ROBUST LEARNING AND EVALUATION IN SEQUENTIAL DECISION MAKING

A DISSERTATION
SUBMITTED TO THE INSTITUTE OF COMPUTATIONAL AND MATHEMATICAL ENGINEERING
AND THE COMMITTEE ON GRADUATE STUDIES
OF STANFORD UNIVERSITY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Ramtin Keramati

June 2021

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Emma Brunskill) Principal Adviser

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Benjamin Van Roy)

I certify that I have read this dissertation and that, in my opinion, it is fully adequate in scope and quality as a dissertation for the degree of Doctor of Philosophy.

(Marco Pavone)

Approved for the Stanford University Committee on Graduate Studies

Abstract

Reinforcement learning (RL), as a branch of artificial intelligence, is concerned with making a good sequence of decisions given experience and rewards in a stochastic environment. RL algorithms, propelled by the rise of deep learning and neural networks, have shown an impressive performance in achieving human-level performance in games like Go, Chess, and Atari. However, when applied to high-stakes real-world applications, these impressive performances are not matched. This dissertation tackles some important challenges around robustness that hinder our ability to unleash the potential of RL to real-world applications. We look at the robustness of RL algorithms in both online and offline settings and introduce new algorithms that may be of particular interest when applying RL to real-world applications such as health care and education.

In the first line of work, we consider an online setting where the agent can interact with the environment and collect experience and rewards to learn the optimal sequence of decisions. In many real-world applications, online interactions are limited, limiting our ability to collect data. That raises the necessity of sample efficient algorithms. In addition, safety concerns highlight the importance of learning risk-sensitive policies in these applications. We, therefore, combine recent advances in distributional reinforcement learning with the principle of optimism in the face of uncertainty to develop a scalable algorithm to learn a CVaR (conditional value at risk) optimal policy in a sample efficient manner to minimize the number of interactions needed with the environment.

In high-stakes real-world applications, often any online interaction is undesirable and we have to be able to perform off-policy policy evaluation (OPE). OPE methods evaluate a new policy (evaluation policy) given the experiences collected using another policy (behavior policy). For example, an agent that aims to learn the adaptive treatment plan for patients in a hospital may not be able to collect any experiences interacting with patients but can use data containing past decisions made by clinicians and their outcomes. OPE is a counterfactual and challenging task that is often solved by making a crucial assumption, sequential ignorability. Sequential ignorability states that the evaluation policy has access to all the information used by the behavior policy to make decisions, in other words, there are no unobserved confounders. This assumption is often violated in observational data, and failure to acknowledge that results in an arbitrary biased estimate of the evaluation policy. In this dissertation, we consider the bounded effect of unobserved confounders and develop a scalable

algorithm to provide bounds on OPE. Our work can be used to raise concerns or certify the superior performance of an evaluation policy under the existence of unobserved confounders and prevents undesirable outcomes of deploying a new decision policy.

One shortcoming of the existing OPE method for sequential decision-making is that they often evaluate the expected performance given a distribution. However, in most real-world applications, we would like to assess if the population's subgroups benefit from a newly suggested policy. In this dissertation, we take a step toward quantifying heterogeneity in OPE for sequential decision making and identify subgroups with similar benefits or harm from the evaluation policy. This information provides essential insight into the performance of the evaluation policy that domain experts can use before deploying a policy.

Acknowledgments

First, I would like to express my appreciation to my advisor Emma Brunskill for her tremendous guidance and support throughout my Ph.D. She taught me how to conduct scientific research and convey my ideas to an audience of researchers. Despite her busy schedule, she always allocated time for discussing research ideas. She always encouraged me to consider the broader impact of my research.

I am thankful to Professor Benjamin Van Roy and Professor Marco Pavone for being part of my reading committee, despite their busy schedule. I would also like to thank Professor Philip Thomas, who kindly agreed to be an oral examiner in my oral examination. Additionally, I am very thankful to Professor Mohsen Bayati for chairing my oral examination.

I am thankful to my fascinating collaborators at Stanford, who made my Ph.D. more fruitful: Christoph Dann, Alex Tamkin, Steve Yadlowsky, Hongseok Namkoong, Evan Liu, Omer Gottesman, Yao Liu, and Finale Doshi-Velez. I am grateful for the help of Indira Choudhury and my friends in the Institute of Computational and Mathematical Engineering (ICME) for helping me navigate through my Ph.D.

I am grateful to my family, who supported me throughout the process with their unconditional love. My friends played an important role in my life during my Ph.D. they offered a tremendous amount of help, support, and love, for which I am very grateful: Behrad Habib Afshar, Khashayar Khosravi, Milad Sharif, Mahdis Mahdieh, Kamyar Azizzadenesheli, Kian Katanforoosh, Victor Saade, Dylan Bourgeois, Sherrie Wang, and Poorya Mehrabinia.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Motivation

Reinforcement learning (RL) is a branch of machine learning and artificial intelligence that is concerned with making an optimal sequence of decisions under uncertainty. The general framework of RL consists of an environment and an agent that interacts with the environment. At every step, the agent observes a state, takes an action, and receives a reward. The goal is to find an optimal sequence of actions to perform a specific task.

This framework applies to many real-world applications. For example, playing games such as Chess [Silver et al., 2018], Go [Silver et al., 2017] and Atari [Mnih et al., 2013] can be formulated as a reinforcement learning problem where the agent's goal is to win the game by taking actions with respect to the rule of the game. The application of RL goes beyond games, for example in health care one can use RL agents to learn an adaptive optimal treatment plan for septic patients in ICU [Komorowski et al., 2018] or find the optimal STI (structured treatment interruption) strategies for HIV patients [Ernst et al., 2006]. In education, RL agents can be used to adaptively sequence material to best keep a student engaged [Mandel et al., 2014]. RL framework can also be used to personalize recommendation systems [Li et al., 2010] and generate dialogues for interactive dialogue systems such as chatbots [Li et al., 2016].

In designing algorithms for reinforcement learning, we need to overcome four main challenges: optimization, exploration, generalization, and delayed feedback.

Optimization: Optimization means that we want the RL agent to find the optimal sequence of actions with respect to some pre-defined objective function. This objective function is often being defined as a risk-neutral expected value of the sum of discounted rewards. Optimization usually involves searching over the policy space, where a policy is defined as a mapping between states and distribution over actions.

Exploration: In reinforcement learning, the agent will only observe a scalar value reward feedback

for the chosen action and no feedback for other actions. Hence, the agent needs to explore different actions to find the optimal action. This is in contrast with supervised learning where the training data indicates the true label of the observation.

Generalization: Often it is impossible to observe all the possible states in an environment. Therefore, we expect that the RL agent acts well in unseen states and can generalize from past experiences. This generalization from observed data is often achieved by the use of function approximation such as neural networks.

Delayed feedback: Another unique challenge to reinforcement learning is that the effect of an action may not be reflected in the immediate reward feedback received by the agent. For example, a patient that undergoes surgery will receive the reward of better health in the future and not in the immediate aftermath of the surgery. This challenge makes the application of myopic algorithms (algorithms that only look at the immediate reward to optimize the long-term sum of rewards) sub-optimal.

Despite these challenges, there has been an impressive empirical success using reinforcement learning algorithms, mostly propelled by the use of neural networks for function approximation. Examples include, super-human performance in Atari [Mnih et al., 2013], human-level performance in the game of Go, Chess, Shogi [Silver et al., 2018], beating the professional players in Start Craft II [Vinyals et al., 2019], and solving Rubik's cube with a robotic hand [Akkaya et al., 2019]. However; when looking at the results of RL algorithms on other applications besides playing a game and robotics, although impressive, these performances are not matched [Gottesman et al., 2019a]. Common property in those applications is either access to a perfect simulator or the ability to collect large amounts of data by running our algorithms online to combat the challenges that we mentioned earlier. However, when applying RL to high-stakes real-world applications like health care and education we do not have access to a high fidelity simulator. Additionally, the availability of interactions and safety concerns hinder our ability to collect a large amount of data. For example, for an agent that tries to learn an adaptive treatment plan for patients, the agent has limited access to interaction with patients, and, it is crucial to avoid catastrophic outcomes in this setting.

To unlock the potential of reinforcement learning for high-stake applications such as health care and education we need to think about the robustness of algorithms and the ability to learn from limited interaction or historical data. In an online setting where the agent is allowed to interact with the environment, it is important to optimize a risk-sensitive policy rather than a risk-neutral policy to minimize the chances of the severe outcome as well as minimizing the number of interactions needed to learn those policies. In an offline setting, the agent is not allowed to interact with the environment and has to use historically collected data to suggest a policy better than common practice. In this setting, a thorough examination of the suggested policy to satisfy the safety criteria that ensure robustness is of prime importance.

1.2 Robust Policy Learning and Evaluation

The goal of this dissertation is to tackle some of the important challenges around robustness of reinforcement learning algorithms in both online and offline settings to make it more applicable to high-stakes real-world applications and unleash the power of reinforcement learning framework. Throughout this dissertation we consider episodic reinforcement learning where the agent acts in episodes and takes action until the end of the episode and starts again. In an online setting, the agent interacts with the environment and collects samples of states, actions, and rewards in every episode and tries to improve its performance. On the other hand, in an offline setting, data is already collected using other agents such as other automated agents or humans acting in the environment and the goal is to use the historical data to learn a better decision policy or evaluate a new decision policy.

In an online setting, the agent can interact with the environment. When acting in high-stakes applications such as health care, the agent should avoid severe outcomes as a consequence of its actions. Hence, the consideration of risk-sensitive objectives in contrast to risk-neutral objectives is important. Previous works have considered altering the reward function to account for severe outcomes [Mihatsch and Neuneier, 2002], or optimization of risk measures such VaR (Value at Risk) and CVaR (Conditional Value at Risk) [Chow et al., 2015, 2017] as an alternative. Two main approaches can be taken, first, optimizing a risk-sensitive objective, second optimizing a risk-neutral objective subject to a risk constraint [Chow, 2017]. In either case, an important factor that is often overlooked is the sample efficiency of these algorithms. Learning a safe policy fast is of prime importance when dealing with real-world applications, as the number of feasible interactions is often limited.

Most of the past literature in risk-sensitive RL often suffers from two main shortcomings. First, they are not scalable to large state spaces when using neural networks as function approximators. Second, they do not offer a strategic exploration scheme for sample efficient learning. In this dissertation, we combine the principle of optimism in the face of uncertainty (OFU) [Brafman and Tennenholtz, 2002] and recent advances in distributional reinforcement learning (DRL) [Dabney et al., 2018, Bellemare et al., 2017] to develop a sample efficient scalable algorithm for learning a risk-sensitive policy. This is in contrast with safe policy learning, where we focus on avoiding sever outcomes while learning.

In applications where it is not feasible to interact with the environment, we need to work in an offline setting. The goal is to evaluate a new decision policy (evaluation policy) using historical data collected by a different policy, the behavior policy. We call this task off-policy policy evaluation (OPE). OPE is inherently a counterfactual and challenging problem where we ask the question of "what if the evaluation policy were used instead of the behavior policy?". A common method for OPE is importance sampling that reweights the historical data based on their likelihood under the evaluation policy and the behavior policy [Mahmood et al., 2014]. Other common methods are

model-based methods that learn a model of the world based on the historical data and evaluates the policy given the learned model, and doubly-robust methods that use the model-based estimate to reduce the variance of importance sampling estimate [Jiang and Li, 2015].

There are some key challenges in applying off-policy policy evaluation to real-world application; confounding factors, heterogeneity, overlap, and ... [Gottesman et al., 2019a]. In this dissertation we tackle two of these limitations.

Confounding factors: Does the RL agent have access to all the variables being used by the behavior policy to make the decision? For example, consider the clinician's decision to administer antibiotics to septic patients upon admission to the ICU. Clinicians may decide whether a patient is healthy based on unrecorded comorbidities or unrecorded conversations with the patient. It is more probable that healthier patients receive a lower dose of antibiotics and they are more likely to show better outcomes. Without access to all the information that the clinician used to make the decision, a lower dose of antibiotics may appear to yield a better outcome but this association may be due to spurious correlation. Failure to consider the effect of unobserved confounders may result in drastic estimation errors in evaluating a new decision policy. Developing an algorithm to measure the robustness of the evaluation process to unobserved confounders can help inform whether to deploy a new decision policy in high-stakes applications such as health care.

In one-time step, non-sequential settings there have been recent works on bounding the value of the treatment effect in observational studies under the assumption of a bounded effect of the unobserved confounder. [Yadlowsky et al., 2018, Rosenbaum, 2014]. Unfortunately, there is very limited work in sequential settings and these methods are not readily applicable to longer horizons. Zhang and Bareinboim [2019] for example derived partial identification bounds on policy performance with limited restrictions on the influence of the unobserved confounder. However, these bounds are too conservative and not scalable to large state spaces. In this dissertation, we take a step toward developing tighter bounds on the evaluation policy in sequential decision-making settings. This bound can be efficiently computed and is scalable to larger state space with the use of function approximators such as neural networks.

Heterogeneity: Does the evaluation policy treat everyone the same? We may estimate that an evaluation policy is better than the behavior policy on average; however, that does not necessarily mean that it is better for everyone. That raises a concern when deploying the evaluation policy. For example, if the historical data is skewed toward one gender, we may evaluate the evaluation policy to be better than the behavior policy but it may be worse for the other gender group. This raises a concern about the heterogeneous effect of the evaluation policy. Do we know subgroups of the population that will benefit or suffer from deploying the evaluation policy? In larger state spaces with multiple features, it is unclear how to discover and evaluate the evaluation policy for these subgroups.

A great deal of work has been done in non-sequential settings to estimate heterogeneous treatment

effect (see e.g. [Athey and Imbens, 2016, Athey and Wager, 2019, Xie et al., 2012]). The majority of these works use recursive partitioning and a tree structure to uncover subgroups with similar treatment effects while having a large feature set. These methods show impressive performance in discovering heterogeneous treatment effects; however, due to challenges like overlap that we discuss shortly, they are not readily applicable to longer horizons.

Overlap: To evaluate the evaluation policy, we rely on samples collected by the behavior policy and reweights those samples based on their likelihood under these policies (importance sampling). Intuitively, when the behavior and the evaluation policy are very different, in other words, they induce very different distributions over the state-action space, samples of the behavior policy are less useful to evaluate the evaluation policy. For example, consider data collected by a clinician that never administers antibiotics in ICU to patients, it is impossible to evaluate the benefit of antibiotics using those samples. The effective sample size is a measure of overlap which indicates the effective number of samples we have to evaluate the evaluation policy and is inversely correlated with the variance of our estimate. For example, if the evaluation policy and the behavior policy are close to each other the effective sample size is higher and we can estimate the value of the evaluation policy better. However, for many evaluation policies, small overlap and small effective sample size hinder our ability to obtain a low variance estimate of their value.

