i=1}^{k-1} z_i\right)$,*

$$\sum_{k=1}^{n} \frac{z_k}{\sqrt{Z_{k-1}}} \leq (\sqrt{2} + 1)\sqrt{Z_n} \tag{5.30}$$

CHAPTER 6

Sequential Patient Allocation for Randomized Controlled Trials

6.1 Introduction

It is very rare that a clinician can understand/determine whether an treatment has a clinically relevant effect. Given uncertain knowledge about the disease and large variations in the patients' outcomes, it is often very challenging to say that an treatment has an effect on the basis of observational/case-control study. Randomized Controlled Trials (RCTs) are the gold standard for comparing the effect and value of treatment(s) because there exists a control group that is comparable to the treatment group in every possible way except the treatment.

An treatment goes through phase I-III of clinical research before it is accepted and implemented. Figure 6.1 summarizes the stages of RCTs. Phase I trials attempt to estimate tolerability and characterize pharmacokinetics and pharmacodynamics with participants of healthy volunteers. Phase II clinical trials attempt to determine the maximum tolerated dose that is the dose in which an adverse event occurs. In this paper, we focus on Phase III clinical trials that aim to determine the efficacy of the treatment with respect to control action. In the most of clinical trials, the designers are interested in answering the question of "*In population W is drug A at daily dose X more efficacious in improving Z by Q amount over a period of time T than drug B at daily dose Y*" [FFD98]. Table 6.1 shows the sample size of the trial, treatment evaluated on the trial, primary outcome used to evaluate the treatment action and result summary of the trial. As seen from the Table 6.1, the sample size of trials are ranging from 1000 to 4000, the type of outcomes are ranging from binary to time to event. Table 6.1 also shows that some of treatments are effective only on the

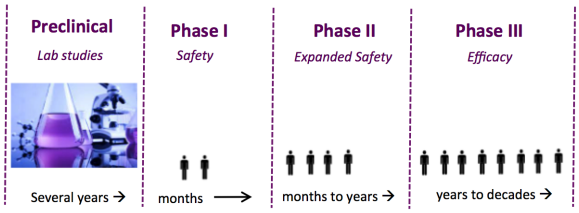


Figure 6.1: Stages of Clinical Trials

subgroups of the patients, not on the entire target population. The existing trials recruit patients from target population and uniformly randomly assign the patients to treatment groups. However, as we show later in the paper, this becomes very suboptimal as the heterogeneity of the patients increases. We propose a novel way of recruiting patients in adaptive way based on observations already made.

In this paper, we propose two different optimization criteria to design RCTs: average treatment effect estimation and efficacious treatment identification. The first criteria is to minimize the estimation error of subgroup-level treatment effect defined as the difference between the expected primary outcome of treatment and control action. We show that the optimal approach is to allocate the patients into treatment and control groups proportional to standard deviation of primary outcomes. Hence, the existing repeated fair coin tossing approach is only optimal if the standard deviations of the treatment and control group is the same. It is very unlikely that these groups have the same variations. Even if they did, it is impossible to verify that before the clinical trial has started. The second criteria is to minimize the convex combination of the type-I and type-II error of identifying the subgroups in which treatment action is efficacious with respect to control action. The optimal policy in this case depends both on the mean and standard deviation of primary outcomes of both treatment and control action and is more complicated. The parameters, hence the standard deviation, of the primary outcome distribution is unknown a priori. Hence, a learning algorithm is required to estimate the parameters and allocate the patients to treatment groups in adaptive way.

We model the RCT design as a finite stage MDP. The designer recruits N patients over K steps in order to optimize the learning criteria described above. We define the state

Study	Size	Treatment	Primary Outcome	Result
[Gro01]	3640	Supplement of vitamins	whether AMD event happened or not	significant reduction for the development of AMD event
[CBM04]	4162	Statin	Time to adverse event (death or stroke)	results consistent among the subgroups (no improvements)
[HWB99]	838	Red cell transfusion	30 days mortality	effective among the patients with Apache 2 score less than equal to 20 and age less than 55
[ROP12]	1272	Use of fingolimod	brain volume loss	subgroup analysis showed that it was effective for a subgroup

Table 6.1: RCT Examples in the Literature

vector as the parameters of the posterior distributions of the treatment and control actions, the transition function as the Bayes update of the parameters and single stage rewards as the improvements on the learning criteria by taking an allocation action. Unfortunately, the Dynamic Programming (DP) solution to solve the MDP is computationally intractable because of the large state and action space. Therefore, we propose Optimistic Knowledge Gradient (Opt-KG), computationally tractable, a greedy, and approximate solution to K-step MDP to achieve more plausible learning power.

Experiments are conducted in a synthetic dataset inspired by real RCTs with Bernoulli and Exponential outcome setting. We show that (i) Opt-KG achieves significant improvements in the mis-classification errors, (ii) achieves the same confidence levels with existing approaches in RCTs in lesser patient budget when the objective is to identify the efficacy of the treatment. When the objective is learning the subgroup level treatment effects, we show that the Opt-KG algorithm achieves significant improvement with respect to existing approaches in RCTs especially when the standard deviations of treatment outcomes differs.

The main contribution of the paper is three folds: (1) we formulate the RCT design problem as a finite stage MDP, and provide an optimal solution via DP, (2) we show an approximate, computationally tractable solution, Optimistic Knowledge Gradient (Opt-KG), (3) our framework can be used with various optimization criteria; including treatment effect estimation and identifying efficacious treatments.

6.2 Related Work

The most commonly used procedure to allocate patients into treatment and control group is "repeated fair coin-tossing". The main drawback is the possibility of imbalanced group sizes in RCTs with small patient budget [FFD98]. Hence, decision-makers follow a restricted procedure in which probability of assigning a patient to underrepresented group is increased, see blocked randomization [LMW88] and adaptive bias-coin randomization [SG02]. A less frequently used procedure is *covariate-adaptive randomization* in which patients are assigned to treatment groups to minimize *covariate imbalance* [MHS12].

The most similar existing approach to ours is *response adaptive randomization* in which the probability of being assigned to a treatment group increases if responses of prior patients in that particular group is favorable [HR06b]. This randomization approach takes the patient's benefit into account. The Multi-Armed Bandits (MABs) are mathematical decision framework for resource (patient) allocation to maximize the cumulative outcomes (patient benefit) [ACF02c, GGW11, AG12b]. [VBW15] shows that MABs achieve cumulative outcomes (patient benefit) but poor learning performance as they allocate most of the resources (patients) to favorable actions (treatments). Hence, [VBW15] proposes a variation on MAB algorithms that allocate patients to control action at times to overcome this issue of low learning performance. In our paper, we mainly focus on learning performance, not the patient benefit because the most of the RCTs aim to test the hypothesis of new drug being more efficacious than the control action [FFD98].

In this chapter, we provide a Bayesian approach to RCT design and model it as a finite stage MDP. The optimal patient allocation policy can be obtained via the Dynamic Programming (DP) that is computationally intractable for large size problems. The Gittins Index [GGW11, Whi80] provides an (optimal) closed form solution to the MDP, however, is again computationally intractable.

Knowledge Gradient(KG) policies provide an approximate, greedy approach to the MDP. Since the action space, containing all possible allocation choices, is very large, the KG policies are not either computationally tractable. Additionally, the KG mostly focus on multivariate

normal priors for the measurements [FPD08, FPD09]. In this paper, we provide a variation of the Opt-KG policies [CLZ13], to handle large action space.

6.3 Randomized Clinical Trial (RCT) Design

In this section, we describe the statistical model for RCTs. Our model for RCTs constitute three components: patient subgroups, treatments, treatment outcome. Define $\mathcal{W}$ as patient population and $\mathcal{X}$ as the set of discrete partition on $\mathcal{W}$ with $|\mathcal{X}| = X$. Let $\mathcal{Y} = \{0, 1\}$ be the action space and 0 indicating control action, 1 indicating treatment action and $\mathcal{Z}$ be the (bounded) outcome space, with $z_{\min}$ and $z_{\max}$ being the minimum and maximum outcome. Our goal in this paper is to recruit N patients over a period of T observing each outcome in T_{obs} period (having $K = T/T_{\text{obs}}$ steps) (i) to estimate the subgroup-level treatment effects, (ii) to identify efficacious treatments for the subgroups.

We model the treatment outcome distribution belonging to exponential family. In the next subsection, we give a brief description of the exponential family and Jeffrey's prior on their parameters.

6.3.1 Exponential Families and Jeffrey's Prior

Define $\mathcal{Z}$, Θ as a general sample space and general parameter space, respectively. Define $G : \mathcal{Z} \to \mathbb{R}$ and $h : \mathcal{Z} \to \mathbb{R}$. The single parameter exponential family with sufficient statistic G and parametrization θ, relative to h is dominated family of distributions by the following densities with respect to λ

$$p(z|\theta) = \Phi(\theta)h(z)\exp(\theta G(z))$$

where $\Phi(\theta)$ is uniquely determined by $\int_{-\infty}^{\infty} p(z|\theta)d\lambda(z) = 1$ hence

$$\Phi(\theta) = \left(\int_{-\infty}^{\infty} h(z)\exp(\theta G(z))d\lambda(z)\right)^{-1}.$$

It is usually written as

$$p(z|\theta) = h(z)\exp(\theta G(z) - F(\theta))$$

where $F(\theta) = -\log \Phi(\theta)$. Expectation and variance of the sufficient statistics can be expressed with respect to derivatives of the normalization function, i.e.,

$$F'(\theta) = \mathbb{E}_{Z|\theta}\left[G(Z)\right], \; F''(\theta) = \mathrm{Var}_{Z|\theta}\left[G(Z)\right].$$

Different choices of these functions generates a different probability density. Choosing $h(z) = 1$, $\theta = \ln \frac{p}{1-p}$, $G(z) = z$ generates a Bernoulli distribution with p. Some other examples of exponential family distribution are poison, and exponential distributions.

In this paper, we'll put a Bayesian "non-informative" prior, which is invariant under reparametrization of the parameter space. We use the Jeffrey's prior, that is proportional to square root of the Fisher information $I(\theta)$. In the case of the exponential family, the Fisher information is the second derivative of the normalization function, i.e., $I(\theta) = \sqrt{F''(\theta)}$. Under the Jeffrey's prior, the posterior on the parameter θ after n observation is given by

$$p(\theta|z_1, \ldots, z_n) \propto \sqrt{F''(\theta)} \exp\left(\theta \sum_{i=1}^{n} G(z_i) - nF(\theta)\right).$$

Given n outcomes $(z_1, z_2, \ldots, z_n)$, we can summarize the information needed for the posterior distribution is $\boldsymbol{s} = [s_0, s_1] = [\sum_{i=1}^{n} G(z_i), n]$. Then, the posterior is proportional to $\sqrt{F''(\theta)} \exp\left(\theta s_0 - s_1 F(\theta)\right)$.

6.3.2 MDP Model for RCT Design

In this subsection, we model the RCT design problem as a finite step in discounted MDP. A finite step MDP consists of number of steps, a state space, an action space, transition dynamics, a reward function. We need to define all the components of the MDP.

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space. As noted above, denote $\mathcal{X}$, $\mathcal{Y}$, $\mathcal{Z}$ as the set of subgroups of the patients, treatment set, and outcome set. Let $\theta_{x,y}$ be the true parameter for subgroup x and treatment y. The value indicating the success of treatment action y on patient subgroup x is referred to as treatment outcome and assume to be drawn according to exponential family distribution, i.e., $Z \sim p(\cdot|\theta_{x,y})$. Define $\mu(\theta_{x,y})$ denote true outcome of the treatment y on subgroup x, i.e., $\mu(\theta) = F'(\theta) = \mathbb{E}_{Z|\theta}\left[G(Z)\right]$ and treatment effect to be $E(\theta_0, \theta_1) = F'(\theta_1) - F'(\theta_0)$ for the parameters θ_0 and θ_1 so that treatment effect on

subgroup x is given by $E(\theta_{0,x}, \theta_{1,x})$. We are given a budget of N patients to be recruited in K steps. At time step k, the designer decides to recruit M_k patients $u_k(x, y)$ of whom are from subgroup x and assigned to treatment y where

$$\sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} u_k(x, y) = M_k.$$

Having made a decision $\boldsymbol{U}_k = \{u_k(x, y)\}$, the designer observes the outcomes $\boldsymbol{W}_k = \{W_k(x, y) = \sum_{j=1}^{u_k(x,y)} G(Z_j) : Z_j \sim \mathbb{P}(\cdot|\theta_{x,y})\}$. Denote $\bar{M}_k = \sum_{k=0}^{k-1} M_k$ as the number of patients recruited in the previous k decision time. Define the filtration $(\mathcal{F}_k)_{k=0}^{K}$ by letting $\mathcal{F}^k$ to be the sigma-algebra generated by the past decisions and observations $\{\boldsymbol{U}^0, \boldsymbol{W}^0, \boldsymbol{U}^1, \boldsymbol{W}^1 \ldots, \boldsymbol{U}^{k-1}, \boldsymbol{W}^{k-1}\}$. We'll define $\mathbb{E}_k[\cdot]$ to indicate $\mathbb{E}\left[\cdot|\mathcal{F}^k\right]$ and $\mathrm{Var}_k[\cdot]$ to indicate $\mathrm{Var}\left[\cdot|\mathcal{F}^k\right]$. Allocation decisions are restricted to be $\mathcal{F}^k$-measurable so that decisions depend only on measurements and decisions made in the past.

We'll use the Bayes rule to form a sequence of posterior predictive distributions for $F'(\theta_{x,y})$ from prior and successive measurements. Let $\mu_{x,y}^k = \mathbb{E}_k[F'(\theta_{x,y})]$ posterior mean outcome statistics and $\sigma_{x,y}^k = \mathrm{Var}_k(F'(\theta_{x,y}))$ be the posterior variance of the parameter of subgroup x and treatment y. Based on these definitions above, the posterior mean and variance of subgroup-level treatment effect are given by $E^k(x) = \mu_{x,0}^k - \mu_{x,1}^k$ and $\sigma_E^k(x) = \left(\sigma_{x,0}^k\right)^2 + \left(\sigma_{x,1}^k\right)^2$, respectively. Define the average treatment effect (ate) as the difference between the outcome of treatment and control action in a target population statistic, that is, $E_{pop}^k = \sum_{x\in\mathcal{X}} w_x E^k(x)$ where w_x is the proportion of target population with subgroup x.

The state space is the space of all possible distributions under consideration for $\{\theta_{x,y}\}$. Let $\boldsymbol{S}^k$ denote *state matrix* of shape $2X \times 2$ that contains the hyper-parameters of posterior distribution of the outcomes for both treatment and control actions for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$ in the kth decision time. Define $\mathcal{S}^k$ to be the all possible states at nth stage, that is,

$$\begin{aligned}\mathcal{S}^k = \{\boldsymbol{S}^k = [\boldsymbol{s}_{x,y}^k]| : \bar{M}_k y_{min} \le s_{x,y,0}^k - s_{x,y,0}^0 \le \bar{M}_k y_{max},\\ 0 \le s_{x,y,1}^k - s_{x,y,1}^0 \le \bar{M}_k,\ \forall(x, y)\}\end{aligned}$$

Our action space in decision time k is the set of all possible pairs of $(M_k, \boldsymbol{U}_k)$ with $M_k \le N - \bar{M}_{k-1}$ and $\sum_{x,y} u_k(x, y) = M_k$. Taking an action $a_k = (M_k, \boldsymbol{U}_k)$ means recruiting

M_k patients $u_k(x, y)$ of whom are from subgroup x and assigned to treatment y. Fix the decision stage as k. When the designer selects one of the actions, we use the Bayes rule to update the distribution of $\theta_{x,y}$ conditioned on $\mathcal{F}^k$ based on outcome observations of $\boldsymbol{W}_k$, obtaining a posterior distribution conditioned on $\mathcal{F}^{k+1}$. Thus, our posterior distribution for $\theta_{x,y}$ is proportional to $\sqrt{F''(\theta)}\exp\left(\theta s^k_{x,y,0} - F(\theta)s^k_{x,y,1}\right)$. The parameters of the posterior distribution can be written as a function of $\boldsymbol{s}^k$ and $\boldsymbol{W}_k$. Define $\boldsymbol{S}^{k+1} = P(\boldsymbol{s}^k, a_k, \boldsymbol{W}_k)$ as the transition function given observed treatment outcome $\boldsymbol{Z}_k$ and $\mathbb{P}(\boldsymbol{S}^{k+1}|\boldsymbol{S}^k, a_k)$ as the posterior state transition probabilities conditioned on the filtration in decision time k. Having taken an allocation action $a_k = (M_k, \boldsymbol{U}_k)$, the state transition probabilities can be given by: $\boldsymbol{S}^{k+1} = \boldsymbol{s}^k + \left[\boldsymbol{W}^k, \boldsymbol{U}^k\right]$ with posterior predictive probability $p(\boldsymbol{W}|\boldsymbol{s}^k)$ where

$$\mathbb{P}(\boldsymbol{W}|\boldsymbol{s}) = \prod_{(x,y)\in\mathcal{X}\times\mathcal{Y}} \mathbb{P}(V(x,y)|\boldsymbol{s}_{x,y})$$

$$\mathbb{P}(W|\boldsymbol{s}_{x,y}) = \int_{\theta\in\Theta} \mathbb{P}\left(\sum_{j=1}^{u_k(x,y)} G(Z_j) = W \middle| \theta\right) \mathbb{P}(\theta|\boldsymbol{s}_{x,y})\, d\theta$$

Our objective is to minimize the weighted expected mean-squared error. The expected mean-squared error of the average treatment effect of subgroup x is given by

$$\begin{aligned}\text{mse}^k(x) &= \mathbb{E}\left[\left(\bar{E}^k(x) - \mathbb{E}\left[\bar{E}^k(x)\right]\right)^2\right] \\ &= \sum_{y=0}^{1} \mathbb{E}\left[\left(\mu^k_{x,y} - \mathbb{E}\left[\mu^k_{x,y}\right]\right)^2\right] = \sum_{y=0}^{1}\left(\sigma^k_{x,y}\right)^2.\end{aligned}$$