Small overlap and high variance estimates of the value of the evaluation policy limits the use of non-sequential setting heterogeneous treatment effect estimators such as the one mentioned earlier for sequential settings. In this dissertation, we take a step towards developing an algorithm to identify subgroups and estimate the heterogeneous treatment effect in sequential settings. We leverage a different viewpoint, and instead of asking "what is the treatment effect for each individual", we ask "what are the subgroups that we can effectively evaluate their treatment effect". This allows us to focus on the part of the state space where we can obtain a low variance estimate of subgroups.

1.3 Thesis Contributions and Outline

This dissertation addresses some of the aforementioned challenges in making reinforcement learning more applicable to high-stakes real-world applications. At the high level, we address challenges around the robustness of RL in both online and offline settings.

- Chapter 2 (Sample Efficient Risk Sensitive Policy Learning): This chapter covers robustness in the online setting. We cover risk-sensitive planning along with distributional reinforcement learning and the principle of optimism in the face of uncertainty. We further develop a novel algorithm for efficient exploration for learning a CVaR (Conditional Value at Risk) optimal policy. Our algorithm leverages Dvoretzky–Kiefer–Wolfowitz (DKW) concentration inequality [Dvoretzky et al., 1956] to obtain an optimistic estimate of CVaR.

  Similarly, in K-arm bandit setting we propose a novel algorithm to learn a CVaR optimal

arm. Our method leverages DKW inequality to obtain an optimistic estimate of CVaR for each arm. We show that our method have logarithmic regret. This chapter is published in the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20) [Keramati et al., 2020]: *Keramati R, Dann C, Tamkin A, Brunskill E. Being optimistic to be conservative: Quickly learning a cvar policy. InProceedings of the AAAI Conference on Artificial Intelligence 2020 Apr 3 (Vol. 34, No. 04, pp. 4436-4443).*

- Chapter 3 (Off-Policy Policy Evaluation Under Unobserved Confounding): We discuss one of the main assumptions of off-policy policy evaluation, sequential ignorability [Chakraborty and Murphy, 2014]. We demonstrate practical examples where this assumption is violated, or in other words when we have unobserved confounders. We describe the failure of common OPE methods when sequential ignorability doesn't hold: under no assumption OPE can be arbitrary biased.

  We introduce bounded confounding model for sequential settings, and assume confounding happens only in one time step. We further develop a scalable algorithm to estimate bounds on OPE under unobserved confounding. Our algorithm can be used in large state space when using function approximators such as neural networks. Additionally, we show our estimator is consistent under some regularity assumptions. This chapter is published in Thirty-fourth Conference on Neural Information Processing Systems (NeurIPS-2020)[Namkoong et al., 2020]: *Namkoong H*, Keramati R*, Yadlowsky S*, Brunskill E. Off-policy Policy Evaluation For Sequential Decisions Under Unobserved Confounding. arXiv preprint arXiv:2003.05623. 2020 Mar 12.*

- Chapter 4 (Identification of Subgroups With Similar Benefits in Off-Policy Policy Evaluation): We describe heterogeneous treatment effects in off-policy policy evaluation in sequential settings. We demonstrate the limitation of the existing method for non-sequential settings when applied to the sequential setting. We leverage recursive partitioning to develop an algorithm to automatically discover subgroups with different treatment effects and estimate the difference between the behavior and the evaluation policy for these subgroups.

We hope that this dissertation provides a step toward more robust reinforcement learning algorithms that can unleash the potential of RL to high-stakes real-world applications such as health care and education.

Chapter 2

Sample Efficient Risk Sensitive Policy Learning

While maximizing expected return is the goal in most reinforcement learning approaches, risk-sensitive objectives such as conditional value at risk (CVaR) are more suitable for many high-stakes applications. However, relatively little is known about how to explore to quickly learn policies with good CVaR. In this section, we present the first algorithm for sample-efficient learning of CVaR-optimal policies in Markov decision processes based on the optimism in the face of uncertainty principle. This method relies on a novel optimistic version of the distributional Bellman operator that moves probability mass from the lower to the upper tail of the return distribution. We prove asymptotic convergence and optimism of this operator for the tabular policy evaluation case. We further demonstrate that our algorithm finds CVaR-optimal policies substantially faster than existing baselines in several simulated environments with discrete and continuous state spaces.

2.1 Introduction

A key goal in reinforcement learning (RL) is to quickly learn to make good decisions by interacting with an environment. In most cases, the quality of the decision policy is evaluated with respect to a risk-neutral objective, its expected (discounted) sum of rewards. However, in many interesting cases, it is important to consider the full distributions over the potential sum of rewards, and the desired objective may be a risk-sensitive measure of this distribution. For example, a patient undergoing surgery for a knee replacement will (hopefully) only experience that procedure once or twice, and may be interested in the distribution of potential results for a single procedure, rather than what may happen on average if he or she were to undertake that procedure hundreds of time. Finance

and (machine) control are other cases where interest in risk-sensitive outcomes is common. Risk-neutral objective fails to address phenomena like loss aversion, where the subjective impact of a loss is often greater than that of an equivalent gain [Tversky and Kahneman, 1992]. This has particular relevance in high-stakes settings. For example, in finance, people may prefer investment strategies that yield modest but stable returns over higher-yield strategies that have a chance of bankrupting them. Likewise, patients undergoing surgery might opt for a procedure with a longer recovery time if it means minimizing the already small chance of a serious complication.

A popular risk-sensitive measure of a distribution of outcomes is the Conditional Value at Risk (CVaR) [Artzner et al., 1999]. Intuitively, CVaR is the expected reward in the worst α-fraction of outcomes, and has seen extensive use in financial portfolio optimization [Zhu and Fukushima, 2009], often under the name *expected shortfall*. While there has been recent interest in the RL community in learning to converge or identify good CVaR decision policies in Markov decision processes [Chow and Ghavamzadeh, 2014, Chow et al., 2015, Tamar et al., 2015b, Dabney et al., 2018], interestingly we are unaware of prior work focused on how to quickly learn such CVaR MDP policies, even though sample efficient RL for maximizing expected outcomes is a deep and well-studied theoretical [Jaksch et al., 2010, Dann et al., 2018] and empirical [Bellemare et al., 2016] topic. Sample efficient exploration seems of equal or even more importance in the case when the goal is risk-averse outcomes.

In this section, we work towards sample efficient reinforcement learning algorithms that can quickly identify a policy with an optimal CVaR. Our focus is on minimizing the amount of experience needed to find such a policy, similar in spirit to probably approximately correct RL methods for expected reward. Note that this is different than another important topic in risk-sensitive RL, which focuses on safe exploration: algorithms that focus on avoiding any potentially very poor outcomes during learning. These typically rely on local smoothness assumptions and do not typically focus on sample efficiency [Berkenkamp et al., 2017, Koller et al., 2018]; an interesting question for future work is whether one can do both safe and efficient learning of a CVaR policy. Our work is suitable for the many settings where some outcomes are undesirable but not catastrophic.

Our approach is inspired by the popular and effective principle of optimism in the face of uncertainty (OFU) in sample efficient RL for maximizing expected outcomes [Strehl and Littman, 2008, Brafman and Tennenholtz, 2002]. Such work typically works by considering uncertainty over the MDP model parameters or state-action value function and constructing an optimistic value function given that uncertainty is then used to guide decision making. To take a similar idea for rapidly learning the optimal CVaR policy, we seek to consider the uncertainty in the distribution of outcomes possible and the resulting CVaR value. To do so, we use the Dvoretzky-Kiefer-Wolfowitz (DKW) inequality. While to our knowledge this has not been previously used in reinforcement learning settings, it is a very useful concentration inequality for our purposes as it provides bounds on the true cumulative distribution function (CDF) given a set of sampled outcomes. We leverage these bounds to compute optimistic estimates of the optimal CVaR.

Our interest is in creating empirically efficient and scalable algorithms that have a theoretically sound grounding. To that end, we introduce a new algorithm for quickly learning a CVaR policy in MDPs and show that at least in the evaluation case in tabular MDPs, this algorithm indeed produces optimistic estimates of the CVaR. We also show that it does converge eventually. We accompany the theoretical evidence with an empirical evaluation. We provide encouraging empirical results on a machine replacement task [Delage and Mannor, 2010], a classic MDP where risk sensitive policies are critical, as well as a well validated simulator for type 1 diabetes [Man et al., 2014] and a simulated treatment optimization task for HIV [Ernst et al., 2006]. In all cases we find a substantial benefit over simpler exploration strategies. To our knowledge this is the first algorithm that performs strategic exploration to learn good CVaR MDP policies.

2.2 Background and Notation

Let X be a bounded random variable with cumulative distribution function $F(x) = \mathbb{P}[X \leq x]$. The *conditional value at risk (CVaR)* at level $\alpha \in (0, 1)$ of a random variable X is then defined as [Rockafellar et al., 2000]:

$$\text{CVaR}_\alpha(X) := \sup_\nu \left\{ \nu - \frac{1}{\alpha}\mathbb{E}[(\nu - X)^+] \right\} \tag{2.1}$$

We define the inverse CDF as $F^{-1}(u) = \inf\{x : F(x) \geq u\}$. It is well known that when X has a continuous distribution [Acerbi and Tasche, 2002].

$$\text{CVaR}_\alpha(X) = \mathbb{E}_{X\sim F}\left[X | X \leq F^{-1}(\alpha)\right]$$

For ease of notation we sometimes write CVaR as a function of the CDF F, $\text{CVaR}_\alpha(F)$.

We are interested in the CVaR of the discounted cumulative reward in a **Markov Decision Process (MDP)**. An MDP is defined by a tuple $(\mathcal{S}, \mathcal{A}, R, P, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are finite state and action space, $r \sim R(s, a)$ is the reward distribution, $s' \sim P(s, a)$ is the transition kernel and $\gamma \in [0, 1)$ is the discount factor. A stationary policy π maps each state $s \in \mathcal{S}$ to a probability distribution over action space $\mathcal{A}$.

Let $\mathcal{Z}$ denote the space of distributions over returns (discounted cumulative rewards) from such an MDP, and assume that these returns are in $[V_{\min}, V_{\max}]$ almost surely, where $V_{\min} \geq 0$. We define $Z_\pi(s, a) \in \mathcal{Z}$ to be the distribution of the return of policy π with CDF $F_{Z_\pi(s,a)}$ and initial state action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$ as

$$Z_\pi(s, a) := \text{Law}_\pi\left(\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s, A_0 = a\right)$$

RL algorithms most commonly optimize policies for expected return and explicitly learn Q-values,

$Q^\pi(s,a) = \mathbb{E}[Z_\pi(s,a)]$ by applying approximate versions of Bellman backups. Instead, we are interested in other properties of the return distribution and we will build on several recently proposed algorithms that aim to learn a parametric model of the entire return distribution instead of only its expectation. Such approaches are known as *distributional RL methods*.

Distributional Reinforcement Learning Distributional RL methods apply a sample-based approximation to distributional versions of the usual Bellman operators. For example, one can define a distributional Bellman operator [Bellemare et al., 2017] as $\mathcal{T}^\pi : \mathcal{Z} \to \mathcal{Z}$ as

$$\mathcal{T}^\pi Z_\pi(s,a) :\overset{D}{=} R(s,a) + \gamma P^\pi Z(s,a) \tag{2.2}$$

where $\overset{D}{=}$ denotes equality in distribution, and the transition operator is defined as $P^\pi Z(s,a) :\overset{D}{=} Z(s',a')$ with $s' \sim P(\cdot|s,a)$, $a' \sim \pi(s)$. The optimality version $\mathcal{T}$ is similarly any $\mathcal{T}Z = \mathcal{T}^\pi Z$ where π is an optimal policy w.r.t. expected return. Note that this is not necessarily unique when there are multiple optimal policies. Rowland et al. [2018] showed that $\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction in the Cramér-metric, $\bar{\ell}_2$

$$\bar{\ell}_2(Z_1, Z_2) = \sup_{s,a} \ell_2\left(Z_1(s,a), Z_2(s,a)\right) = \sup_{s,a} \left(\int (F_{Z_1(s,a)}(u) - F_{Z_2(s,a)}(u))^2 du \right)^{1/2} \tag{2.3}$$

One of the canonical algorithms in distributional RL is CDRL or C51 [Bellemare et al., 2017] which represent the return distribution Z^π as a discrete distribution with fixed support on N atoms $\{z_i = V_{\min} + i\Delta z : 0 \leq i < N\}, \Delta z := \frac{V_{\max} - V_{\min}}{N-1}$ the discrete distribution is parameterized as $\theta : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^N$:

$$Z_\theta(s,a) = z_i \quad \text{w.p.} \quad p_i(s,a) = \frac{e^{\theta_i(s,a)}}{\sum_j e^{\theta_j(s,a)}}$$

Essentially, C51 uses a sample transition (s,a,r,s') to perform an approximate Bellman backup $Z \leftarrow \Pi_{\mathcal{C}} \hat{\mathcal{T}} Z$, where $\hat{\mathcal{T}}$ is a sample-based Bellman operator and $\Pi_{\mathcal{C}}$ is a projection back onto the support of discrete distribution $\{z_0, \ldots, z_{N-1}\}$.

2.3 Optimistic Distributional Operator

In contrast to the typical RL setup where an agent tries to maximize its expected return, we seek to learn a stationary policy that maximizes the CVaR_α of the return at risk level α. Note that the CVaR-optimal policy at any state can be non-stationary [Shapiro et al., 2009], as it depends on the sum of rewards achieved up to that state. For simplicity, as Dabney et al. [2018] we instead seek a stationary policy, which generally can be suboptimal but typically still achieve high CVaR,

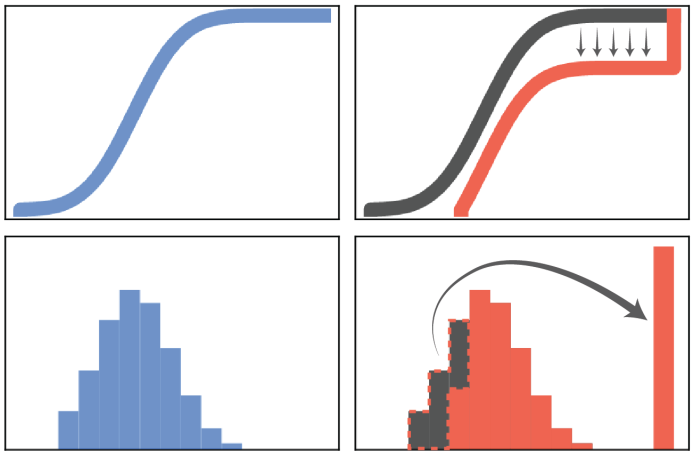


Figure 2.1: Top-left: Empirical CDF Top-right: The lower DKW confidence band (a shifted-down version of the empirical CDF). Bottom-left: Empirical PDF. Bottom-right: Optimistic PDF.

as observed in our experiments. To find such policies quickly, we follow the optimism in the face of uncertainty (OFU) principle and introduce optimism in our CVaR estimates to guide exploration.

While adding a bonus to rewards is a popular approach for optimism in the standard expected return case [Ostrovski et al., 2017], we here follow a different approach and introduce optimism into our return estimates by shifting the empirical CDFs. Formally, consider a return distribution $Z(s,a) \in \mathcal{Z}$ with CDF $F_{Z(s,a)}(x)$. We define the optimism operator $O_c : \mathcal{Z} \to \mathcal{Z}$ as

$$F_{O_c Z(s,a)}(x) = \left(F_{Z(s,a)}(x) - c \frac{\mathbf{1}\{x \in [V_{\min}, V_{\max})\}}{\sqrt{n(s,a)}} \right)^+ \tag{2.4}$$

where c is a constant and $(\cdot)^+$ is short for $\max\{\cdot, 0\}$. In the definition above, $n(s,a)$ is the number of times the pair (s,a) has been observed so far or an approximation such as pseudo-counts [Bellemare et al., 2016]. By shifting the cumulative distribution function down, this operator essentially puts probability mass from the lower tail to the highest possible value $V_{\max}$. An illustration is provided in Figure 2.1. This approach to optimism is motivated by an application of the DKW-inequality to the empirical CDF. As shown recently by Thomas and Learned-Miller [2019], this can yield tighter upper confidence bounds on the CVaR. Our approach also has the advantage that the scaling c does not depend on the risk level α, unlike a potential reward bonus approach that would likely require more parameter tuning.

2.4 Theoretical Analysis

The optimistic operator introduced above operates on the entire return distribution and our algorithm introduced in the next section combines this optimistic operator to estimated return-to-go distributions. As such, it belongs to the family of distributional RL methods [Dabney et al., 2018]. These methods are a recent development and come with strong asymptotic convergence guarantees when used for *policy evaluation* in tabular MDPs [Rowland et al., 2018]. Yet, finite sample guarantees such as regret or PAC bounds still remain elusive for distributional RL *policy optimization* algorithms.

A key technical challenge in proving performing bounds for distributionally robust policy optimization during RL is that convergence of the distributional Bellman optimality operator can generally not be guaranteed. Prior results have only showed that if the optimization process itself is to compute a policy which maximizes expected returns, such as Q-learning, then convergence of the distirbutional Bellman optimality operator is guaranteed to converge [Rowland et al., 2018, Theorem 2]. Note however that if the goal is to leverage distributional information to compute a policy to maximize something other than expected outcomes, such as a risk sensitive policy like we consider here, no prior theoretical results are known in the reinforcement learning setting to our knowledge. However, it is promising that there is some empirical evidence that one can compute risk-sensitive policies using distributional Bellman operators [Dabney et al., 2018] which suggests that more theoretical results may be possible.

Here we take a first step towards this goal. Fortunately, convergence issues of the distributional Bellman optimality have not been observed empirically which suggests that better theoretical characterizations could enable finite-sample guarantees eventually. Developing such characterizations is an important problem but not the focus of this chapter. Our primary aim in this work is to provide tools to introduce optimism into distributional return-to-go estimates to guide sample-efficient exploration for CVaR. Therefore, our theoretical analysis focuses on showing that this form of optimism does not harm convergence and is indeed a principled way to obtain optimistic CVaR estimates.

First, we prove that the optimism operator is a non-expansion in the Cramér distance. This results shows that this operator can be used with other contraction operators without negatively impacting the convergence behaviour. Specifically we can guarantee convergence with distributional Bellman backup.

Proposition 1. *For any c, the O_c operator is a non-expansion in the Cramér distance $\bar{\ell}_2$. This implies that optimistic distributional Bellman backups $O_c\mathcal{T}^\pi$ and the projected version $\Pi_{\mathcal{C}} O_c \mathcal{T}^\pi$ are $\sqrt{\gamma}$-contractions in $\bar{\ell}_2$ and iterates of these operators converge in $\bar{\ell}_2$ to a unique fixed-point.*

Proof. Consider $Z, Z' \in \mathcal{Z}$, any state $s \in \mathcal{S}$ and action $a \in \mathcal{A}$ with CDFs $F_{Z(s,a)}$ and $F_{Z'(s,a)}$ and

consider the application of the optimism operator O_c:

$$\int (F_{O_c Z(s,a)}(x) - F_{O_c Z'(s,a)}(x))^2 dx = \int_{V_{\min}}^{V_{\max}} ([F_{Z(s,a)}(x) - c]^+ - [F_{Z'(s,a)}(x) - c]^+)^2 dx \tag{2.5}$$

Generally, for any $a \geq b$ we have

$$([a-c]^+ - [b-c]^+)^2 = \begin{cases} (a-b)^2 & \text{if } a, b \geq c \\ (a-c)^2 \leq (a-b)^2 & \text{if } a > c \geq b \\ 0 & \text{if } c \geq a, b \end{cases} \tag{2.6}$$

and applying this case-by-case bound to the quantity in the integral above, we get

$$\int (F_{O_c Z(s,a)}(x) - F_{O_c Z'(s,a)}(x))^2 dx \leq \int_{V_{\min}}^{V_{\max}} (F_{Z(s,a)}(x) - F_{Z'(s,a)}(x))^2 dx \tag{2.7}$$

By taking the square root on both sides as well as a max over states and actions, we get that O_c is a non-expansion in $\bar{\ell}_2$. The rest of the statement follows from the fact that $\mathcal{T}^\pi$ is a $\sqrt{\gamma}$-contraction and $\Pi_{\mathcal{C}}$ a non-expansion [Rowland et al., 2018] and the Banach fixed-point theorem. □

This result does not include Bellman backups with the Bellman-optimality operator $\mathcal{T}$ or its generalization $\mathcal{T}_\alpha$ to CVaR-greedy policies. Here, $\mathcal{T}_\alpha$ is any operator that satisfies $\mathcal{T}_\alpha Z = \mathcal{T}^\pi Z$ for any policy π that is greedy w.r.t. CVaR at level α, i.e., $\pi(s) \in \operatorname{argmax}_a \operatorname{CVaR}_\alpha(Z(s,a))$. Note that $\mathcal{T}_\alpha = \mathcal{T}$ for $\alpha = 1$. The lack of general convergence guarantees in the control case is because optimality operators are generally not contractions [Bellemare et al., 2017] and not a limitation of our optimism operator and we have not observed any convergence issues in our experiments.

Next, we provide theoretical evidence that this operator indeed produces optimistic CVaR estimates. Consider here batch policy evaluation in MDPs M with finite state- and action-spaces. Assume that we have collected a fixed number of samples $n(s,a)$ (which can vary across states and actions) and build an empirical model $\hat{M}$ of the MDP. For any policy π, let $\hat{\mathcal{T}}^\pi$ denote the distributional Bellman operator in this empirical MDP. Then we indeed achieve optimistic estimates by the following result:

Theorem 1. *Let the shift parameter in the optimistic operator be sufficiently large which is* $c = O\left(\ln(|\mathcal{S}||\mathcal{A}|/\delta)\right)$. *Then with probability at least* $1-\delta$, *the iterates* $\operatorname{CVaR}_\alpha((O_c\hat{\mathcal{T}}^\pi)^m Z_0)$ *converges for any risk level* α *and initial* $Z_0 \in \mathcal{Z}$ *to an optimistic estimate of the policy's conditional value at risk. That is, with probability at least* $1-\delta$,

$$\forall s, a : \operatorname{CVaR}_\alpha((O_c\hat{\mathcal{T}}^\pi)^\infty Z_0(s,a)) \geq \operatorname{CVaR}_\alpha(Z_\pi(s,a)).$$

Proof of this theorem is presented in Appendix A.2. This theorem uses the DKW inequality which

to the best of our knowledge has not been used for MDPs. Note, that the statement guarantees optimism for all risk levels $\alpha \in [0, 1]$ without paying a penalty for it. Since we estimate the transitions and rewards for each state and action separately, one generally does not expect to be able to use a shift parameter smaller than $\Omega(\ln(|\mathcal{S}||\mathcal{A}|/\delta))$. Thus, Theorem 1 is unimprovable in that sense. Specifically, we avoid a polynomial dependency on the number of states $|\mathcal{S}|$ in the shift parameter c by combining two techniques: (1) concentration inequalities w.r.t. the optimal CVaR of the next state for a certain finite set of alphas and (2) a covering argument to get optimism for all infinitely many $\alpha \in [0, 1]$. This is substantially more involved than the expected reward case.

These results are a key step towards finite-sample analyses. In future work it would be very interesting to obtain a convergence analysis for distributional Bellman optimality operators in general, though this is outside the scope of this current paper. Such a result could lead to sample-complexity guarantees when combined with our existing analysis.

2.5 Algorithm

In the policy evaluation case where we would like to compute optimistic estimates of the CVaR of a given observed policy π, our algorithm essentially performs an approximate version of the optimistic Bellman update $O_c\mathcal{T}^\pi$ where $\mathcal{T}^\pi$ is the distributional Bellman operator. from Section 2.2, this operator can be composed by distributional Bellman operator $\mathcal{T}^\pi$ which results in optimistic Bellman evaluation update . We further analyze the theoretical properties of this operator and show that an optimistic estimate of CVaR can be obtained with empirical transition kernel $\hat{P}$ and reward distribution $F_{\hat{R}}$.

For the control case where we would like to learn a policy that maximizes CVaR, we instead define a distributional Bellman optimality operator $\mathcal{T}_\alpha$. Analogous to prior work [Bellemare et al., 2017], $\mathcal{T}_\alpha$ is any operator that satisfies $\mathcal{T}_\alpha Z = \mathcal{T}^\pi Z$ for some policy π that is greedy w.r.t. CVaR at level α. Our algorithm then performs an approximate version of the optimistic Bellman backup $O_c\mathcal{T}_\alpha$, shown in Algorithm 1.

The main structure of our algorithm resembles categorical distributional reinforcement learning (C51) [Bellemare et al., 2017]. In a similar vein, our algorithm also maintains a return distribution estimate for each state-action pair, represented as a set of N weights $p_i(s, a)$ for $i \in [N]$. These weights represent a discrete distribution with outcomes at N equally spaced locations $z_0 < z_1 < \cdots < z_{N-1}$, each $\Delta z = \frac{V_{\max} - V_{\min}}{N-1}$ apart. The current probability assigned to outcome z_i in (s, a) is denoted by $p_i(s, a)$, where the atom probabilities $p_{1:N}(s, a)$ are given by a differentiable model such as a neural network, similar to C51. Note that other parameterized representations of the weights are straightforward to incorporate [Bellemare et al., 2017].

The main differences between Algorithm 1 and existing distributional RL algorithms (e.g. C51) are highlighted in red. We first apply an optimism operator to our successor distribution $F_{Z(s_{t+1},a)}$

Algorithm 1: CVaR-MDP

Input: Parameters: γ, risk level $\alpha \in [0,1]$, $c \geq 0$, density model ρ,

1 **for** *t=1, ...* **do**

2 Observe transition s_t, a_t, r_t, s_{t+1};

3 **for** $a' \in \mathcal{A}$ **do**

/* emp. CDF of return for (s_{t+1}, a') */

4 $\hat{F}^{a'}(x) := \sum_{j=0}^{N-1} p_j(s_{t+1}, a')\mathbf{1}\{x \geq z_j\}$;

/* Pseudo-counts using density model */

5 $\hat{n} = \frac{1}{\exp(\kappa t^{-1/2}\alpha(\nabla \log \rho_\theta(s_{t+1},a'))^2)-1}$ /* Optimistic CDF */

6 $\tilde{F}^{a'}(x) := \left[\hat{F}^{a'}(x) - \frac{c\mathbf{1}\{x \in [V_{\min}, V_{\max})\}}{\sqrt{\hat{n}}}\right]^+$;

7 **if** *Control* **then**

$a^\star \leftarrow \text{argmax}_{a \in \mathcal{A}} \text{CVaR}_\alpha(\tilde{F}^a)$

if *Evaluation* **then**

$a^\star \sim \pi(.|s_{t+1})$

8 $m_i = 0$ for $i \in \{0, \ldots, N-1\}$;

9 **for** $j \in 0, \ldots, N-1$ **do**

/* optimistic PDF from opt. CDF */

10 $\tilde{p}_j \leftarrow \tilde{F}^{a^\star}\left(z_j + \frac{\Delta z}{2}\right) - \tilde{F}^{a^\star}\left(z_j - \frac{\Delta z}{2}\right)$;

/* Project on support of $\{z_i\}$ */

11 $\tilde{\mathcal{T}} z_j \leftarrow [r_t + \gamma z_j]_{V_{\min}}^{V_{\max}}$;

/* Distribute prob. of $\tilde{\mathcal{T}}(z_j)$ */

12 $b_j \leftarrow (\tilde{\mathcal{T}} z_j - V_{\min})/(\Delta z)$;

13 $l \leftarrow \lfloor b_j \rfloor$; $u \leftarrow \lceil b_j \rceil$;

14 $m_l \leftarrow m_l + \tilde{p}_j(u - b_j)$;

15 $m_u \leftarrow m_u + \tilde{p}_j(b_j - l)$;

16 Update return weights $p_{1:N}$ by optimization step on cross-entropy loss $-\sum_{j=0}^{N-1} m_j \log p_j(s_t, a_t)$;

/* Take next action */

17 $a_{t+1} \leftarrow a^\star$;

18 Update density model for ρ with additional observation of (s_{t+1}, a_{t+1});

(Lines 4–6) to form an optimistic CDF $\tilde{F}_{Z(s_{t+1},a)}$ for all actions $a \in \mathcal{A}$. This operator should encourage exploring actions that might lead to higher CVaR policies for our input α. These optimistic CDFs are also used to decide on the successor action in the control setting (Line 7). Then, similar to C51 we apply the Bellman operator $\hat{\mathcal{T}} z_i$ for $i \in [N]$ and distribute the probability of $\tilde{p}_i$ to the immediate neighbours of $\hat{\mathcal{T}} z_i$, where we calculate the probability mass $\tilde{p}_i$ with the optimistic CDF $\tilde{F}_{Z(s_{t+1},a^\star)}$ (Line 10).

Following Bellemare et al. [2017], we train this model using the cross-entropy loss, which for a particular state transition at time t is

$$-\sum_{j=0}^{N-1} m_j \log p_j(s_t, a_t) \tag{2.8}$$

where $m_{0:N-1}$ are the weights of the target distribution computed in Lines 8–15 in Algorithm 1. In the tabular setting we can directly update the probability mass p_j by

$$p_j(s_t, a_t) = (1 - \beta)p_j(s_t, a_t) + \beta m_j(s_t, a_t)$$

where β is the learning rate.

In tabular settings, the counts $n(s, a)$ can be directly stored and used; however, this is not the case in continuous settings. For this reason, we adopt the pseudo-count estimation method proposed by Ostrovski et al. [2017] and replace $n(s, a)$ by a pseudo-count $\hat{N}_t(s, a)$ in the optimistic distributional operator (Equation 2.4). Let ρ be a density model and $\rho_t(s, a)$ the probability assigned to the state action pair (s, a) by the model after t training steps. The prediction gain PG of ρ is defined

$$PG_t(s, a) = \log \rho'_t(s, a) - \log \rho_t(s, a) \tag{2.9}$$

Where $\rho'_t(s, a)$ is the probability assigned to (s, a) if it were trained on that same (s, a) one more time. Now we define the pseudo count of (s, a) as

$$\hat{N}_t(s, a) = (\exp(\kappa t^{-\frac{1}{2}}(PG(s, a))_+ - 1)^{-1} \tag{2.10}$$

where κ is a constant hyper-parameter, and $(PG(s, a))_+$ thresholds the value of the prediction gain at 0.