For a state $\boldsymbol{s} = (\tilde{\boldsymbol{u}}, \tilde{\boldsymbol{w}})$, the action space is the set of pairs of $(m, \boldsymbol{u})$ with m is less than equal to patient budget left, and elements in $\boldsymbol{u}$ summing up to m, that is given by

$$\mathcal{A}(\boldsymbol{s}) = \{(m, \boldsymbol{u}) : 1 \leq m < N - \mathbf{1}^T\tilde{\boldsymbol{u}}\mathbf{1}, \mathbf{1}^T\boldsymbol{u}\mathbf{1} = m\}$$

We define the reward function as the decrease on the mean-squared error by taking an action a in state $\boldsymbol{s}$. The reward function $R : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ is then given by

$$R(\boldsymbol{s}, a) = \sum_{x\in\mathcal{X}} w_x \mathbb{E}\left[\text{mse}^k(x) - \text{mse}^{k+1}(x)\middle|\boldsymbol{s}, a\right]$$

where expectation is taken with respect to outcome distribution of treatment and control actions under state vector $\boldsymbol{s}$. An allocation policy is a mapping from state space to action

space, prescribing the number of patients to recruit from each subgroup and how to assign them to treatment groups, that is $\pi : \mathcal{S} \to \mathcal{A}$. The value function of a policy π starting from state of $\boldsymbol{S}^0$ is given by

$$\begin{aligned} V^{\pi}(\boldsymbol{S}^0) &= B(\boldsymbol{S}^0) - \sum_{(x,y)} \mathbb{E}^{\pi}\left[w_x(\sigma^K_{x,y})^2\right] \\ &= \sum_{k=0}^{K} \mathbb{E}^{\pi}\left[R(\boldsymbol{S}^k, A_k)\right] \end{aligned} \tag{6.1}$$

where $B(\boldsymbol{S}^0) = \sum_{x\in\mathcal{X}} w_x \,\mathrm{mse}^0(x)$ and the expectation is taken with respect to allocation policy π. Our learning problem is K-stage of MDP with tuple:

$$\{T, \{\mathcal{S}^k\}, \{\mathcal{A}(\boldsymbol{s})\}, P(\boldsymbol{S}^k, a_k, \boldsymbol{W}_k), R(\boldsymbol{S}^k, a_k)\}$$

and our goal is to solve the following optimization problem:

$$\begin{aligned} &\text{maximize}_{\pi}\ \mathbb{E}^{\pi}\left[R(\boldsymbol{S}^n, \pi(\boldsymbol{S}^n))\right] \\ &\text{subject to } \forall n:\ \pi(\boldsymbol{S}_n) \in \mathcal{A}(\boldsymbol{S}_n) \end{aligned}$$

In the next subsection, we propose a Dynamic Programming (DP) approach for solving the K-stage MDP defined above.

6.3.3 Dynamic Programming (DP) solution

In the dynamic programming approach, the value function is defined as the optimal value function given a particular state $\boldsymbol{S}^k$ at a particular stage k, and may be determined recursively through Bellman's equation. If the value function can be computed efficiently, the optimal policy can also be computed from it. The optimal value function at the terminal stage $K-1$ (the stage in which last patient is recruited) is given by:

$$V^{K-1}(\boldsymbol{s}) = \max_{a\in\mathcal{A}(\boldsymbol{s})} R(\boldsymbol{s}, a)$$

The dynamic programing principle tells us that the loss function at other indices $0 \le n < K-1$ is given recursively by

$$\begin{aligned} Q^k(\boldsymbol{s}, a) &= \mathbb{E}_n\left[V^k(P(\boldsymbol{s}, a, \boldsymbol{W}))\right], \\ V^k(\boldsymbol{s}) &= \max_{a\in\mathcal{A}(\boldsymbol{s})} Q^k(\boldsymbol{s}, a) \end{aligned}$$

where the expectation is taken with respect to randomness of the cumulative treatment outcomes $\boldsymbol{W}$. The dynamic programming principle tells us that any policy that satisfies the following is optimal: $A_k^* = \arg\max_{a \in \mathcal{A}(\boldsymbol{S}^k)} = Q^k(\boldsymbol{S}^k, a)$.

It is computationally intractable to solve the DP because (i) the state space contains all possible distributions under consideration, hence very large, (ii) the value and Q functions need to be re-computed each step the state vector is updated.

In the rest of the paper, we show an (approximate) solution to the case in which the number of the patients to recruit in each step is fixed and determined by the designer initially. However, we show in the experiments that choosing m can have a large impact on the performance. We propose a greedy solution to solve the MDP approximately in computationally tractable way. Our proposed solution, Opt-KG, computes the maximum one -stage reward that can be obtained by taking each action $a \in \mathcal{A}$, and then selects the action with the maximum one-stage reward

6.4 A greedy solution for fixed recruitment: Optimistic Knowledge Gradient (Opt-KG)

In the rest of the paper, we focus on the setting where the number of patients to recruit is fixed and equal to $M = T/K$. Then, the allocation decision is all possible recruitments and treatment assignments of M patients. The action set is

$$\mathcal{A} = \left\{ \boldsymbol{u} : \sum_x \sum_y \boldsymbol{u}(x, y) = M \right\}.$$

and size of this action set is $|\mathcal{A}| = \binom{M+2X-1}{2X-1}$. The Rand-KG algorithm computes an expected improvement in the terminal value function by taking an action $a \in \mathcal{A}$. We can write the value function at terminal stage can be decomposed into improvements in the value function, that is,

$$\begin{aligned} V^K(\boldsymbol{S}^K) = &\left[V^K(\boldsymbol{S}^K) - V^K(\boldsymbol{S}^{K-1})\right] \\ &+ \ldots + \left[V^K(\boldsymbol{S}^{k+1}) - V^K(\boldsymbol{S}^k)\right] + V^K(\boldsymbol{S}^k) \end{aligned}$$

The Knowledge Gradient (KG) is the policy that selects the action that will make the maximum improvements in the value function at each instance, that is,

$$\begin{aligned} A_k^{KG}(\boldsymbol{s}) &= \arg\max_{a\in\mathcal{A}(\boldsymbol{s})} \mathbb{E}_k\left[V^K(P(\boldsymbol{s},a,\boldsymbol{W})) - V^K(\boldsymbol{s})\right] \\ &= \arg\max_{a\in\mathcal{A}(\boldsymbol{s})} \int_{\boldsymbol{w}} V^K(P(\boldsymbol{s},a,\boldsymbol{w}))\mathbb{P}(\boldsymbol{w}|\boldsymbol{S}^k)\,d\boldsymbol{w} \end{aligned}$$

As seen from above, computing the KG policy requires computing the posterior predictive distribution and posterior expectation integral for each action in $\mathcal{A}$. This is also computationally expensive procedure since the action space is in the order of $\mathcal{O}\left(M^{2X-1}\right)$. That is why we focus on the Optimistic Knowledge Gradient (Opt-KG) Algorithm that computes optimistic improvements in the value functions. Algorithm 8 shows a tractable way of computing the optimistic functions. At each iteration m, the procedure computes the maximum improvement in the value function by an additional sample from the pair (x, y) and increments that index by 1. The complexity of computing $A^{Opt-KG}(\boldsymbol{s})$ is $\mathcal{O}(MX)$

At each decision step k, the algorithm computes the best action using the procedure in Algorithm 8 and recruits the patients accordingly. At the end of decision step, the state vector is updated based on treatment outcome observations. When the patient budget is exhausted, our algorithm outputs the subgroup-level treatment effect and average treatment effect for whole population. At the terminal stage K, ATE for the subgroup x is given by

$$\bar{E}^K(x) = \mathbb{E}\left[F'(\theta)|\boldsymbol{S}_{x,0}^K\right] - \mathbb{E}\left[F'(\theta)|\boldsymbol{S}_{x,1}^K\right],$$

and ATE for whole population is given by

$$E_{pop}^K = \sum_{x\in\mathcal{X}} w_x \bar{E}^K(x)$$

where w_x is the proportion of the population with subgroup x. The pseudo-code for the Opt-KG is given in Algorithm 9.

6.5 Extension to Treatment Efficacy Identification in RCTs

In this section, we modify our problem to identify the subgroups in which the treat action is effective with respect to control action. Efficacy of treat action with respect to control

Algorithm 8 Optimistic Action Computation

Input : Current state vector: $\boldsymbol{s}$

Set optimal action $\boldsymbol{u}^* = \mathbf{0}$

for $m = 1, \ldots, M$ **do**

  for $(x, y) \in \mathcal{X} \times \mathcal{Y}$ **do**

    Set $\boldsymbol{s}_1 = P(\boldsymbol{s}, \boldsymbol{u}^*, \boldsymbol{u}^* z_{max})$

    Set $\boldsymbol{s}_2 = P(\boldsymbol{s}, \boldsymbol{u}^*, \boldsymbol{u}^* z_{min})$

    Set $\tilde{\boldsymbol{u}} = \boldsymbol{u}^* + \mathbf{1}_{(x,y)}$

    Compute $v_1 = V^K(P(\boldsymbol{s}, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}} y_{max})) - V^K(\boldsymbol{s}_1)$

    Compute $v_2 = V^K(P(\boldsymbol{s}, \tilde{\boldsymbol{u}}, \tilde{\boldsymbol{u}} y_{min})) - V^K(\boldsymbol{s}_2)$

    Compute $q(x, y) = \max(v_1, v_2)$.

  end for

  Compute $(x^*, y^*) = \arg\max_{x,y} q(x, y)$.