Our setting differs from Ostrovski et al. [2017] in the sense that we have to compute the count before taking the action a. A naive way would be to try all actions and train the model to compute the counts but this method is slow and requires the environment to support an undo action. Instead, we can estimate PG for all actions as follows. Consider the density model parametrized by θ, $\rho(s, a; \theta)$. After observing (s, a), the training step to maximize the log likelihood will update the parameters by $\theta' = \theta + \alpha \nabla_\theta \log \rho(s, a; \theta)$, where α is the learning rate. So we can approximate the new log probability using a first-order Taylor expansion

$$\begin{aligned}
\log \rho'_t(s, a) &= \log \rho(s, a; \theta') \\
&\approx \log \rho(s, a; \theta) + \nabla_\theta \log \rho(s, a; \theta)(\theta' - \theta) \\
&= \log \rho(s, a; \theta) + \alpha(\nabla_\theta \log \rho(s, a; \theta))^2.
\end{aligned}$$

This calculation suggests that the prediction gain can be estimated just by computing the gradient of the log likelihood given a state-action pair, i.e., $PG(s, a) \approx \alpha(\nabla_\theta \log \rho(s, a; \theta))^2$. As discussed in Graves et al. [2017] this estimate of prediction gain is biased, but empirically we have found this method to perform well.

2.6 Experimental Evaluation

We validate our algorithm empirically in three simulated environments against baseline approaches. Finance, health and operations are common areas where risk-sensitive strategies are important, and we focus on two health domains and one operations domain.

The majority of prior risk-sensitive RL work has not focused on efficient exploration, and there has been very little deep distributional RL work focused on risk sensitivity. Our key contribution is to evaluate the impact of more strategic exploration on the efficiency with which a risk-sensitive policy can be learned. We compare to following approaches:

1. ϵ-greedy CVaR: In this benchmark we use the same algorithm, except we do not introduce an optimism operator, instead using an ϵ-greedy approach for exploration. This benchmark can be viewed as analogous to the distributional RL methods of C51 [Bellemare et al., 2017] if the computed policy had optimized for CVaR instead of expected reward.

2. IQN-ϵ-greedy CVaR: In this benchmark we use implicit quantile network (IQN) that also uses ϵ-greedy method for exploration [Dabney et al., 2018]. We adopted the dopamine implementation of IQN [Castro et al., 2018].

3. CVaR-AC: An actor-critic method proposed by [Chow and Ghavamzadeh, 2014] that maximizes the expected return while satisfying an inequality constraint on the CVaR. This method relies on the stochasticity of the policy for exploration.

Note that a comparison to an expectation maximizing algorithm is uninformative since such approaches are maximizing different (non-risk-sensitive) objectives.

All of these algorithms use hyperparameters, and it is well recognized that ϵ-greedy algorithms can often perform quite well if their hyperparameters are well-tuned. To provide a fair comparison, we evaluated across a number of schedules for reducing the ϵ parameter for both ϵ-greedy and IQN, and a small set of parameters (4-7) for the optimism value c for our method. We used the specification described in [Chow and Ghavamzadeh, 2014] for CVaR-AC.

All results are averaged over 10 runs and we report 95% confidence intervals. We report the performance of ϵ-greedy at evaluation time (setting $\epsilon = 0$), which is the best performance of ϵ-greedy.

Machine Replacement Machine repair and replacement is a classic example in the risk sensitive literature, though to our knowledge no prior work has considered how to quickly learn a good risk-sensitive policy for such domains. Here we consider a minor variant of a prior setting [Delage and Mannor, 2010].

Specifically, as shown in Figure 2.2, the environment consists of a chain of n (25 in our experiments) states. There are two actions: *replace* and *don't replace* the machine. Choosing *replace* at

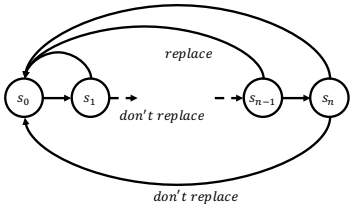


Figure 2.2: Machine Replacement: This environment consists of a chain of n states, each affording two actions: *replace* and *don't replace*.

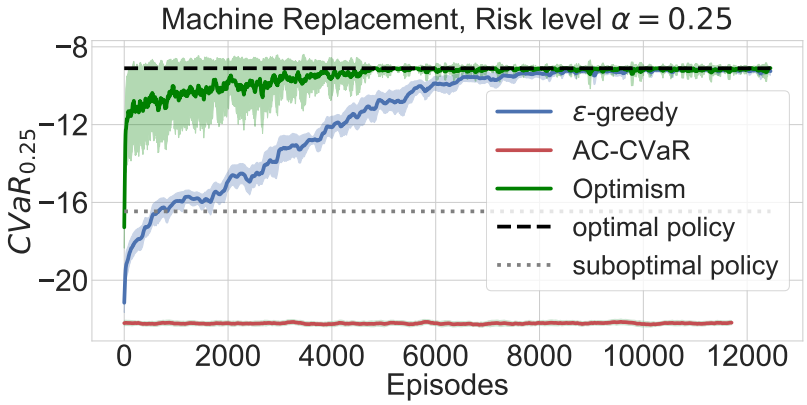


Figure 2.3: Machine Replacement: The thick grey dashed line is the $\text{CVaR}_{0.25}$-optimal policy. The thin dashed lines labeled as the suboptimal policy is the optimal expectation-maximizing policy. The shaded area shows the 95% confidence intervals.

any state terminates the episode, while choosing *don't replace* moves the agent to the next state in the chain. At the end of the chain, choosing *don't replace* terminates the episode with a high variance cost, and choosing *replace* terminates the episode with a higher cost but lower variance. This environment is especially a challenging exploration task due to the chain structure of the MDP, as well as the high variance of the reward distributions when taking actions in the last state. Additionally in this MDP it is feasible to exactly compute the $\text{CVaR}_{0.25}$-optimal policy, which allows us to compare the learned policy to the true optimal CVaR policy. Note here that the optimal policy for maximizing $\text{CVaR}_{0.25}$ is to *replace* on the final state in the chain to avoid the high variance alternative; in contrast, the optimal policy for expected return always chooses *don't replace*.

Specifically, the environment in our experiment consist of n states (we use $n = 25$ in the experiment), where action *replace* transitions to the terminal state with cost $\mathcal{N}(\mu_{r,t}, \sigma_{r,t})$ at state t, where $\mu_{r,t} = r_{max} - \frac{t}{n}(r_{max} - r_{min})$ and $\sigma_{r,t} = 0.1 + 0.01t$. Action *don't replace* has cost $\mathcal{N}(0, 1e-2)$ and transitions to state $t+1$. In our experiment we used $r_{max} = 23, r_{min} = 10$. However, for the last state n action *don't replace* has cost $\mathcal{N}(\mu_r, 10)$, where we used $\mu_r = 8$, and transitions to the terminal state. For C51 algorithm we use $V_{min} = -50, V_{max} = 50$, $\gamma = 0.99$, learning rate 0.01 and

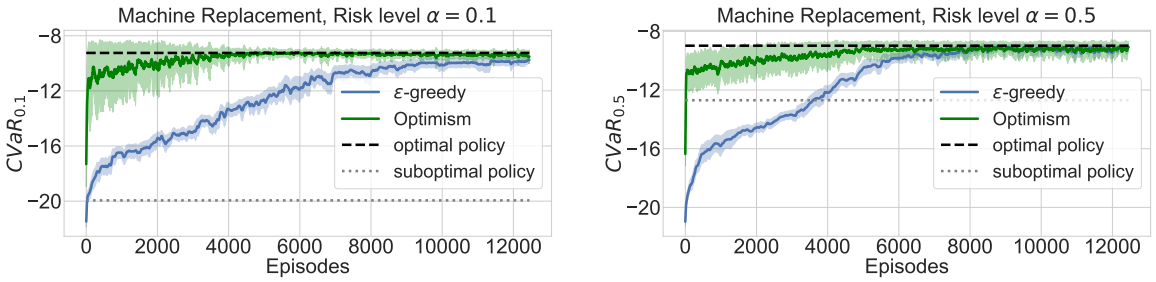


Figure 2.4: Machine Replacement with different risk levels. Left: risk level $\alpha = 0.1$, Right: risk level $\alpha = 0.5$

51 atoms.

Tuning: We use ϵ-greedy with schedule $(\text{start}, \text{end}, \text{n}) = (0.9, 0.1, 5000)$ that starts with $\epsilon = \text{start}$ and decays linearly to $\epsilon = \text{end}$ in n time steps, staying constant afterwards. This schedule achieved the best performance in our experiments when compared to other linear schedules $\{(0.9, 0.3, 5000)$, $(0.9, 0.1, 10000)$, $(0.9, 0.1, 15000)$, $(0.9, 0.05, 5000)\}$, and exponential decays with schedule in the form of $(\epsilon_0, d, step)$: $\{(0.9, 0.99, 5)$, $(0.9, 0.99, 20)$, $(0.9, 0.99, 2)$, $(0.9, 0.99, 30)$, $(0.5, 0.99, 5)\}$ where $\epsilon = \epsilon_0 \times d^{episode/step}$. We have also tried our algorithm with optimism values of $c = [0.25, 0.5, 1, 2]$. For the actor critic method we use the CVaR limit as -10, radial basis function as kernel and other set of hyper-parameters are as described in the appendix of Chow and Ghavamzadeh [2014].

Results: In Machine Replacement (Figure 2.3) we see that our method quickly converges to the optimal CVaR performance.

Unfortunately despite our best efforts, our implementation of CVaR-AC did not perform well even on the simplest environment, so we did not show the performance of this method on other environments. One challenge here is that CVaR-AC has a significant number of hyper-parameters, including 3 different learning rates schedule for the optimization process, initial Lagrange multipliers and the kernel functions.

Additional to the risk level $\alpha = 0.25$, we observe the same gain in the performance for other risk levels. As shown in figure 2.4, optimism based exploration shows a significant gain over $\epsilon-$greedy exploration for risk levels $\alpha = 0.1$ and $\alpha = 0.5$.

HIV Treatment In order to test our algorithm on a larger continuous state space, we leverage an HIV Treatment simulator. The environment is based on the implementation by [Geramifard et al., 2015] of the physical model described in [Ernst et al., 2006]. The patient state is represented as a 6-dimensional continuous vector and the reward is a function of number of free HIV viruses, immune response of the body to HIV, and side effects. There are four actions, each determining which drugs are administered for the next 20 day period: Reverse Transcriptase Inhibitors (RTI), Protease Inhibitors (PI), neither, or both. Typical treatments for HIV patients utilize cocktails consist of one

or more RTIs in combination with a PI and patients taking these drugs experience many common and sometimes highly undesirable side effects, often leading to poor compliance. So it is important to minimize the side effects.

There are 50 time steps in total per episode, for a total of 1000 days. We chose here a larger number of days per time step compared to the typical setup (200 steps of 5 days each) to facilitate faster experimentation. This design choice also makes the exploration task harder, since taking one wrong action can drastically destabilize a patient's trajectory. The original proposed model was deterministic, which makes the CVaR policy identical to the policy optimizing the expected value. Such simulators are rarely a perfect proxy for real systems, and in our setting we add Gaussian noise $\sim \mathcal{N}(0, 0.01)$ to the efficacy of each drug (RTI: ϵ_1 and PI: ϵ_2 in Ernst et al. [2006]). This change necessitates risk-sensitive policies in this environment.

The environment is an implementation of the physical model described in [Ernst et al., 2006]. The state space is of dimension 6 with and action space is of size 4, indicating the efficacy of being on the treatment. ϵ_1, ϵ_2 described in [Ernst et al., 2006] takes values as $\epsilon_1 \in \{0, 0.7\}$ and $\epsilon_2 \in \{0, 0.3\}$. We have also added the stochasticity to the action by a random gaussian noise. So the efficacy of a drug is computed as $\epsilon_i + \mathcal{N}(0, 0.01)$.

The reward structure is defined similar to the prior work [Ernst et al., 2006]. And we simulate for 1000 time steps, where agent can take action in 50 steps (each 20 simulation step) and actions remains constant in each interval. While trianing we normalize the reward by dividing them by $1e6$.

Categorical Distributional RL: The C51 model consist of 4 hidden layers each of size 128 and ReLU activation function, followed by of $|\mathcal{A}|$ each with 151 neurons followed by a softmax activation, for representing the distribution of each action. We used Adam optimizer with learning rate decay schedule from $1e-3$ to $1e-4$ in half a number of episodes, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e-8$. We set $V_{max} = 40$, $V_{min} = -10$, 151 probability atoms, and used batch size of 32. For computing the CVaR we use 50 samples of the return.

Implicit Quantile Network: IQN model consists of 4 hidden layers with size 128 and ReLU activation. Then an embedding of size 64 computed by $\text{ReLU}(\sum_{i=0}^{n-1} \cos(\pi i \tau) w_{ij} + b_j)$. Then we take the element wise multiplication of the embedding and the output of 4 hidden layers, followed by a fully connected layer with size 128 and ReLU activation, and a softmax layer. We used 8 samples for N and N' and 32 quantiles.

Density Model: For log likelihood density model we used realNVP [Dinh et al., 2016] with 3 layers each of size 64. The input of the model is a concatenated vector of (s, a). We used same hyper parameters for optimizer as in C51 model. We have used constant $\kappa = 1e-5$ for computing the pseudo-count.

Tuning: We have tuned our method, ϵ-greedy and IQN. For ϵ-greedy we tried 5 different linear schedule of ϵ-greedy, {(0.9, 0.05, 10), (0.9, 0.05, 8), (0.9, 0.05, 5), (0.9, 0.05, 4), (0.9, 0.05, 2)} where first element is the initial ϵ, second element is the final ϵ and the third element is episode ratio (i.e.