  Update $\boldsymbol{u}^* = \boldsymbol{u}^* + \mathbf{1}_{(x^*, y^*)}$.

end for

Return $A^{Opt-KG}(\boldsymbol{s}) = \boldsymbol{u}^*$.

Algorithm 9 The Opt-KG Algorithm for ATE Estimation

Input : $\{w_x\}$, K, $\boldsymbol{S}^0$

for $k = 1, \ldots, K$ **do**

  Compute $\boldsymbol{U}_k^* = A^{Opt-KG}(\boldsymbol{S}^k)$ using Algorithm 1.

  Recruit the patients based on $\boldsymbol{U}_k^*$, observe cumulative treatment outcome $\boldsymbol{W}_k^*$.

  Update $\boldsymbol{S}^{k+1} = \boldsymbol{S}^k + (\boldsymbol{U}_k^*, \boldsymbol{W}_k^*)$.

end for

Compute $\mu_{x,y}^K = \mathbb{E}\left[F'(\theta) | \boldsymbol{S}_{x,y}^K\right]$

Compute the treatment effect $\bar{E}^K(x) = \mu_{x,1}^K - \mu_{x,0}^K$.

Compute ATE as $\bar{E}_{pop}^K = \sum_{x \in \mathcal{X}} w_x \bar{E}^K(x)$.

Output : $\{\bar{E}^K(x)\}$, E_{pop}^K.

action is defined as the indicator in which treat outcome is τ-percent better than control outcome, that is,

$$\nu(x) = \begin{cases} 1 & \text{if } \frac{[\mu(\theta_{x,0}) - \mu(\theta_{x,1})]}{\mu(\theta_{x,0})} \geq \tau \\ 0 & \text{if otherwise} \end{cases}$$

We define the positive set $\mathcal{H}^* = \{x \in \mathcal{X} : \nu(x) = 1\}$ as the set of subgroups in which treat action is effective and $\bar{\mathcal{H}}^* = \mathcal{X} \setminus \mathcal{H}^*$ as set of subgroups in which treat action is ineffective. In this setting, the output is an estimated set of subgroups in which treat action is effective with respect to control action, that is, $\mathcal{H}^K$. Define $\bar{\mathcal{H}}^K$ as the set of subgroups in which treat action is estimated to be ineffective. We call any subgroup $x \in \mathcal{H}^*$ as the effective subgroup and any subgroup $x \in \mathcal{H}^K$ as estimated effective subgroup. There are typically two types of errors: type-I error in which an actual effective subgroup is found to be ineffective and type-II error in which an actual ineffective subgroup is found to effective when the patient budget is exhausted. The formal definitions for type-I and type-II errors are given by:

$$\mathrm{e}_1^K = \sum_{x \in \mathcal{X}} 1\left(x \in \mathcal{H}^*\right) 1\left(x \notin \mathcal{H}^K\right)$$
$$\mathrm{e}_2^K = \sum_{x \in \mathcal{X}} 1\left(x \notin \mathcal{H}^*\right) 1\left(x \in \mathcal{H}^K\right)$$

The total error is the convex combination of type-I and type-II error and is given by: $\mathrm{e}^K = \lambda \mathrm{e}_1^K + (1-\lambda)\mathrm{e}_2^K$ where $\lambda \in (0,1)$ is the tradeoff parameter between type-I and type-II error. This is a parameter that is selected by the designer based on the actual costs of Type-I and Type-II errors. Selecting a large λ will potentially decrease the Type-I error but increase the Type-II error. Selecting a small λ will do the opposite changes in Type-I and Type-II errors.

Based on the outcome distributions at a terminal stage, our objective is to identify the subgroups in which treatment action is efficacious with respect to control action. We can

formally define our objective as:

$$\mathcal{H}_K = \arg\min_H \sum_{x\in\mathcal{X}} \mathbb{E}\big[\lambda 1(x \notin H)1(x \in \mathcal{H}^*) + (1-\lambda)1(x \in H)1(x \notin \mathcal{H}^*)\big|\mathcal{F}^K\big] \tag{6.2}$$

The optimization problem stated above can be solved by a naive way of iterating all possible H. However, this solution will be exponential in the number of patient subgroups. We show a solution to (6.2) that is linear in the number of patient subgroups. Before we present the solution to (6.2), we define the posterior probability on the treatment efficacy at decision stage k as P_x^k, This is the probability that efficacy is 1 conditional on $\theta_{x,0}$ and $\theta_{x,1}$ drawn according to posterior distributions. Formally,

$$P_x^k = \mathbb{P}\left(\frac{\mu(\theta_0)-\mu(\theta_1)}{\mu(\theta_0)} \geq \tau \middle| \begin{matrix}\theta_0\sim P_{\theta|S_{x,0,0},S_{x,0,1}} \\ \theta_1\sim P_{\theta|S_{x,1,0},S_{x,1,1}}\end{matrix}\right).$$

This is the probability that treatment action is τ-percent effective with respect to control action given the posterior probabilities of $\boldsymbol{S}_{x,y}^k$.

A simple solution to (6.2) is $\mathcal{H}_K = \{x \in \mathcal{X} : P_x^K \geq 1-\lambda\}$. Define g as

$$g(x;\lambda) = \lambda(1-x)1(x \geq 1-\lambda) + (1-\lambda)x1(x < 1-\lambda).$$

Then, the total error can be written as $\mathrm{e}^K = \sum_{x\in\mathcal{X}} g(P_x^K)$. where $B(\boldsymbol{S}^0) = \sum_{x\in\mathcal{X}} g(P_x^0)$. To complete our MDP formulation, define the reward function $R(\boldsymbol{s}, a)$ as the decrease in the error by taking an allocation action a in state $\boldsymbol{s}$, that is,

$$R(\boldsymbol{s}, a) = \mathbb{E}\left[g(P_x^k;\lambda) - g(P_x^{k+1};\lambda)\big|\boldsymbol{s}, a\right].$$

The value function of a policy π starting from $\boldsymbol{S}^0$ is then given by:

$$\begin{aligned} V^\pi(\boldsymbol{S}^0) &= B(\boldsymbol{S}^0) - \sum_{x\in\mathcal{X}} \mathbb{E}^\pi\left[g(P_x^K)\right] \\ &= \mathbb{E}^\pi\left[R(\boldsymbol{S}^k, a_k)\right]. \end{aligned}$$

where $B(\boldsymbol{S}^0) = \sum_{x\in\mathcal{X}} g(P_x^0)$ and the expectation is taken with respect to allocation policy π. Our learning problem can then be modified to use new reward functions.

We can then modify the DP, and the OPT-KG algorithm for this new reward definition. At the end of the allocation, the subgroups in which the treatment action is efficacious with respect to control action can be determined by $\mathcal{H}_K = \{x \in \mathcal{X} : P_x^K \geq 1 - \lambda\}$.

6.6 Experiments

In our experiments, we use two different types of primary outcomes: time to adverse event (Exponential distribution) and an indicator for an adverse event (Bernoulli distribution). An example for the Exponential outcome is clinical study in [CBM04]. In this study, 4162 patients who had been hospitalized for an acute coronary syndrome within the preceding 10 days are recruited. The treatments of 40 mg of pravastatin daily (control action) and 80 mg of atorvastatin daily (treatment action) are compared with respect to the primary outcome of death from any cause. An example for the Bernoulli outcome case is clinical study in [HWB99]. In this study, 838 critically ill patients are recruited and randomly assigned to restrictive strategy and liberal strategy of transfusion. The primary outcome of 30-day mortality is used to evaluate these two strategies.

We use the exponential outcomes with the treatment and control actions without subgroups in the first simulation setting and Bernoulli outcomes with treatment and control action with 4 subgroups. In all the experiments below, we assume 100 patients are recruited in each step. We generate 1000 different experiments and report the average performance metrics. We report the Root Mean Squared Error (RMSE), that is rmse $= \sqrt{\frac{1}{1000}\sum_{j=1}^{1000}(\bar{E}_j^K - E_j)^2}$ in the first simulation setting and type-I, type-II and total error rate, that is,

$$\begin{aligned} \text{err}_1 &= \frac{\sum_{j=1}^{1000}\sum_{x\in\mathcal{X}} 1\left(x \in \mathcal{H}_j^*\right) 1\left(x \notin \mathcal{H}_j^K\right)}{\sum_{j=1}^{1000}\sum_{x\in\mathcal{X}} 1\left(x \in \mathcal{H}_j^*\right)} \\ \text{err}_2 &= \frac{\sum_{j=1}^{1000}\sum_{x\in\mathcal{X}} 1\left(x \notin \mathcal{H}_j^*\right) 1\left(x \in \mathcal{H}_j^K\right)}{\sum_{j=1}^{1000}\sum_{x\in\mathcal{X}} 1\left(x \notin \mathcal{H}_j^*\right)} \end{aligned} \tag{6.3}$$

and total error rate is 1- accuracy.

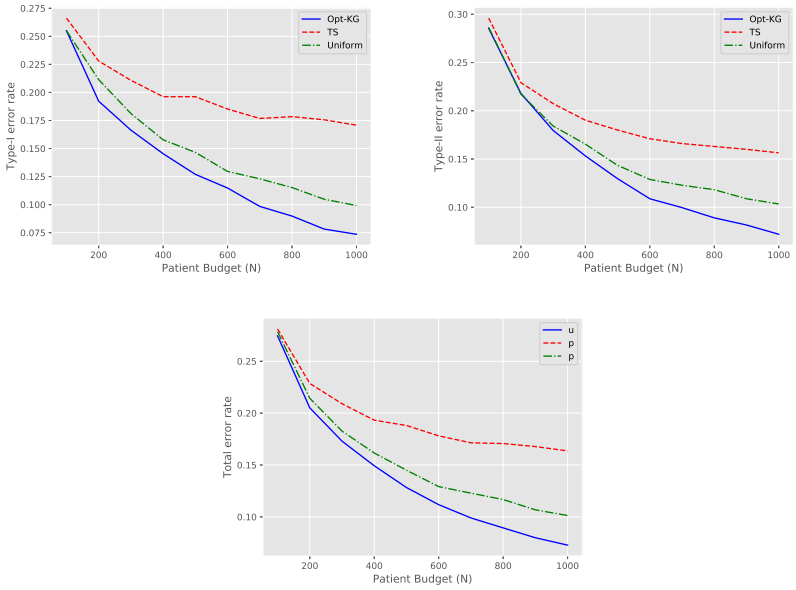


Figure 6.2: Error Comparisons with Benchmarks

We compare our algorithm with Uniform Allocation (UA), aka repeated fair coin tossing, that uniformly randomly recruits the patients from subgroups and uniformly assigns the patients to treatment groups, and Thompson Sampling (TS) [AG12b] that draws the parameters of the outcomes for each action and then selects the action with the maximum sample.

6.6.1 Results on Treatment Efficacy Identification

In this subsection, we perform experiments on Treatment efficacy identification problem. In this first experiment, we assume 2 subgroups with $\theta_{x,0} = 0.5$ for all $x \in \{0, 1\}$, and $\theta_{1,1} = 0.7$ and vary $\theta_{0,1}$ from 0.51 to 0.6. Our goal is to identify the best treatment for each subgroup. In this setting, classifying the subgroup 0 is more challenging than classifying subgroup 1. As seen from the Figure 6.3, our algorithm outperforms the uniform allocation in the cases when the gap between treatment and control among the subgroups is different from each other. This is improvement is achieved by sampling more from the subgroups in which the gap is smaller.

In the rest of the experiments on this subsection, we assume 4 different subgroups with

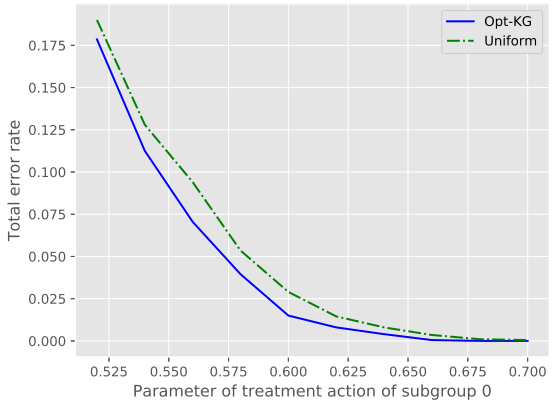


Figure 6.3: Total error rates for different parameter

$\theta_{x,0} = 0.5$ for all $x \in \mathcal{X}$, and $\{\theta_{x,1}\} = (0.3, 0.45, 0.55, 0.7)$. Our goal is to identify the subgroups in with treatment action is more efficient than the control action. We use this setting because there are 2 subgroups in which treatment action is inefficient and 2 subgroups in which treatment action is efficient and some subgroups are easier to classify than others. For example, subgroup 1 is easier to classify than subgroup 3. In the second experiment, we show the Type-I, Type-II and total error of the Opt-KG, TS and UA algorithms for different patient budget. As seen from Figure 6.2, Opt-KG significantly outperforms the UA and TS algorithm for all budgets. We note that UA outperforms the TS algorithm when the patient budget is large. This is because TS aims to maximize the patient benefit, not the learning performance. Hence, it allocates all the patients to better treatment, and eventually estimates the action with lower performance poorly.

In the third experiment, Figure 6.4 shows the trade-off between the Type-I and Type-II errors. As seen from the Figure 6.4, Type-I error of the Opt-KG algorithm decreases with λ, and Type-II error of the Opt-KG increases with λ. Therefore, investigators can achieve trade-offs between the errors by setting λ.

The next experiment shows the performance of the OPT-KG algorithm in the case when m patients are recruited sequentially till the patient budget of 500. As shown from the Table 6.2, the error of the Opt-KG increases with m. When $m = 250$ patients are recruited in one time, the Opt-KG only adapts the allocation policy in the second 250 patients and hence

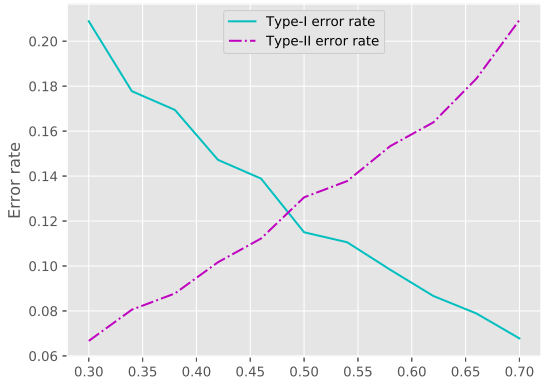


Figure 6.4: Tradeoffs between Type-I and Type-II errors

has larger error.

m	10	25	50	100	250
Opt-KG	0.1209	0.1245	0.1281	0.1292	0.1411
UA	0.1484	0.1484	0.1484	0.1484	0.1484

Table 6.2: Error Metric for different patient m

In the experiment, we illustrate the importance of the informative prior on patient recruitment in RCTs. We generate informative priors by sampling 10 patients from each subgroup and treatment group (without using from the patient budget). As seen from the Table 6.3, our improvement with respect to Uniform Allocation is larger in the case of informative prior. The informative prior is useful for the Opt-KG in two ways: (i) the informative prior helps the Opt-KG to make more informed decisions on patient recruitment in the first steps, (ii) the informative prior helps the Opt-KG to focus more on the subgroups with smaller gaps.

In the experiment, we show a slightly different result. As in the previous steps, we assume that 100 patients are recruited in each step. And the trial runs until a confidence level is reached, that is $\frac{1}{X}\sum_{x\in\mathcal{X}} g(P_x^n) \leq \beta$ for some confidence level β. As seen from Table IV, the Opt-KG recruits less number of patients than the UA to achieve the same confidence level.

Budget	200	500	1000
Non-informative	0.0429	0.1151	0.2813
Informative	0.0748	0.1349	0.3263

Table 6.3: Improvement score for different budgets

we show the number of patients recruited by Opt-KG and uniform allocation until $\frac{1}{X}\sum_{x\in\mathcal{X}} g(P_x^n) \leq \beta$. In this experiment, we again use the Bernoulli outcome model and report the average number of patients who are recruited by Opt-KG and UA for different values of β. As seen from Table 6.4, the UA achieves the same confidence level as Opt-KG by recruiting 2 times more patients than Opt-KG, or running the trial 2 time more than UA. This means that designer might save tremendous resources by implementing Opt-KG

Algorithm	$\beta = 0.05$	$\beta = 0.1$
Opt-KG	11.3	6.1
UA	21.7	9.1

Table 6.4: Comparison of trial length for a confidence level

6.6.2 Results on ATE estimation

In this subsection, we perform experiments on ATE estimation problem. The first experiment is designed to show the effectiveness of our algorithm for different parameter of treatment outcome where we set $\theta_0 = 1$, and vary θ_1. Figure 6.5 shows the RMSE improvement of Opt-KG with respect to UA for different θ_1, and hence different standard deviation of treatment outcome distribution. Figure 6.5 shows that the Opt-KG achieves lower RMSE by allocating more patients to the treatment group with higher standard deviation. The Opt-KG allocates 87% of the patients to treatment group for $\theta_1 = 0.01$ (and standard deviation of 100) and achieves more than 25% improvement in RMSE with respect to UA. In the case when the standard deviations of both actions are the same, the RMSE performance of Opt-KG is the

same as UA.

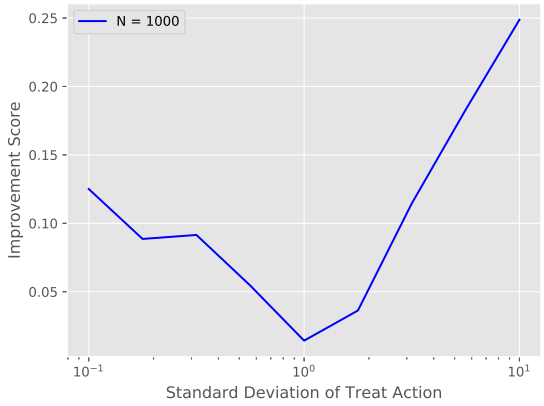


Figure 6.5: RMSE improvement of Opt-KG with respect to UA

Figure 6.6 shows the RMSE performance of Opt-KG, UA and TS for different patient budget for parameter of control action $\theta_0 = 1$, parameter of the treatment action $\theta_1 = 0.1$. As seen from the Figure, the Opt-KG outperforms both algorithms in all budgets. Note again that the performance of the TS algorithm is even worse than the UA algorithm for larger budgets.