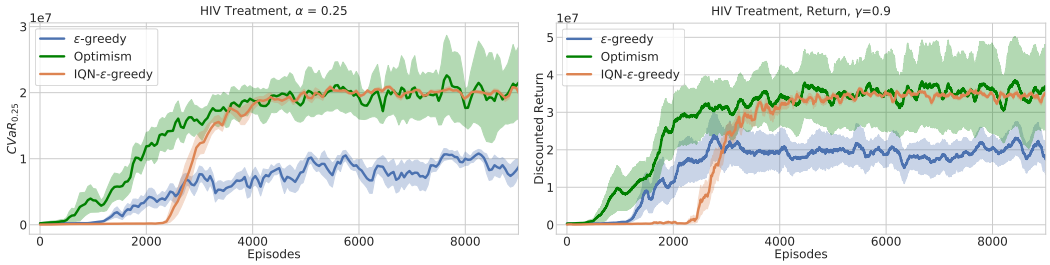


Figure 2.5: Comparison of our approach against an ϵ-greedy and IQN baseline. All models were trained to optimize the $\text{CVaR}_{0.25}$ of the return on a stochastic version of the HIV simulator [Ernst et al., 2006]. Top: Objective $\text{CVaR}_{0.25}$; Bottom: Discounted expected return of the same policies as in top plot.

epsilon starts from initial and reaches to the final value in episode ratio fraction of total number of episodes, linearly). Additionally we have tried 5 different exponential decay schedule for ϵ-greedy and IQN in the form of (ϵ_0, d, $step$): {(1.0, 0.9, 10), (1.0, 0.9, 100), (1.0, 0.9, 500), (1.0, 0.99, 10), (1, 0.99, 100)} where $\epsilon = \epsilon_0 \times d^{episode/step}$. The first of the linear decay set preformed the best. We have also tested our algorithm with constant optimism values of $(0.2, 0.4, 0.5, 0.8, 1, 2, 5)$ where we picked the best value 0.8.

Results: In the HIV Treatment we also see a clear and substantial benefit to our optimistic approach over the baseline ϵ-greedy approach and IQN(Figure 2.5).

Diabetes 1 Treatment Patients with type 1 diabetes regulate their blood glucose level with insulin in order to avoid hypoglycemia or hyperglycemia (very low or very high blood glucose level, respectively).

An open source implementation of type 1 diabetes simulator [Man et al., 2014] simulates 30 different virtual patients, 10 child, 10 adolescent and 10 adult. For our experiments in this paper we have used `adult#003`, `adult#004` and `adult#005`. Additionally we have used `"Dexcom"` sensor for CGM (to measure blood glucose level) and `"Insulet"` as a choice of insulin pump. All simulations are 10 hours for each patient and after 10 hour, patient resets to the initial state. Each step of simulation is 3 minutes.

State space is a continous vector of size 2 (glucose level, meal size) where glucose level is the amount of glucose measured by `"Dexcom"` sensor and meal size is the amount of Carbohydrate in each meal.

Action space is defined as (bolus, basal=0) where amount of bolus injection discretized by 6 bins between 30 (max bolus, a property of the `"Insulet"` insulin pump) and 0 (no injection). Additionally we inject two source of stochasticity to the taken action, assume action $a = (a_b, 0)$ at

time t is the agent's decision, then we take the action $a = (a'_b, 0)$ at time t' where:

$$a'_b = a_b + \mathcal{N}(0, 1)$$
$$t' = t + c - \lfloor x \times c \rfloor$$

Where $x \sim P(x; 1) = 2x^{-1}$ is drawn from the power law distribution and $c = 5$. Note that this means delay the action at most 5 step where the probability of taking the action at time t is higher than time $t + i, i \geq 1$ following the power law. Since each step of simulation is 3 minutes, patient might take the insulin up to 15 minutes after the prescribed time by the agent.

Reward structure is defined similar to the prior work [Bastani, 2014] as following:

$$r(bg) = \begin{cases} -\frac{(bg'-6)^2}{5} & if\ bg' < 6 \\ -\frac{(bg'-6)^2}{10} & if\ bg' \geq 6 \end{cases}$$

Where $bg' = bg/18.018018$ which is the estimate of bg (blood glucose) in mmol/L rather than mg/dL. Additionally if the amount of glucose is less than 39 mg/dL agent incurs a penalty of -10.

We generated a meal plan scenario for all the patients that is meal of size 60, 20, 60, 20 CHO with the schedule 1 hour, 3 hours, 5 hours and 7 hours after starting the simulation. Notice that this will make the simulation horizon 200 steps and 5 actionable steps (initial state, and after each meal).

Categorical Distributional RL: The C51 model consist of 2 hidden layers each of size 32 and ReLU activation function, followed by of $|\mathcal{A}|$ each with 51 neurons followed by a softmax activation, for representing the distribution of each action.

We used Adam optimizer with learning rate $1e-3$, $\beta_1 = 0.9$,$\beta_2 = 0.999$ and $\epsilon = 1e-8$. We set $V_{max} = 15$, $V_{min} = -40$, 51 probability atoms, and used batch size of 32. For computing the CVaR we use 50 samples of the return.

Density Model: For log likelihood density model we used realNVP [Dinh et al., 2016] with 3 layers each of size 64. The input of the model is a concatenated vector of (s, a). We used same hyper parameters for optimizer as in C51 model. We have used constant $\kappa = 1e-5$ for computing the pseudo-count.

Tuning: We have tuned our method and ϵ-greedy on patient `adult#001` and used the same parameters for the other patients. We tried 5 different linear schedule of ϵ-greedy, {(0.9, 0.1, 2), (0.9, 0.05, 4), (0.9, 0.05, 6), (0.9, 0.3, 4), (0.9, 0.3, 4), (0.9, 0.05, 10)} where first element is the initial ϵ, second element is the final ϵ and the third element is episode ratio (i.e. epsilon starts from initial and reaches to the final value in episode ratio fraction of total number of episodes, linearly). Additionally we have tried 5 different exponential decay schedule for ϵ-greedy in the form of (ϵ_0, d, *step*): {(0.9, 0.99, 5), (0.9, 0.99, 20), (0.9, 0.99, 2), (0.9, 0.99, 30), (0.5, 0.99, 5)} where $\epsilon = \epsilon_0 \times d^{episode/step}$. The first of the exponential decay set preformed the best. We have also tested

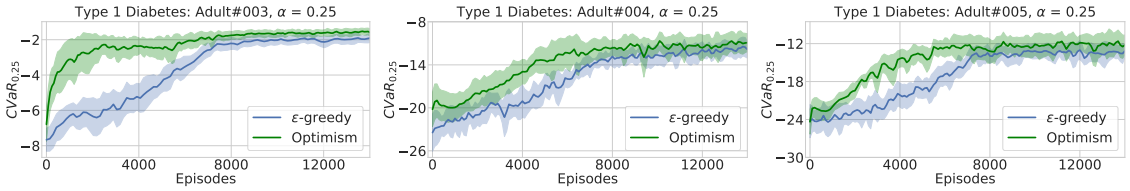


Figure 2.6: Type 1 diabetes simulator: $\text{CVaR}_{0.25}$ for three different adults. Plots are averaged over 10 runs with 95% CI.

our algorithm with constant optimism values of $[0.2, 0.4, 0.5, 0.8, 1, 2, 5]$ where we picked the best value 0.5.

This domain also readily offers a suite of related tasks, since the environment simulates 30 patients with slightly different dynamics. Tuning hyper-parameters on the same task can be misleading [Henderson et al., 2018], as is the case in our two previous benchmarks. In this setting we tune baselines and our method on one patient, and test the performance on different patients.

Figure 2.6 is particularly encouraging, as it shows the results for the diabetes simulator across 3 patients, where the hyperparameters were fixed after optimizing for a separate patient. Since in real settings it would be commonly necessary to fix the hyperparameters in advance, this result provides a nice demonstration that the optimistic approach can consistently equal or significantly improve over an ϵ-greedy policy in related settings, similar to the well known results in Atari in which hyperparameters are optimized for one game and then used for multiple others.

***Safer* Exploration.** Our primary contribution is a new algorithm to learn risk-sensitive policies quickly, with less data. However, an interesting side benefit of such a method might be that the number of extremely poor outcomes experienced over time may also be reduced, not due to explicitly prioritizing a form of safe exploration, but because our algorithm may enable a faster convergence to a safe policy.

To evaluate this, we consider a risk measure proposed by [Clarke and Kovatchev, 2009], which quantifies the risk of a severe medical condition based on how close their glucose level is to hypoglycemia (blood glucose, $\leq$3.9 mmol/l) and hyperglycemia (blood glucose, $\geq$10 mmol/l).

Table 2.1 shows the fraction of episodes in which each patient experienced a severely poor outcome for each algorithm while learning. Optimism-based exploration approximately halves the number of episodes with severely poor outcomes, highlighting a side benefit of our optimistic approach of more quickly learning a good safe policy.

2.7 CVaR Bandit

In a K-armed bandit problem, an agent samples from one of K distributions at each turn. Informally, the goal of the agent is to choose from "good" arms as often as possible. This requires the agent

	ϵ-greedy	CVaR-MDP
Adult#003	11.2% ± 3.6%	**4.2% ± 2.3%**
Adult#004	2.3% ± 0.3%	**1.4% ± 0.6%**
Adult#005	3.3% ± 0.3%	**1.7% ± 0.6%**

Table 2.1: Type 1 Diabetes simulator, percent of episodes where patients experienced a severe medical condition (hypoglycemia or hyperglycemia), averaged across 10 runs

to balance exploration of all possible arms with exploitation of the knowledge it has gained from previous samples. This theoretical framework has applications ranging from clinical trial design [Villar et al., 2015] to financial portfolio optimization [Shen et al., 2015] to optimizing websites [White, 2012].

The multi-armed bandits literature has traditionally used the expected value of an arm's reward distribution as the proxy for the *goodness* of that arm. However, in many interesting cases, it is important to consider the full distributions over the potential rewards, and the desired objective may be a risk-sensitive measure of this distribution. For example, a patient undergoing a surgery for a knee replacement will (hopefully) only experience that procedure once or twice, and may well be interested in the distribution of potential results for a single procedure, rather than what may happen on average if he or she were to undertake that procedure hundreds of time. Finance and (machine) control are other cases where interest in risk-sensitive outcomes are common.

Similar to previous sections, we aim to develop a sample efficient algorithm for CVaR-Bandits where we aim to find the CVaR optimal arm using efficient exploration. Existing OFU approaches for risk-sensitive bandits derive bonus terms as upper-confidence bounds on the CVaR of the rewards [Cassel et al., 2018], one might wonder whether tighter bonuses could be obtained by taking the shape of the arm distribution into account. This may be especially relevant for an objective like CVaR which is more sensitive to one side of the distribution than the other; for example, if the distribution of observed samples is skewed towards the minimum observed reward, one might be more confident in the corresponding CVaR estimate than if the samples were skewed in the other direction.

We present a method that applies optimism directly at the sample level, generating a new set of *optimistic samples* for each arm *before* computing the sample CVaR from them. This process enables our algorithm to take the shape of the distribution into account, as opposed to count-based reward bonuses added to the sample CVaR, which are agnostic to it. To construct these optimistic samples, we use the Dvoretzky–Kiefer–Wolfowitz (DKW) concentration inequality [Dvoretzky et al., 1956], which provides bounds on the true cumulative distribution function (CDF) given a set of sampled outcomes. These bounds take the form of confidence bands around the empirical distribution function (EDF), and we show how to modify our samples to match the optimistic band.

We will shortly illustrate how DKW allows us to both preserve strong theoretical guarantees and

enable better empirical performance by being more sensitive to the underlying CDF distribution, compared to bounds that are agnostic to this shape. Empirically, we observe that our DKW approach of bounding the CDF achieves an order of magnitude lower CVaR-regret than Cassel et al. [2018]'s approach which uses direct bonuses on the sample CVaR. These improvements on a variety of simulated environments showcase the importance of distributionally-aware exploration for quickly learning risk-sensitive policies.

2.7.1 Notation

We consider a stochastic K-armed bandit setting with rewards contained in $[0, U]$. $T_i(n)$ is the number of times arm i has been pulled up to round n; A_t is the action taken during round t; $[m]$ denotes the set $\{1, \dots m\}$ for any m; and P_i is the PDF of the distribution of rewards from the ith arm. Let $(X_{i,t})_{i\in[k],t\in[n]}$ denote a collection of independent random variables (the samples of our arms), with the pdf of X_{it} equal to P_i. $X_t = X_{A_t,T_{A_t}(t)}$ is the reward in round t. The empirical distribution function of $X_{i,s}$ is $\hat{F}_{i,t}(x) = \frac{1}{t}\sum_{s=1}^{t} \mathbb{I}\{X_{i,s} \leq x\}$.

Here, we define the *CVaR-regret* at time n as

$$R_n^\alpha = n \max_i \left(\text{CVaR}_\alpha(F_i)\right) - \mathbb{E}\left[\sum_{t=1}^{n} \text{CVaR}_\alpha(F_{A_t})\right] = \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}^\alpha\right] \tag{2.11}$$

where F_a is the CDF of the distribution of rewards from the ath arm, A_t is the action taken at time t and $\Delta_a^\alpha = \max_j(\text{CVaR}_\alpha(F_j)) - \text{CVaR}_\alpha(F_a)$ is the suboptimality of arm a with respect to CVaR. We also consider an alternate notion of regret for the CVaR setting from Cassel et al. [2018] in Section 2.7.3 and Appendix A.3.3.

2.7.2 Algorithm

We present our algorithm, CVaR-UCB, in Algorithm 2. CVaR-UCB computes an optimistic estimate of the CVaR of each arm from available samples, and then chooses the arm with the highest upper-confidence bound in each turn. This optimistic estimate is based on the concentration of the empirical cumulative distribution function (CDF) via the Dvoretzky-Kiefer-Wolfowitz [DKW; Dvoretzky et al., 1956] inequality. DKW provides a deviation bound on the maximum difference between the empirical CDF and the true CDF as a function of the number of samples observed. We can use the DKW bound to compute an optimistic estimate of the CVaR value of an arm by computing the CVaR of the optimistic confidence band around the CDF. This optimistic CVaR can be found very simply by shifting the lowest-reward samples to the maximum reward U and then taking the empirical CVaR of the resulting "optimistic samples".

We now show that this simple method has strong regret bounds. To our knowledge this is the first time that DKW-based optimism has been used in risk-sensitive RL or bandits. For proofs see

Algorithm 2: CVaR-UCB

Input: Risk level α, reward range U, horizon n

Choose each arm once;
Set $\hat{F}_a$ as the CDFs of each arm a on $[0, U]$ for all $a \in [K]$;
Set $T_a \leftarrow 1$;
for $t = 1, \ldots, n$ **do**
  for $a = 1, \ldots, K$ **do**
    $\epsilon_a \leftarrow \sqrt{\frac{\ln(2n^2)}{2T_a}}$;
    $\tilde{F}_a(x) \leftarrow \left(\hat{F}_a(x) - \epsilon_i \mathbf{1}\{x \in [0, U)\}\right)^+$; // optimistic empirical CDF
    $\mathrm{UCB}_a^{\mathrm{DKW}}(t) \leftarrow \mathrm{CVaR}_\alpha(\tilde{F}_a)$;
  Play action $A_t = \mathrm{argmax}_i \, \mathrm{UCB}_i^{\mathrm{DKW}}(t)$;
  $T_{A_t} \leftarrow T_{A_t} + 1$;
  Update empirical CDF $\hat{F}_{A_t}$ of arm A_t;

Appendix A.3.1.