In the next experiment, we focus on extreme case in which the designer has only two steps to recruit $N = 1000$ patients for parameter of control action $\theta_0 = 1$ and θ_1 varying from $1, 5, 10, 25$. The designer recruits M patients in the first step and $N - M$ patients in the second step. Figure 6.7 shows the RMSE for different values of M. As seen from the Figure, the minimum RMSE for ATE is achieved in $M^* = 85, 150, 235$ for the parameters $5, 10, 25$ respectively. If the designer recruited less patients than M^*, the designer would not be able to estimate the parameters effectively and hence would not be able to recruit the patients in an optimal way. If the designer recruited more patients than M^*, the designer would not have sufficient number of patient budget left to achieve the optimal allocation.

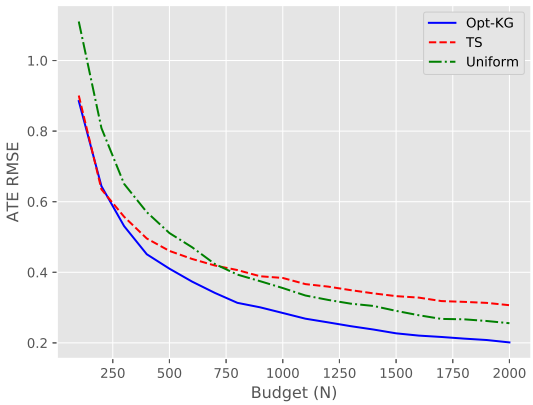


Figure 6.6: RMSE performance of Opt-KG with respect to UA and TS

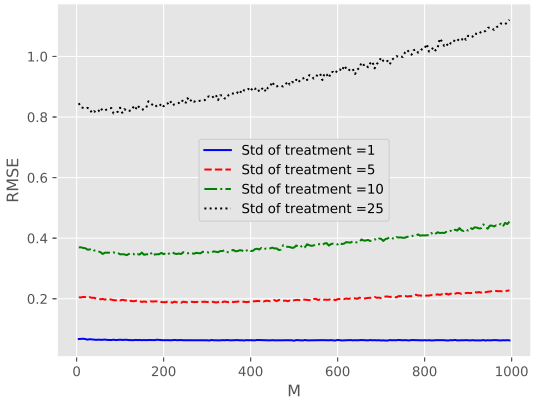


Figure 6.7: RMSE performance of recruiting M patients in the first step

CHAPTER 7

Concluding Remarks

In this thesis, we studied two different ways to learn personalized treatment policies: learning by conducting a clinical trial, learning from observational studies. In the case of clinical trials, we proposed algorithms both to maximize the patient benefit (Chapter 2, Chapter 5) and learning power (Chapter 6). In the case of observational studies, we proposed efficient deep learning algorithms in which selection bias is handled with feature selection (Chapter 3) and representation learning (Chapter 4). Although our focus in this thesis was medical informatics, there are wide range of applications in which our algorithms are applicable – from education to recommender systems.

[ACF95] Peter Auer, Nicolo Cesa-Bianchi, Yoav Freund, and Robert E Schapire. "Gambling in a rigged casino: The adversarial multi-armed bandit problem." In *Annual Symposium on Foundations of Computer Science*, pp. 322–331, 1995.

[ACF02a] P. Auer, N. Cesa-Bianchi, and P. Fischer. "Finite-time analysis of the multiarmed bandit problem." *Machine Learning*, **47**:235–256, 2002.

[ACF02b] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time Analysis of the Multi-armed Bandit Problem." *Machine Learning*, **47**:235–256, 2002.

[ACF02c] Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. "Finite-time analysis of the multiarmed bandit problem." *Machine learning*, **47**:235–256, 2002.

[AG12a] Shipra Agrawal and Navin Goyal. "Analysis of Thompson Sampling for the Multi Armed Bandit Problem." In *Proc. COLT*, 2012.

[AG12b] Shipra Agrawal and Navin Goyal. "Analysis of thompson sampling for the multi-armed bandit problem." In *Conference on Learning Theory*, pp. 1–39, 2012.

[AG13] Shipra Agrawal and Navin Goyal. "Thompson sampling for contextual bandits with linear payoffs." In *Proc. ICML*, 2013.

[AI15] Susan Athey and Guido W Imbens. "Recursive Partitioning for Heterogeneous Causal Effects." *arXiv preprint arXiv:1504.01132*, 2015.

[AJS18] Onur Atan, James Jordon, and Mihaela van der Schaar. "Deep-Treat: Learning Optimal Personalized Treatments From Observational Data Using Neural Networks." In *AAAI Conference on Artificial Intelligence*, pp. 2071–2078, 2018.

[AMS09] Jean-Yves Audibert, Rémi Munos, and Csaba Szepesvári. "Exploration–exploitation tradeoff using variance estimates in multi-armed bandits." *Theoretical Computer Science*, **410**(19):1876–1902, 2009.

[APS11] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. "Improved algorithms for linear stochastic bandits." In *Advances in Neural Information Processing Systems*, pp. 2312–2320, 2011.

[APS12] Yasin Abbasi-Yadkori, David Pal, and Csaba Szepesvari. "Online-to-Confidence-Set Conversions and Application to Sparse Stochastic Bandits." In *International Conference on Artificial Intelligence and Statistics*, pp. 1–9, 2012.

[AS16] Onur Atan and Mihaela van der Schaar. "Data-Driven Online Decision Making with Costly Information Acquisition." *arXiv preprint arXiv:1602.03600*, 2016.

[AS17] Ahmed M Alaa and Mihaela van der Schaar. "Bayesian Inference of Individualized Treatment Effects using Multi-task Gaussian Processes." In *Advances in Neural Information Processing Systems (NIPS)*, 2017.

[ATS15a] Onur Atan, Cem Tekin, and Mihaela Schaar. "Global multi-armed bandits with Hölder continuity." In *Artificial Intelligence and Statistics*, pp. 28–36, 2015.

[ATS15b] Onur Atan, Cem Tekin, and Mihaela van der Schaar. "Global multi-armed bandits with Hölder continuity." In *Proc. AISTATS*, pp. 28–36, 2015.

[ATS18] Onur Atan, Cem Tekin, and Mihaela van der Schaar. "Global bandits." *IEEE Transactions on Neural Networks and Learning Systems*, 2018.

[Aue02] P. Auer. "Using confidence bounds for exploitation-exploration trade-offs." *Journal of Machine Learning Research*, pp. 397–422, 2002.

[AZF16] Onur Atan, William R Zame, Qiaojun Feng, and Mihaela van der Schaar. "Constructing Effective Personalized Policies Using Counterfactual Inference from Biased Data Sets with Many Features." *arXiv preprint arXiv:1612.08082*, 2016.

[BBC07] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. "Analysis of representations for domain adaptation." In *Advances in neural information processing systems*, pp. 137–144, 2007.

[BC12] S. Bubeck and N. Cesa Bianchi. "Regret Analysis of Stochastic and Nonstochastic Multi-armed Bandit Problems." *Machine Learning*, 2012.

[BCK08] John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. "Learning bounds for domain adaptation." In *Advances in neural information processing systems*, pp. 129–136, 2008.

[BL09] Alina Beygelzimer and John Langford. "The offset tree for learning with partial labels." In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 129–138, 2009.

[BL13] S. Bubeck and C. Y. Liu. "Prior-free and prior-dependent regret bounds for Thompson Sampling." In *Advances in Neural Information Processing Systems*, pp. 638–646. 2013.

[BPC13] Léon Bottou, Jonas Peters, Joaquin Quinonero Candela, Denis Xavier Charles, Max Chickering, Elon Portugaly, Dipankar Ray, Patrice Y Simard, and Ed Snelson. "Counterfactual reasoning and learning systems: the example of computational advertising." *Journal of Machine Learning Research*, **14**(1):3207–3260, 2013.

[BV04] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[CBM04] Christopher P Cannon, Eugene Braunwald, Carolyn H McCabe, Daniel J Rader, Jean L Rouleau, Rene Belder, Steven V Joyal, Karen A Hill, Marc A Pfeffer, and Allan M Skene. "Intensive versus moderate lipid lowering with statins after acute coronary syndromes." *New England journal of medicine*, **350**(15):1495–1504, 2004.

[CK11] Nicolo Cesa-Bianchi and Sham Kakade. "An optimal algorithm for linear bandits." *arXiv preprint arXiv:1110.4322*, 2011.

[CLR11] Wei Chu, Lihong Li, Lev Reyzin, and Robert E Schapire. "Contextual Bandits with Linear Payoff Functions." In *Proc. AISTATS*, volume 15, pp. 208–214, 2011.

[CLZ13] Xi Chen, Qihang Lin, and Dengyong Zhou. "Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing." In *International Conference on Machine Learning*, pp. 64–72, 2013.

[CSS11] Nicolo Cesa-Bianchi, Shai Shalev Shwartz, and Ohad Shamir. "Efficient learning with partially observed attributes." *Journal of Machine Learning Research*, **12**:2857–2878, 2011.

[CWY13] Wei Chen, Yajun Wang, and Yang Yuan. "Combinatorial multi-armed bandit: General framework, results and applications." In *Proc. ICML*, pp. 151–159, 2013.

[Dau09] Hal Daumé III. "Frustratingly easy domain adaptation." *arXiv preprint arXiv:0907.1815*, 2009.

[DB04] Jennifer G Dy and Carla E Brodley. "Feature selection for unsupervised learning." *Journal of machine learning research*, **5**(845–889), 2004.

[DHK08] Varsha Dani, Thomas P Hayes, and Sham M Kakade. "Stochastic Linear Optimization under Bandit Feedback." In *Proc. COLT*, pp. 355–366, 2008.

[DHK11] Miroslav Dudik, Daniel Hsu, Satyen Kale, Nikos Karampatziakis, John Langford, Lev Reyzin, and Tong Zhang. "Efficient optimal learning for contextual bandits." *arXiv preprint arXiv:1106.2369*, 2011.

[DHS12] Richard O Duda, Peter E Hart, and David G Stork. *Pattern classification*. John Wiley & Sons, 2012.

[DLL11] Miroslav Dudík, John Langford, and Lihong Li. "Doubly robust policy evaluation and learning." In *International Conference on Machine Learning (ICML)*, 2011.

[FCG11] S. Filippi, O. Cappe, A. Garivier, and C. Szepesvari. "Parametric bandits: The generalized linear case." In *Advances in Neural Information Processing Systems*, pp. 586–594, 2011.

[FFD98] Lawrence M Friedman, Curt Furberg, David L DeMets, David Reboussin, and Christopher B Granger. *Fundamentals of clinical trials*, volume 3. Springer, 1998.

[FPD08] Peter I Frazier, Warren B Powell, and Savas Dayanik. "A knowledge-gradient policy for sequential information collection." *SIAM Journal on Control and Optimization*, **47**(5):2410–2439, 2008.

[FPD09] Peter Frazier, Warren Powell, and Savas Dayanik. "The knowledge-gradient policy for correlated normal beliefs." *INFORMS journal on Computing*, **21**(4):599–613, 2009.

[GC11] A. Garivier and O. Cappe. "The KL-UCB algorithm for bounded stochastic bandits and beyond." In *Proc. COLT*, 2011.

[GGW11] John Gittins, Kevin Glazebrook, and Richard Weber. *Multi-armed bandit allocation indices*. John Wiley & Sons, 2011.

[GK10] Daniel Golovin and Andreas Krause. "Adaptive Submodularity: A New Approach to Active Learning and Stochastic Optimization." In *COLT*, pp. 333–345, 2010.

[GKJ12] Y. Gai, B. Krishnamachari, and R. Jain. "Combinatorial network optimization with unknown variables: Multi-armed bandits with linear rewards and individual observations." *IEEE/ACM Trans. Netw*, **20**(5):1466–1478, 2012.

[Gro01] Age-Related Eye Disease Study Research Group et al. "A randomized, placebo-controlled, clinical trial of high-dose supplementation with vitamins C and E, beta carotene, and zinc for age-related macular degeneration and vision loss." *Archives of ophthalmology*, **119**(10):1417, 2001.

[GUA16] Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario Marchand, and Victor Lempitsky. "Domain-adversarial training of neural networks." *The Journal of Machine Learning Research*, **17**(1), 2016.

[Hal99] Mark A Hall. *Correlation-based feature selection for machine learning*. PhD thesis, The University of Waikato, 1999.

[HCN05] Xiaofei He, Deng Cai, and Partha Niyogi. "Laplacian score for feature selection." In *Advances in neural information processing systems*, pp. 507–514, 2005.

[Hil11] Jennifer L Hill. "Bayesian nonparametric modeling for causal inference." *Journal of Computational and Graphical Statistics*, **20**(1), 2011.

[HK12] Elad Hazan and Tomer Koren. "Linear Regression with Limited Observation." In *Proc. 29th Int. Conf. on Machine Learning*, pp. 807–814, 2012.

[HR06a] Jennifer Hill and Jerome P Reiter. "Interval estimation for treatment effects using propensity score matching." *Statistics in Medicine*, **25**(13):2230–2256, 2006.

[HR06b] Feifang Hu and William F Rosenberger. *The theory of response-adaptive randomization in clinical trials*, volume 525. John Wiley & Sons, 2006.

[HWB99] Paul C Hébert, George Wells, Morris A Blajchman, John Marshall, Claudio Martin, Giuseppe Pagliarello, Martin Tweeddale, Irwin Schweitzer, Elizabeth Yetisir, and Transfusion Requirements in Critical Care Investigators for the Canadian Critical Care Trials Group. "A multicenter, randomized, controlled clinical trial

of transfusion requirements in critical care." *New England Journal of Medicine*, **340**(6):409–417, 1999.

[Ion08] Edward L Ionides. "Truncated importance sampling." *Journal of Computational and Graphical Statistics*, **17**(2):295–311, 2008.

[IW09] Guido W Imbens and Jeffrey M Wooldridge. "Recent developments in the econometrics of program evaluation." *Journal of economic literature*, **47**(1):5–86, 2009.

[JL16] Nan Jiang and Lihong Li. "Doubly Robust Off-policy Evaluation for Reinforcement Learning." In *International Conference on Machine Learning (ICML)*, 2016.

[JOA10] Thomas Jaksch, Ronald Ortner, and Peter Auer. "Near-optimal Regret Bounds for Reinforcement Learning." *Journal of Machine Learning Research*, **11**:1563–1600, 2010.

[JS16] Thorsten Joachims and Adith Swaminathan. "Counterfactual Evaluation and Learning for Search, Recommendation and Ad Placement." In *International ACM SIGIR conference on Research and Development in Information Retrieval*, pp. 1199–1201, 2016.

[JSS16] Fredrik Johansson, Uri Shalit, and David Sontag. "Learning Representations for Counterfactual Inference." In *International Conference on Machine Learning (ICML)*, 2016.

[Kal01] Dan Kalman. "A generalized logarithm for exponential-linear equations." *The College Mathematics Journal*, **32**(1):2, 2001.

[KB14] Diederik Kingma and Jimmy Ba. "Adam: A method for stochastic optimization." In *International Conference on Learning Representations*, 2014.

[KKE13] N. Korda, E. Kaufmann, and R. E., Munos. "Thompson Sampling for 1-Dimensional Exponential Family Bandits." In *Advances in Neural Information Processing Systems*, 2013.

[KOG12] E. Kaufmann, Cappe O., and A. Garivier. "On Bayesian upper confidence bounds for bandit problems." In *Proc. AISTATS*, 2012.

[KR92] Kenji Kira and Larry A Rendell. "A practical approach to feature selection." In *Proceedings of the ninth international workshop on Machine learning*, pp. 249–256, 1992.

[KS96] Daphne Koller and Mehran Sahami. "Toward optimal feature selection." 1996.

[LCL10] Lihong Li, Wei Chu, John Langford, and Robert E Schapire. "A contextual-bandit approach to personalized news article recommendation." In *Proceedings of the 19th international conference on World wide web*, pp. 661–670, 2010.

[LM14] Tor Lattimore and Rémi Munos. "Bounded Regret for Finite-Armed Structured Bandits." In *Advances in Neural Information Processing Systems*, pp. 550–558, 2014.

[LMW88] John M Lachin, John P Matts, and LJ Wei. "Randomization in clinical trials: conclusions and recommendations." *Controlled clinical trials*, **9**(4):365–374, 1988.

[LPP10] Tyler Lu, Dávid Pál, and Martin Pál. "Contextual multi-armed bandits." In *International Conference on Artificial Intelligence and Statistics*, pp. 485–492, 2010.

[LR78] TL Lai and Herbert Robbins. "Adaptive design in regression and control." *Proceedings of the National Academy of Sciences*, **75**(2):586–587, 1978.

[LR85] T. Lai and H. Robbins. "Asymptotically efficient adaptive allocation rules." *Advances in applied mathematic*, **6**(1):4–22, 1985.

[LZ07] John Langford and Tong Zhang. "The epoch-greedy algorithm for contextual multi-armed bandits." *Advances in Neural Information Processing Systems (NIPS)*, **20**:1096–1103, 2007.

[LZ08] J. Langford and T. Zhang. "The epoch-greedy algorithm for contextual multi-armed bandits." In *Advances in Neural Information Processing Systems*, pp. 1096–1023, 2008.

[LZ13] K. Liu and Q. Zhao. "Distributed Learning in Multi-Armed Bandit with Multiple Players." *IEEE Trans. Signal Process.*, **58**:5547–5567, 2013.

[MHS12] David Moher, Sally Hopewell, Kenneth F Schulz, Victor Montori, Peter C Gøtzsche, PJ Devereaux, Diana Elbourne, Matthias Egger, and Douglas G Altman. "CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials." *International Journal of Surgery*, **10**(1):28–55, 2012.

[MP09] A Maurer and M Pontil. "Empirical bernstein bounds and sample variance penalization." In *The 22nd Conference on Learning Theory*, 2009.

[MRT09] A. Mersereau, P. Rusmevichientong, and J.N. Tsitsiklis. "A structured multi-armed bandit problem and the greedy policy." *IEEE Trans. Autom. Control*, **54**:2787–2802, 2009.

[MS11] Shie Mannor and Ohad Shamir. "From bandits to experts: On the value of side-observations." In *Advances in Neural Information Processing Systems*, pp. 684–692, 2011.