Theorem 2. *Consider CVaR-UCB on a stochastic K-armed bandit problem with rewards bounded in* $[0, U]$*. For any given horizon* n *the expected CVaR-regret after this horizon is bounded as*

$$R_n^\alpha \leq \sum_{i \in [K]: \Delta_i^\alpha > 0} \frac{4U^2 \ln(\sqrt{2}n)}{\alpha^2 \Delta_i^\alpha} + 3 \sum_{i=1}^{K} \Delta_i^\alpha; \qquad R_n^\alpha \leq \frac{4U}{\alpha} \sqrt{nK \ln(\sqrt{2}n)} + 3KU \tag{2.12}$$

Note that the bounds differ on their dependence on the number of samples n and risk level α: the problem-dependent bound is $O\left({}^{U^2 \log n}/_{\alpha^2}\right)$, while the problem-independent bound grows as $O\left({}^{U\sqrt{n}}/_{\alpha}\right)$. Observe that for $\alpha = 1$, we recover (in dominant terms) the well known upper confidence bound regret results for comparing to the arm with the best expected reward [Bubeck et al., 2012].

2.7.3 Comparison with Direct Bonuses on the CVaR

In our proposed approach we compute an optimistic estimate of the CDF and then extract a CVaR from this optimistic estimate. In contrast, in standard bandit methods optimizing for expected outcomes, one simply adds a direct bonus to the mean to compute the upper-confidence bound. Therefore, a natural alternative to our proposal, as introduced by Cassel et al. [2018], is to directly compute the empirical CDF, extract the empirical CVaR and then add a bonus based on the number of samples. Procedurally this is equivalent to right-shifting each observed sample, in contrast to our algorithm in which we use DKW to compute a lower bound on the empirical CDF, effectively shifting probability mass from the lower-reward tail to the max reward.

Interestingly we will shortly observe that empirically there is a significant difference between these two styles of approaches. In particular, our approach of first computing a bound on the

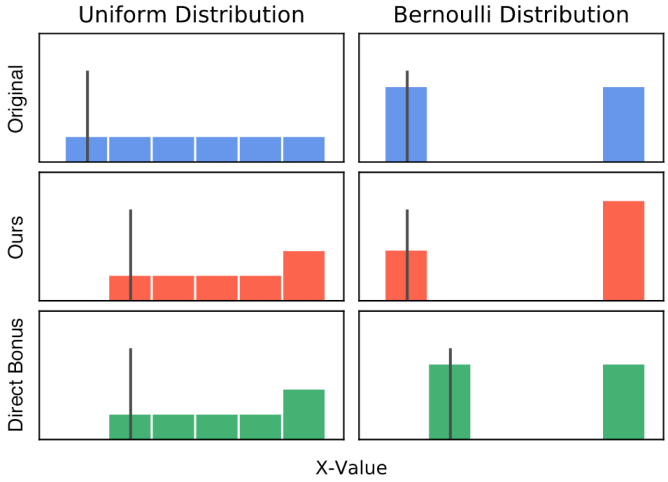


Figure 2.7: While our method shifts the lowest-reward samples to the maximum value, direct bonuses on the sample CVaR effectively shift all samples to the right equally. For the uniform distribution (left), both have the same effect, leading to an equivalent CVaR estimate (vertical black line). However, for a Bernoulli distribution, our method can leave the empirical CVaR estimate unchanged while direct bonuses always result in a looser estimate.

empirical CDF will yield a CVaR transformation that depends on the shape of the CDF itself. In contrast, adding a bonus directly to the empirical CVaR is agnostic of the CDF structure, and relies only on the number of samples observed. To gain some geometric intuition for the superiority of our method, consider a distribution where the probability mass in the low-reward tail is clustered around a single point. In this scenario, one would hope to estimate the CVaR quite quickly as the tail variance [Valdez, 2004] is small, and indeed the lowest α fraction of optimistic samples will be almost unchanged by our optimistic perturbation. Figure 2.7 illustrates this by looking at the optimistic CVaR estimates generated for two different distributions.

Cassel et al. [2018] show that their U-UCB approach achieves a problem-dependent "proxy regret" bound of order $O\left(U^2 \log n/\alpha^2\right)$ which matches our bound above.(To be exact, the proxy regret as presented in Cassel et al. [2018] is not a cumulative notion, and thus the bound is presented as $O\left(U^2 \log n/\alpha^2 n\right)$, with an additional n present in the denominator. For clarity's sake, we deal with the cumulative version here.) Note however that their proxy regret is potentially a slightly less standard notion of regret. While we show in Appendix A.3.3 that it is an upper bound on our CVaR-regret, our CVaR-regret is amenable to a simpler analysis and is still a good measure of the algorithm's performance; e.g., to achieve sub-linear CVaR-regret the algorithm needs to play the optimal arm more and more frequently. However, to rule out the possibility that our algorithm's superior performance is due to the use of a different objective, Proposition 2 shows that CVaR-UCB achieves the same dependency on α and n as U-UCB, even when evaluating on proxy regret.

We further illustrate the benefits of our approach over direct bonuses by devising a variant of

U-UCB with even better dependence on risk level α. This algorithm, Brown-UCB, leverages a CVaR concentration inequality from Brown [2007] to compute a bonus. We show in the Appendix A.3.2 that its problem-dependent CVaR-regret grows at $O\left(U^2 \log n/\alpha\right)$. Similar to Thomas and Learned-Miller [2019]'s observation in the concentration inequalities setting, we will soon see that while our direct bonus approach has a slightly improved dependency on the risk level compared to the CVaR-UCB above, the latter's empirical advantages can be significant, highlighting the practical significance of leveraging the specific CDF structure when computing a bonus.

Proposition 2. *Consider a stochastic K-armed bandit problem with rewards bounded in* $[0, U]$. *For any given horizon* n *and risk level* α, *both* CVaR-*UCB and U-UCB incur proxy regret with* $O(\frac{\log n}{n})$ *and* $O(1/\alpha^2)$ *dependency on the horizon and risk level, respectively.*

2.7.4 Empirical Evaluations

We present a series of experiments demonstrating the superior performance of our algorithm across several types of distributions and numbers of arms.

Truncated Normal Environments First, we present results on 3-armed truncated-normal bandits with varying risk levels and CVaR-regret gaps. We compare our CVaR-UCB with four others:

1. an ϵ-greedy algorithm, which chooses a random arm with probability $\epsilon = 0.1$, and otherwise the arm with highest empirical CVaR.
2. the CVaR best-arm identification algorithm from Kolla et al. [2019].
3. the U-UCB algorithm from Cassel et al. [2018].
4. a variant of U-UCB called Brown-UCB presented in Appendix A.3.2

In each distribution, the arms correspond to truncated normal distributions with different means and variances. The parameters of these distributions, along with their CVaRs, are shown in Table 2.2. The difference between the Easy and Hard environments is that the arms in the Hard environment feature a smaller difference between their CVaR values compared to those in the Easy environment. We also assess our algorithm's performance on two runs of the Hard environment with different levels of α. Note that we began each experiment by pulling each arm $\lceil 1/\alpha \rceil$ times to ensure there are enough samples so that the empirical CVaR is properly defined.

The results in Figure 2.8 show that the CVaR-regret of our algorithm is an order of magnitude lower than the reward bonus-based algorithms we compared against (note the log-scale). Furthermore, as expected, the cumulative-CVaR regret of our algorithm grows logarithmically, as expected. This is in contrast to the linear growth of ϵ-greedy and Kolla et al. [2019], which was not designed for regret minimization.

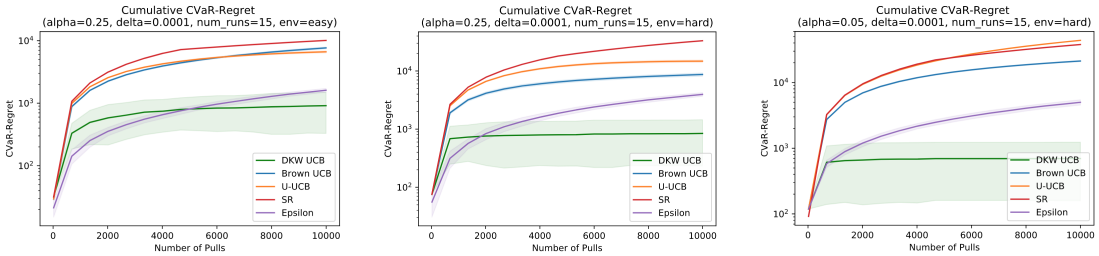
Environment	Arm 1	Arm 2	Arm 3
Easy, $\alpha = 0.25$	3.18	**9.32**	-1.02
Hard, $\alpha = 0.25$	6.59	**9.32**	6.91
Hard, $\alpha = 0.05$	4.91	**6.52**	2.43

(a)

Parameter	Arm 1	Arm 2	Arm 3
Easy & Hard μ	10	15	16
Easy σ	6	5	15
Hard σ	3	5	8

(b)

Table 2.2: Setup for the Three-Arm Multi-Arm Bandit Environments. (a) CVaR values of different arms in our simulations. In all cases, Arm 2 (bolded) has the highest CVaR. (b) Means and standard deviations for each arm for the Easy and Hard environments. All distributions were truncated at $\pm 1\sigma$.

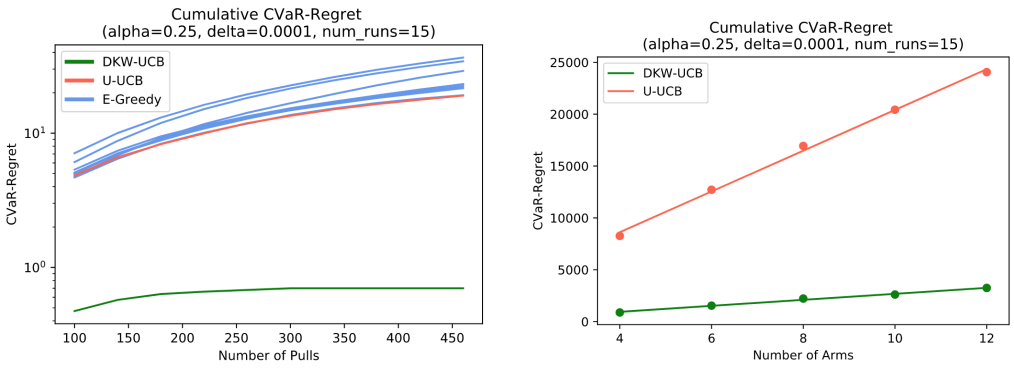


(a) (b) (c)

Figure 2.8: Cumulative CVaR-regret of CVaR-UCB (green; our algorithm), ϵ-greedy (purple), Cassel et al. [2018]'s U-UCB (orange), Brown-UCB Brown [2007] (blue), and Kolla et al. [2019]'s successive rejects algorithm (red) for different bandit setups. While the arm bonuses in rightshifting algorithms are only dependent on the number of samples gathered, the those in DKW-UCB depend further on the particular values of those samples, leading to larger variance than the other algorithms. Means and 95% confidence intervals shown for fifteen runs, with $\delta = 10^{-4}$. Y-axis has log scale. (a) Easy Bandit with $\alpha = 0.25$ (b) Hard Bandit with $\alpha = 0.25$ (c) Hard Bandit with $\alpha = 0.05$

Comparison against a Tuned ϵ-Greedy Baseline In practice, the parameters used for ϵ-greedy can have a substantial impact on its performance. For example, in the risk-neutral case, knowledge of the optimality gaps can be leveraged to create an decaying ϵ-greedy algorithm with logarithmic regret growth [Bietti et al., 2018]. Thus, to demonstrate the practical relevance of our algorithm, it is important to check that finding a successful decay schedule for ϵ-greedy is not easy. Thus, we introduce the Bernoulli Bandit environment, which is designed to penalize algorithms which explore either too little or too much. In addition to comparing against our algorithm, we also compare against the U-UCB algorithm from Cassel et al. [2018].

The environment consists of two arms: a deterministic arm that always returns 0.1 reward, and a stochastic that returns reward 1 with probability 0.8 and returns 0 otherwise. The $\text{CVaR}_{0.25}$ of the stochastic arm is 0.2, compared to 0.1 for the deterministic arm. This environment highlights an algorithm's shortcomings in handling exploration: an suboptimal algorithm will either not explore enough and settle for the inferior deterministic arm, or explore for too long and incur large CVaR-regret. We ran a decaying ϵ-greedy algorithm with a range of exponential decay constants spanning



(a) (b)

Figure 2.9: (a) Cumulative CVaR-regret of our algorithm (green), ϵ-greedy (blue), and Cassel et al. [2018]'s U-UCB (red) on the Bernoulli Bandit environment. The ϵ-greedy algorithm was run with a wide range of starting epsilons and decay constants. Results averaged over 15 runs. Y-axis has log scale. (b) Cumulative CVaR-regret of our algorithm on the One Good Arm environment for different numbers of arms. Values were collected after 3500 pulls and averaged over 15 runs.

from 10^{-1} to 10^{-6} and initial epsilons ranging from 1 to 10^{-3}.

The results of our experiments are shown in Figure 2.9 (a), and show that all algorithms incur at least an order of magnitude higher cumulative CVaR-regret than ours. Note that the successive rejects algorithm requires upper and lower bounds of the CVaR-regret gaps between the optimal arm and all other arms. To ensure as favorable a comparison as possible, we provide the tightest possible bounds based on the true CVaRs of the arms.

Dependence on Number of Arms We also performed an empirical evaluation of our algorithm's dependence on number of arms K in an environment called One Good Arm. In Figure 2.9 (b), we plot the cumulative CVaR-regret of our algorithm and Cassel et al. [2018]'s U-UCB after 3500 pulls for various values of k. The empirical results show what our bound predicts: the CVaR-regret grows linearly with number of arms K. Moreover, our algorithm continues to significantly outperform U-UCB as the number of arms increases.

The One Good Arm environment consists of one normal distribution with mean 0 and variance 1 ($\text{CVaR}_{0.25} \approx -0.703$), and the rest of the arms with mean 0.1 and standard deviation 10 ($\text{CVaR}_{0.25} \approx -6.928$). Distributions were truncated at $\pm 1\sigma$.

2.8 Related Work

Optimizing policies for risk sensitivity in MDPs has been long studied, with policy gradient [Tamar et al., 2015b,a], actor critic [Tamar and Mannor, 2013] and TD methods [Tamar and Mannor, 2013, Sato et al., 2001]. While most of this work considers mean-variance trade objectives, [Chow

et al., 2015] establish a connection between a optimizing CVaR and robustness to modeling errors, presenting a value iteration algorithm.

In contrast, we do not assume access to transition and rewards models. [Chow and Ghavamzadeh, 2014] present a policy gradient and actor-critic algorithm for an expectation-maximizing objective with a CVaR constraint. None of these works considers systematic exploration but rely on heuristics such as ϵ-greedy or on the stochasticity of the policy for exploration. Instead, we focus on how to explore systematically to find a good CVaR-policy.

Our work builds upon recent advances on distributional RL [Bellemare et al., 2017, Rowland et al., 2018, Dabney et al., 2018] which are still concerned with optimizing expected return. Notably, [Dabney et al., 2018] aims to train risk-averse and risk-seeking agents, but does not address the exploration problem or attempts to find optimal policies quickly.

Dilokthanakul and Shanahan [2018] uses risk-averse objectives to guide exploration for good performance w.r.t. expected return. [Moerland et al., 2018] leverages the return distribution learned in distributional RL as a means for optimism in deterministic environments. Mavrin et al. [2019] follow a similar pattern but can handle stochastic environments by disentangling intrinsic and parametric uncertainty. While they also evaluate the policy that picks the VaR-greedy action in one experiment, their algorithm still optimizes expected return during learning. In general, these approaches are fundamentally different from ours which learns CVaR policies in stochastic environments efficiently by introducing optimism *into* the learned return distribution.

CVaR-Bandit Much work on risk-aware multi-armed bandits [Sani et al., 2012, Vakili and Zhao, 2016, Zimin et al., 2014, Vakili and Zhao, 2015] considers mean-variance objectives. We here consider CVaR due to its advantage as a coherent risk measure [Artzner et al., 1999]. Galichet et al. [2013] consider a CVaR-optimizing framework, but only analyze the case where $\alpha \to 0$, which corresponds to finding the arm distribution with the greatest essential infimum. In a pure exploration setting, Kolla et al. [2019] consider the task of finding the arm with the optimal CVaR with a successive rejects algorithm.

Outside the bandits framework, Brown [2007] and Thomas and Learned-Miller [2019] consider the problem of quantifying the uncertainty of a CVaR estimate from a set of samples. Both present concentration inequalities for CVaR, with the latter showing that the DKW inequality can provide much tighter bounds in practice.

There has been very recent related work Cassel et al. [2018] for bandits with more general risk measures. However, for learning CVaR bandit policies, their work relies on a more generic upper confidence bound that works in the exploration bonus framework, whereas we apply optimism at the sample level.

2.9 Summary and Conclusion

We present a new algorithm for quickly learning CVaR-optimal policies in Markov decision processes. This algorithm is the first to leverage optimism in combination with distributional reinforcement learning to learn risk-averse policies in a sample-efficient manner. Unlike existing work on expected return criteria which rely on reward bonuses for optimism, We introduce optimism by directly modifying the target return distribution and provide a theoretical justification that in the evaluation case for finite MDPs, this indeed yields optimistic estimates. We further empirically observe significantly faster learning of CVaR-optimal policies by our algorithm compared to existing baselines on several benchmark tasks. This includes simulated healthcare tasks where risk-averse policies are of particular interest: HIV medication treatment and insulin pump control for diabetes type 1 patients.

In Bandit setting, We present *distributionally-aware optimism* as a simple and effective way to reap the benefits of optimism in risk-sensitive multi-armed bandits. We provide theoretical results matching state-of-the-art and empirical results showing that our algorithm achieves an order-of-magnitude lower CVaR-regret than exploration bonus-based alternatives and other baselines. In the future, we aim to expand our analysis to contextual linear bandits, which capture the high-dimensionality and individualized treatment desired in many applications, including healthcare.

Chapter 3

Off-Policy Policy Evaluation Under Unobserved Confounding

When observed decisions depend only on observed features, off-policy policy evaluation (OPE) methods for sequential decision problems can estimate the performance of evaluation policies before deploying them. However, this assumption is frequently violated due to unobserved confounders, unrecorded variables that impact both the decisions and their outcomes. We assess robustness of OPE methods under unobserved confounding by developing worst-case bounds on the performance of an evaluation policy. When unobserved confounders can affect every decision in an episode, we demonstrate that even small amounts of per-decision confounding can heavily bias OPE methods. Fortunately, in a number of important settings found in healthcare, policy-making, and technology, unobserved confounders may directly affect only one of the many decisions made, and influence future decisions/rewards only through the directly affected decision. Under this less pessimistic model of one-decision confounding, we propose an efficient loss-minimization-based procedure for computing worst-case bounds, and prove its statistical consistency. On simulated healthcare examples—management of sepsis and interventions for autistic children—where this is a reasonable model, we demonstrate that our method invalidates non-robust results and provides meaningful certificates of robustness, allowing reliable selection of policies under unobserved confounding.

3.1 Introduction

New technology and regulatory shifts have allowed the collection of vast amounts of data on sequential trajectories of past decisions and associated rewards, ranging from healthcare decisions and outcomes to product recommendations and purchase histories. This presents unique opportunities for using off-policy methods to inform better sequential decision-making. Leveraging prior data to

evaluate the performance of a sequential decision policy (which we call the *evaluation policy*) before deploying it can reduce the need for online experimentation when doing so is expensive or risky.

A central challenge in off-policy policy evaluation (OPE) is that the estimand is inherently counterfactual: what would the rewards be *if an alternate policy had been used* (the counterfactual) instead of the behavior policy that generated the observed data (the factual). As a result, OPE requires causal reasoning about whether observed rewards were caused by observed decisions, or by a common causal variable that simultaneously affects observed decisions and states / rewards [Hernán and Robins, 2020]. In order to make counterfactual evaluations possible, a standard assumption—albeit often overlooked and unstated—is to require that the behavior policy does not depend on any unobserved variables that also affect the future states/rewards (no unobserved confounding). We refer to this assumption as *sequential ignorability*, following the line of works on dynamic treatment regimes [Robins, 1986, Murphy, 2003].

Sequential ignorability, however, is often violated in OPE problems where the behavior policy is unknown. In medicine, business operations, and automated systems in tech, decisions depend on unlogged features correlated with future outcomes. As an example, clinicians use visual observations or discussions with patients to inform treatment, but such information is typically not recorded; they also may rely on heuristics that are hard to quantify, and can over-extrapolate from past experience [McDonald, 1996]. In judicial decisions, psychological factors affect bail and parole decisions [Dhami, 2003, Danziger et al., 2011]. Heuristics are prevalent in business contexts, for example in venture capital investments [Åstebro and Elhedhli, 2006], and customer targeting [Wübben and Wangenheim, 2008]. Even automated policies in tech firms depend on unlogged features [Agarwal et al., 2016], and complex software and data infrastructures often introduce confounding.

In this chapter, we study a framework for quantifying the impact of unobserved confounders on OPE estimates, developing worst-case bounds on the performance of an evaluation policy. OPE estimates are often used to inform policy selection, and we are particularly interested in methods that can guide when we may be confident (or not) that an alternate decision policy should be preferred. Since OPE is generally impossible under arbitrary unobserved confounding, we begin by positing a model that explicitly limits their influence on decisions. Our proposed model is a natural extension of an influential confounding model for a single binary decision Rosenbaum [2002] to the multi-action sequential decision making setting. When unobserved confounders can affect all decisions, we illustrate in Section 3.4 that even small amounts of confounding can have an exponential (in the number of decisions) impact on the bias of OPE. In this sense, the accuracy of OPE can be highly unreliable under the presence of unobserved confounding that affect all decisions in multi-step horizon problems.

Fortunately, in a number of important applications, unobserved confounders may only directly affect a single decision, and influence future decisions/rewards only through this directly affected decision. As we detail shortly, in healthcare this happens when a high-level expert makes an initial

decision potentially using unrecorded information, after which a standard set of protocols are followed based on recorded inputs. In financial services, this happens when humans initially screen new clients for fraud, after which decisions are made based on standard logged features. For online services, personalized systems can adapt to individual users given enough data, but an initially poor experience leads users to stop using the service before the system can learn [Hashimoto et al., 2018]. Demographic characteristics—many of which are unobservable to typical online services—can significantly affect initial performance [Amodei et al., 2016, Grother et al., 2010, Hovy and Søgaard, 2015, Blodgett et al., 2016, Sapiezynski et al., 2017, Tatman, 2017], and is an important unobserved confounder. In these scenarios, future outcomes typically depend on unseen demographics only through the initial engagement levels, in that the availability of individualized data can improve the personalized system's subsequent performance. In order to evaluate new policies (e.g. fully automated systems), we need to account for unobserved confounding in the initial human-conducted screening decisions; in this scenario, the unobserved features affect subsequent decisions and rewards only through the initial decision and outcome. In other instances, it may be the case that an unobserved confounder at a particular time step is observed in the next period.

Under our less pessimistic model of single-decision confounding, we develop bounds on the expected cumulative rewards under the evaluation policy. We use functional convex duality to derive a dual relaxation, and show that it can be computed by solving a loss minimization problem. The single-decision confounding model allows us to efficiently evaluate these bounds even for continuous states, unlike the general case which requires solving an intractable nonconvex problem over likelihood ratios (which are infinite dimensional for continuous states). We prove the empirical approximation of our procedure is consistent, allowing estimation from observed past decisions.

The single-decision confounding model may not fully describe scenarios where unobserved confounders affect decisions through multiple periods. Our sensitivity analysis method is nevertheless a meaningful tool even in such scenarios, as certifying the robustness of OPE against single-decision confounding is a *necessary* condition for the conclusion of OPE to withstand multi-decision confounding. As we present in the sequel, we observe conclusions of OPE are often invalidated even under less conservative single-decision confounding, raising substantial concern for robustness of OPE under violations of sequential ignorability.

On examples of dynamic treatment regimes for autism and sepsis management, we illustrate how our single-decision confounding model allows informative bounds over meaningful amounts of confounding. Our approach provides certificates of robustness by identifying the level of unobserved confounding at which the potential bias in OPE estimates raise concerns about the validity of selecting the best policy among a set of candidates. Compared to our informative bounds, the naïve approach is prohibitively conservative and lose robustness certificates for even negligible amounts of confounding.

3.2 Motivating example: managing sepsis patients

Sepsis in ICU patients accounts for one third of deaths in hospitals [Howell and Davis, 2017]. Sepsis treatment decisions are made by a clinical care team, including nurses, residents, and ICU attending physicians and specialists [Rhodes et al., 2017]. Difficulties of care often lead to decisions based on imperfect information, and AI-based approaches provide an opportunity for automated management of important medications for sepsis, including antibiotics and vasopressors. These approaches can decide to notify the care team when a patient should be placed on a mechanical ventilator, freeing the care team to allocate more time to critical cases. Motivated by these opportunities, and the availability of data from MIMIC-III [Johnson et al., 2016], several AI-based approaches for sepsis management have been proposed [Futoma et al., 2018, Komorowski et al., 2018, Raghu et al., 2017].

Due to safety concerns, new policies need to be evaluated offline before online clinical validation. Confounding, however, is a serious issue in data generated from ICUs: patients in emergency departments often do not have an existing record in the hospital's electronic health system, leaving a substantial amount of patient-specific information unobserved in subsequent offline analysis. As a prominent example, comorbidities that significantly complicate the cases of sepsis [Brent, 2017] are often unrecorded. Private communication with an emergency department physician revealed that *initial* treatment of antibiotics at admission to the hospital are often confounded by unrecorded factors that affect the eventual outcome. For example, comorbidities such as undiagnosed heart failure can delay diagnosis of sepsis, leading to slower use of antibiotics. There is considerable discussion in the medical literature on the importance of quickly beginning antibiotic treatment, with frequently noted concerns about confounding, as these discussions are largely based on observational data [Seymour et al., 2017, Sterling et al., 2015].

We consider choosing between two automated policies that differ only in initially *avoiding*, or *prescribing* antibiotics, and otherwise similarly improve current standard care. Antibiotics often are viewed as a better treatment for sepsis. In this example, unobserved factors critically effect the decision to prescribe antibiotics upon arrival; since the care team is highly trained, we assume they follow standard protocols based on recorded measurements in subsequent time steps. In Section 3.6, we assess the impact of confounding on OPE in this decision process with a single step of confounding.

3.3 Formulation

Notation conventions vary in the diverse set of communities interested in learning from (sequential) observational data. We use the potential outcomes [Rubin, 2005] notation to make explicit which sequence of actions we wish to evaluate versus which sequence of actions were actually observed. In this approach, we posit all potential states and rewards exist for each possible sequence of actions, but we only observe the one corresponding to the actions taken (also known as partial, or bandit feedback), making the other potential states and rewards counterfactual. Literature in batch off

policy reinforcement learning almost always assumes sequential ignorability, in which case the distribution of *potential* states and rewards are independent of the action taken by the behavior policy, conditional on the observed history. This allows us to consistently estimate counterfactuals simply based on observed outcomes. However, since our aim is to consider the impact of hypothetized confounding, clarifying the difference between the potential and observed states and rewards is cumbersome, but important.

We focus on domains modeled by episodic stochastic decision processes with a discrete set of actions. Let $\mathcal{A}_t$ be a finite set of actions available at time $t = 1, .., T$. Denote a sequence of actions $a_1 \in \mathcal{A}_1, .., a_T \in \mathcal{A}_T$ by $a_{1:T}$ (and similarly $a_{t:t'}$ for arbitrary indices $1 \leq t \leq t' \leq T$ with $a_{1:0} = \varnothing$). For any sequence of actions $a_{1:t}$, let $S_t(a_{1:t-1})$ be the continuous-valued vector of state features, including previous rewards, $R_t(a_{1:t})$ be the reward at time t, and the return $Y(a_{1:T}) := \sum_{t=1}^{T} \gamma^{t-1} R_t(a_{1:t})$ be the discounted sum of rewards. We denote by $W(a_{1:T}) = (S_1, .., S_T(a_{1:T-1}), R_1(a_1), .., R_T(a_{1:T}))$ the sequence of all potential outcomes (over rewards and states) associated with the action sequence $a_{1:T}$. Any sum $\sum_{a_{1:T}}$ over action sequences is over all $a_{1:T} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_T$.

In the off-policy setting, we observe actions $A_1, .., A_T$ generated by an unknown behavior policy $\pi_1, .., \pi_T$. Let H_t denote the observed history, with $H_1 := S_1$ and

$$H_t := (S_1, A_1, S_2(A_1), A_2, .., S_t(A_{1:t-1})) \quad \text{for} \quad t \geq 2.$$

For any fixed sequence of actions $a_{1:t-1}$, denotes an instantiation of H_t following the action sequence by $H_t(a_{1:t-1}) := (S_1, A_1 = a_1, S_2(a_1), .., A_{t-1} = a_{t-1}, S_t(a_{1:t-1}))$.

Let $\mathcal{H}_t$ be the set of all histories. When there is *unobserved confounding* U_t, the behavioral policy draws actions $A_t \sim \pi_t(\cdot \mid H_t, U_t)$. Let $\pi_t(\cdot \mid H_t)$ be the conditional distribution of A_t given only the observed history H_t, marginalizing out the unobserved confounder U_t.

Our goal is to bound the performance of an evaluation policy $\bar{\pi}_1, .., \bar{\pi}_T$ in a confounded sequential off-policy environment. Let $\bar{A}_t \sim \bar{\pi}_t(\cdot \mid \bar{H}_t)$ be the actions generated by the evaluation policy at time t, where we use $\bar{H}_t := (S_1, \bar{A}_1, S_2(\bar{A}_1), \bar{A}_2, .., S_t(\bar{A}_{1:t-1}))$ and $\bar{H}_t(a_{1:t-1}) := (S_1, \bar{A}_1 = a_1, S_2(a_1), \bar{A}_2 = a_2, .., S_t(a_{1:t-1}))$ to denote the history under the evaluation policy, analogously to the shorthands $H_t, H_t(a_{1:t-1})$; note that $\bar{H}_t$ are counterfactuals never observed in the behavioral data. We are interested in estimation of the expected cumulative reward

$$V^{\bar{\pi}} = \mathbb{E}\left[Y(\bar{A}_{1:T})\right]$$

under the evaluation policy. Because we only observe potential outcomes $W(A_{1:t})$ evaluated at the actions $A_{1:t}$ taken by the behavior policy π_t, we need to express $\mathbb{E}[Y(\bar{A}_{1:T})]$ in terms of observable data generated by the behavioral policy π_t. To do so, we use the following definitions and assumptions.