[OA07] P Ortner and R Auer. "Logarithmic online regret bounds for undiscounted reinforcement learning." In *Advances in Neural Information Processing Systems*, 2007.

[OVW16] Ian Osband, Benjamin Van Roy, and Zheng Wen. "Generalization and Exploration via Randomized Value Functions." In *International Conference on Machine Learning*, 2016.

[Pea17] Judea Pearl. "Detecting latent heterogeneity." *Sociological Methods & Research*, **46**(3):370–389, 2017.

[PLD05] Hanchuan Peng, Fulmi Long, and Chris Ding. "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy." *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, **27**(8):1226–1238, 2005.

[Pre76] Ross Prentice. "Use of the logistic model in retrospective studies." *Biometrics*, pp. 599–606, 1976.

[RK03] Marko Robnik-Šikonja and Igor Kononenko. "Theoretical and empirical analysis of ReliefF and RReliefF." *Machine learning*, **53**(1-2):23–69, 2003.

[ROP12] Ernst-Wilhelm Radue, Paul O'connor, Chris H Polman, Reinhard Hohlfeld, Peter Calabresi, Krystof Selmaj, Nicole Mueller-Lenke, Catherine Agoropoulou, Frederick Holdbrook, Ana De Vera, et al. "Impact of fingolimod therapy on magnetic resonance imaging outcomes in patients with multiple sclerosis." *Archives of neurology*, **69**(10):1259–1269, 2012.

[RR83] Paul R Rosenbaum and Donald B Rubin. "The central role of the propensity score in observational studies for causal effects." *Biometrika*, **70**(1):41–55, 1983.

[RT10] P. Rusmevichientong and J.N. Tsitsiklis. "Linearly Parameterized Bandits." *Mathematics of Operations Research*, **5**:395–411, 2010.

[Rub05] Donald B Rubin. "Causal inference using potential outcomes: Design, modeling, decisions." *Journal of the American Statistical Association*, **100**(469):322–331, 2005.

[RV15] Daniel Russo and Benjamin Van Roy. "An information-theoretic analysis of Thompson sampling." *Journal of Machine Learning Research*, 2015.

[RZ10] Philippe Rigollet and Assaf Zeevi. "Nonparametric Bandits with Covariates." In *Proc. COLT*, 2010.

[SBB07] Debbie Saslow, Carla Boetes, Wylie Burke, Steven Harms, Martin O Leach, Constance D Lehman, Elizabeth Morris, Etta Pisano, Mitchell Schnall, Stephen Sener, et al. "American Cancer Society guidelines for breast screening with MRI as an adjunct to mammography." *CA: a cancer journal for clinicians*, **57**(2):75–89, 2007.

[SG02] Kenneth F Schulz and David A Grimes. "Allocation concealment in randomised trials: defending against deciphering." *The Lancet*, **359**(9306):614–618, 2002.

[SJ15a] Adith Swaminathan and Thorsten Joachims. "Batch learning from logged bandit feedback through counterfactual risk minimization." *Journal of Machine Learning Research*, **16**:1731–1755, 2015.

[SJ15b] Adith Swaminathan and Thorsten Joachims. "Counterfactual Risk Minimization: Learning from Logged Bandit Feedback." In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 814–823, 2015.

[SJ15c] Adith Swaminathan and Thorsten Joachims. "The self-normalized estimator for counterfactual learning." In *Advances in Neural Information Processing Systems*, pp. 3231–3239, 2015.

[SJS17] Uri Shalit, Fredrik Johansson, and David Sontag. "Estimating individual treatment effect: generalization bounds and algorithms." In *International Conference on Machine Learning (ICML)*, 2017.

[Sli11] Aleksandrs Slivkins. "Contextual Bandits with Similarity Information." In *24th Annual Conference On Learning Theory*, 2011.

[Sli14a] A. Slivkins. "Contextual Bandits with Similarity Information." In *Journal of Machine Learning Research*, volume 15, pp. 2533–2568, 2014.

[Sli14b] Aleksandrs Slivkins. "Contextual bandits with similarity information." *Journal of Machine Learning Research*, **15**(1):2533–2568, 2014.

[SLL10] Alex Strehl, John Langford, Lihong Li, and Sham M Kakade. "Learning from logged implicit exploration data." In *Advances in Neural Information Processing Systems*, pp. 2217–2225, 2010.

[SSG12] Le Song, Alex Smola, Arthur Gretton, Justin Bedo, and Karsten Borgwardt. "Feature selection via dependence maximization." *Journal of Machine Learning Research*, **13**(May):1393–1434, 2012.

[STS16] Linqi Song, Cem Tekin, and Mihaela van der Schaar. "Online learning in large-scale contextual recommender systems." *IEEE Transactions on Services Computing*, **9**(3):433–445, 2016.

[TAG12] L Tian, A Alizadeh, A Gentles, and R Tibshirani. "A Simple Method for Detecting Interactions between a Treatment and a Large Number of Covariates." *arXiv preprint arXiv:1212.2995*, 2012.

[TAL14] Jiliang Tang, Salem Alelyani, and Huan Liu. "Feature selection for classification: A review." *Data Classification: Algorithms and Applications.*, 2014.

[Tho33] W. R. Thompson. "On the likelihood that one unknown probability exceeds another in view of the evidence of two samples." *Biometrika*, pp. 285–294, 1933.

[TS14] Cem Tekin and Mihaela van der Schaar. "Discovering, Learning and Exploiting Relevance." In *Advances in Neural Information Processing Systems*, pp. 1233–1241, 2014.

[TS15a] Cem Tekin and Mihaela van der Schaar. "Distributed Online Learning via Cooperative Contextual Bandits." *IEEE Trans. Signal Process.*, **63**(14):3700–3714, 2015.

[TS15b] Cem Tekin and Mihaela van der Schaar. "Distributed online learning via cooperative contextual bandits." *IEEE Transactions on Signal Processing*, **63**(14):3700–3714, 2015.

[TS15c] Cem Tekin and Mihaela van der Schaar. "RELEAF: An algorithm for learning and exploiting relevance." *IEEE Journal of Selected Topics in Signal Processing*, **9**(4):716–727, 2015.

[TYS17] Cem Tekin, Jinsung Yoon, and Mihaela van der Schaar. "Adaptive ensemble learning with confidence bounds." *IEEE Transactions on Signal Processing*, **65**(4):888–903, 2017.

[TZ09] Alexandre B Tsybakov and Vladimir Zaiats. *Introduction to nonparametric estimation*. Springer, 2009.

[VBW15] Sofía S Villar, Jack Bowden, and James Wason. "Multi-armed bandit models for the optimal design of clinical trials: benefits and challenges." *Statistical science: a review journal of the Institute of Mathematical Statistics*, **30**(2):199, 2015.

[WA15] Stefan Wager and Susan Athey. "Estimation and inference of heterogeneous treatment effects using random forests." *arXiv preprint arXiv:1510.04342*, 2015.

[WES03] Jason Weston, André Elisseeff, Bernhard Schölkopf, and Mike Tipping. "Use of the zero-norm with linear models and kernel methods." *Journal of machine learning research*, **3**:1439–1461, 2003.

[Whi80] Peter Whittle. "Multi-armed bandits and the Gittins index." *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 143–149, 1980.

[WOS03] Tsachy Weissman, Erik Ordentlich, Gadiel Seroussi, Sergio Verdu, and Marcelo J Weinberger. "Inequalities for the L1 deviation of the empirical distribution." *Hewlett-Packard Labs, Tech. Rep*, 2003.

[XKL10] Zenglin Xu, Irwin King, Michael Rung-Tsong Lyu, and Rong Jin. "Discriminative semi-supervised feature selection via manifold regularization." *IEEE Transactions on Neural Networks*, **21**(7):1033–1047, 2010.

[XTZ15] Jie Xu, Cem Tekin, Simpson Zhang, and Mihaela van der Schaar. "Distributed Online Learning Based on Global Feedback." *IEEE Trans. Signal Process.*, **63**(9):2225–2238, 2015.

[YDS16] J Yoon, C Davtyan, and M van der Schaar. "Discovery and Clinical Decision Support for Personalized Healthcare." *IEEE journal of biomedical and health informatics*, 2016.

[YL03] Lei Yu and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." In *International Conference on Machine Learning (ICML)*, volume 3, pp. 856–863, 2003.

[ZBG13] Navid Zolghadr, Gábor Bartók, Russell Greiner, András György, and Csaba Szepesvári. "Online learning with costly features and labels." In *Advances in Neural Information Processing Systems (NIPS)*, pp. 1241–1249, 2013.

[ZLM14] Daphney-Stavroula Zois, Marco Levorato, and Urbashi Mitra. "Active classification for POMDPs: A Kalman-like state estimator." *IEEE Transactions on Signal Processing*, **62**(23):6209–6224, 2014.

[ZM17] Daphney-Stavroula Zois and Urbashi Mitra. "Active State Tracking With Sensing Costs: Analysis of Two-States and Methods for n-States." *IEEE Transactions on Signal Processing*, **65**(11):2828–2843, 2017.

[ZSM13] Kun Zhang, Bernhard Schölkopf, Krikamol Muandet, and Zhikun Wang. "Domain adaptation under target and conditional shift." In *International Conference on Machine Learning*, pp. 819–827, 2013.