Assumption 1. *Overlap holds with respect to the conditional distributions over actions given only*

the histories between the behavior policy and the evaluation policy. That is, $\pi_t(a_t \mid H_t) > 0$ *whenever* $\bar{\pi}_t(a_t \mid \bar{H}_t) > 0$*, for all* t *and* a_t*, and almost every* H_t*.*

Definition 1 (Sequential Ignorability)**.** *A policy satisfies sequential ignorability if* $\forall t$*, conditional on the history generated by the policy, the action generated by the policy is independent of the potential outcomes* $R_t(a_{1:t}), S_{t+1}(a_{1:t}), R_{t+1}(a_{1:t+1}), .., S_T(a_{1:T-1}), R_T(a_{1:T})$ *for all* $a_{1:T} \in \mathcal{A}_1 \times \cdots \mathcal{A}_T$*.*

Sequential ignorability is a natural condition required for the evaluation policy to be well-defined: any additional randomization used by the evaluation policy $\bar{\pi}_t(\cdot \mid \bar{H}_t)$ cannot depend on unobserved confounders. We assume that the evaluation policy always satisfies this assumption.

Assumption 2. *The evaluation policy satisfies sequential ignorability (Definition 1).*

OPE fundamentally requires counterfactual reasoning since we only observe states $S_t(A_{1:t-1})$ and rewards $R_t(A_{1:t})$ generated by the behavioral policy. The canonical assumption in batch off-policy RL is that sequential ignorability holds for the evaluation *and the behavior policy* [Robins, 1986, 2004, Murphy, 2003]. For example, sequential ignorability allows the application of the standard importance sampling formula using the observed data; we give its proof in Section B.2.1. To ease notation, let

$$\rho_t := \frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t-1}))}{\pi_t(A_t \mid H_t)}.$$

Lemma 1. *If sequential ignorability holds for both* π *and* $\bar{\pi}$*, then*

$$\mathbb{E}\left[Y(\bar{A}_{1:T})\right] = \mathbb{E}\left[Y(A_{1:T}) \prod_{t=1}^{T} \rho_t\right]$$

3.4 Bounds under unobserved confounding

Despite the advantageous implications, it is often unrealistic to assume that the behavior policy π_t satisfies sequential ignorability (Definition 1). If the unobserved confounder U_t, introduced in Section 3.3, contains information about unseen potential rewards, then sequential ignorability doesn't hold.

We now relax the sequential ignorability of the behavior policy, and instead posit a model of bounded confounding, then develop worst-case bounds on the evaluation policy performance $\mathbb{E}[Y(\bar{A}_{1:T})]$ under this model.

Without loss of generality, let the confounder U_t be such that the potential outcomes are independent of A_t when conditioning on U_t alongside the observed states. Such an unobserved confounder always exists since we can define U_t to be the tuple of all unseen potential outcomes.

Assumption 3. *For all* $t = 1, .., T$*, there is a random vector* U_t *s.t. conditional on the history* H_t *generated by the behavior policy* ***and*** U_t*,* $A_t \sim \pi_t(\cdot \mid H_t, U_t)$ *is independent of the potential outcomes* $R_t(a_{1:t}), S_{t+1}(a_{1:t}), R_{t+1}(a_{1:t+1}), .., S_T(a_{1:T-1}), R_T(a_{1:T})$ *for all* $a_{1:T} \in \mathcal{A}_1 \times \cdots \mathcal{A}_T$*.*

Under arbitrary unobserved confounding, $\mathbb{E}[Y(\bar{A}_{1:T})]$ is not identifiable, meaning that many different values are consistent with the observed data. However, it is often plausible to posit that the unobserved confounder U_t has limited influence on the decisions of the behavior policy. In such a case, we may expect OPE estimates that (incorrectly) assume sequential ignorability may not be too biased.

Consider the following model of unobserved confounding for sequential decision making problems, which bounds confounder's influence on the behavior policy's decisions.

Assumption 4. *For $t = 1, .., T$, there is a $\Gamma_t \geq 1$ such that for any $a_t, a'_t \in \mathcal{A}_t$*

$$\frac{\pi_t(a_t \mid H_t, U_t = u_t)}{\pi_t(a'_t \mid H_t, U_t = u_t)} \frac{\pi_t(a'_t \mid H_t, U_t = u'_t)}{\pi_t(a_t \mid H_t, U_t = u'_t)} \leq \Gamma_t \tag{3.1}$$

*almost surely over H_t, and u_t, u'_t, and sequential ignorability holds conditional on H_t **and** U_t.*

The bound (Equation 3.1) is a natural extension of a classical model of confounding proposed by Rosenbaum [2002] for a single decision ($T = 1$) to sequential problems. For binary actions $\mathcal{A}_t = \{0, 1\}$, our bounded unobserved confounding assumption is equivalent [Rosenbaum, 2002] to the logistic model

$$\log \frac{\mathbb{P}(A_t = 1 \mid H_t, U_t)}{\mathbb{P}(A_t = 0 \mid H_t, U_t)} = \kappa(H_t) + (\log \Gamma_t) b(U_t)$$

for some function $\kappa(\cdot)$ and a bounded $b(\cdot)$ taking values in $[0, 1]$.

In the sequential setting where $T > 1$, confounding can lead to exponentially large (in T) over-sampling of large (or small) rewards, introducing a large un-correctable bias. As an illustration, consider the simplified setting where for a single unobserved confounder $U \sim \text{Unif}(\{0, 1\})$, behavioral actions $A_1, \ldots, A_T \in \{0, 1\}$ are drawn conditionally on U, but independent of one another, with the conditional distribution $\mathbb{P}(A_t = 1|U = 1) = \sqrt{\Gamma}/(1 + \sqrt{\Gamma})$ and $\mathbb{P}(A_t = 1|U = 0) = 1/(1 + \sqrt{\Gamma})$. Let the return be $Y(a_{1:T}) = U$ for all action sequences $a_{1:T}$. Although the actions do not affect the outcome, the likelihood of observing $((A_t = 1)_{t=1}^T, Y = 1)$ is $\Gamma^{T/2}/(2(1 + \sqrt{\Gamma})^T)$, whereas the likelihood of observing $((A_t = 1)_{t=1}^T, Y = 0)$ is $1/(2(1 + \sqrt{\Gamma})^T)$. Even in the limit of infinite observations, OPE will mistakenly estimate that always taking $\bar{A}_t = 1$ leads to better rewards than always taking $\bar{A}_t = 0$. The effect of confounding is salient even in this toy example where states don't exist and rewards don't depend on actions. This has important implications for off-policy policy selection or optimization, where systematic biases can lead to selection of a poorly performing policy.

3.5 Confounding in a single decision

In many important applications, it is realistic to assume there is only a single step of confounding at a known time step t^*. Under this assumption, we outline in this section how we obtain a computationally and statistically feasible procedure for computing a lower (or upper) bound on the

value $\mathbb{E}[Y(\bar{A}_{1:T})]$ of an evaluation policy $\bar{\pi}$. Robustness of the policy value with respect to this class of confounding effects is also a necessary (although not sufficient) requirement for robustness to confounding at multiple time steps, making this proposed method a valuable starting place for evaluating the potential effects of confounding even if one believes that there may be confounding at multiple times. After introducing precisely our model of confounding, we show in Proposition 3 how the evaluation policy value can be expressed using likelihood ratios over potential outcomes that can be used to relate the potential outcomes over observed (factual) actions with counterfactual actions not taken. These likelihood ratios over potential outcomes are unobserved, but a lower bound on the evaluation policy value can be computed by minimizing over all feasible likelihood ratios that satisfy our model of bounded confounding. Towards computational tractability, we derive a dual relaxation that can be represented as a loss minimization procedure.

We define the confounding model for when there is an unobserved confounding variable U that only affects the behavior policy's action at a single time $t^\star \in [T]$. For example, in looking at the impact of confounders on antibiotics in sepsis management, it is plausible to assume that while confounders may influence the first decision when the patient arrives, later treatment decisions are un-confounded.

Assumption 5. *For all $t \neq t^\star$, conditional on the history H_t generated by the behavior policy, A_t is independent of the potential outcomes $R_t(a_{1:t}), S_{t+1}(a_{1:t}), R_{t+1}(a_{1:t+1}), .., S_T(a_{1:T-1}), R_T(a_{1:T})$ for all $a_{1:T} \in \mathcal{A}_1 \times \cdots \mathcal{A}_T$. For $t = t^\star$, there exists a random variable U such that the same conditional independence holds only when conditional on the history H_t **and** U.*

Restricting Assumption 4 to a single time step t^*, we assume U has bounded influence on $A_{t^\star}$.

Assumption 6. *There is a $\Gamma \geq 1$ such that for any $a_{t^\star}, a'_{t^\star} \in \mathcal{A}_{t^\star}$, and a.s. over $H_{t^\star}$, and u, u'*

$$\frac{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star}, U = u)}{\pi_{t^\star}(a'_{t^\star} \mid H_{t^\star}, U = u)} \frac{\pi_{t^\star}(a'_{t^\star} \mid H_{t^\star}, U = u')}{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star}, U = u')} \leq \Gamma. \tag{3.2}$$

Selecting the amount of unobserved confounding Γ is a modeling task, and the above confounding model's simplicity and interpretability makes it advantageous for modelers to argue a plausible value of Γ. As in any applied modeling problem, the amount of unobserved confounding Γ should be chosen with expert knowledge (e.g. by consulting doctors that make behavioral decisions). In Section 3.6, we give application contexts where a realistic range of Γ can be posited. One of the most interpretable ways to assess the level of robustness to confounding is via the *design sensitivity* of the analysis [Rosenbaum, 2010]: the value of Γ at which the bounds on the evaluation policy's value crosses a landmark threshold (e.g. performance of behavior policy or some known safety threshold).

By directly applying the bound (Equation 3.2) to adjust importance weights, we obtain a simple naive lower bound on the evaluation policy performance $\mathbb{E}[Y(\bar{A}_{1:T})]$. Details are in Section B.3.1.

Lemma 2. *Let Assumptions 2, 5, 6 hold. Then, we have*

$$\mathbb{E}[Y(\bar{A}_{1:T})] \geq \mathbb{E}\Bigg[Y(A_{1:T}) \times \big(\Gamma \mathbf{1}\{Y(A_{1:T}) < 0\} + \Gamma^{-1}\mathbf{1}\{Y(A_{1:T}) > 0\}\big) \times \prod_{t=1}^{T} \rho_t\Bigg]. \tag{3.3}$$

This bound is often prohibitively conservative, as we illustrate in Section 3.6. Instead we derive a tighter bound on the evaluation policy performance based on a constrained convex optimization formulation over counterfactual distributions. Under Assumption 6, the likelihood ratio between observed and unobserved distribution at $t^\star$ can at most vary by a factor of Γ. Recall that $W(a_{1:T})$ is the tuple of all potential outcomes associated with the actions $a_{1:T}$. The following observation is due to Yadlowsky et al. [2018, Lemma 2.1].

Lemma 3. *Under Assumptions 5, 6, for all $a_{t^\star} \neq a'_{t^\star}$, the likelihood ratio over the tuple of potential outcomes $W := \{W(a_{1:T})\}_{a_{1:T}}$ exists,*

$$\mathcal{L}(\cdot; H_{t^\star}, a_{t^\star}, a'_{t^\star}) := \frac{dP_W(\cdot \mid H_{t^\star}, A_{t^\star} = a'_{t^\star})}{dP_W(\cdot \mid H_{t^\star}, A_{t^\star} = a_{t^\star})}$$

and for $\mathbb{P}_W(\cdot \mid H_{t^\star}, A_{t^\star} = a_{t^\star})$-a.s. for all w, w',

$$\mathcal{L}(w; H_{t^\star}, a_{t^\star}, a'_{t^\star}) \leq \Gamma \mathcal{L}(w'; H_{t^\star}, a_{t^\star}, a'_{t^\star}). \tag{3.4}$$

We let $\mathcal{L}(\cdot; H_{t^\star}, a_{t^\star}, a_{t^\star}) \equiv 1$. Using these (unknown) likelihood ratios, we can express the value of the evaluation policy, $\mathbb{E}[Y(\bar{A}_{1:T})]$. The proof is given in Section B.2.2.

Proposition 3. *Under Assumptions 2, 5, 6, $\mathbb{E}[Y(\bar{A}_{1:T})]$ is equal to*

$$\mathbb{E}[Y(\bar{A}_{1:T})] = \mathbb{E}\Bigg[\prod_{t=1}^{t^\star-1} \rho_t \sum_{a_{t^\star}, a'_{t^\star}} \bar{\pi}_{t^\star}(a_{t^\star} \mid \bar{H}_{t^\star}(A_{1:t^\star-1})) \pi_{t^\star}(a_{t^\star} \mid H_{t^\star}) \times \mathbb{E}\Big[\mathcal{L}(W; H_{t^\star}, a_{t^\star}, a'_{t^\star}) Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t \;\Big|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\Big]\Bigg],$$

Proposition 3 implies a natural bound on the evaluation policy value $\mathbb{E}[Y(\bar{A}_{1:T})]$ under bounded unobserved confounding. Since the likelihood ratios $\mathcal{L}(\cdot; \cdot, a_{t^\star}, a'_{t^\star})$ are fundamentally unobservable, due to their counterfactual nature we take a worst-case approach over all likelihood ratios that vary by at most a factor of Γ (per Lemma 3), and derive a bound that only depends on observable distributions. Taking the infimum over the inner expectation in the expression derived in Proposition 3, and noting that it does not depend on $a'_{t^\star}$,

$$\eta^\star(H_{t^\star}; a_{t^\star}) := \inf_{L \in \mathfrak{L}} \mathbb{E}\left[L(W; H_{t^\star}) Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\right]$$

where $\mathfrak{L}$ is the set of measurable mappings $L : \mathcal{W} \times \mathcal{H}_{t^\star} \to \mathbb{R}_+$ satisfying $L(w; H_{t^\star}) \leq \Gamma L(w'; H_{t^\star})$ a.s. all w, w', and $\mathbb{E}[L(W; H_{t^\star}) \mid H_{t^\star}, A_{t^\star} = a_{t^\star}] = 1$. Since the above optimization is over infinite-dimensional likelihoods, it is difficult to compute. We use functional convex duality to derive a dual relaxation that can be computed by solving a *loss minimization* problem over any well-specified model class. This allows us to compute a meaningful lower bound to $\mathbb{E}[Y(\bar{A}_{1:T})]$ even when rewards and states are continuous, by simply fitting a model using standard supervised methods. For $(s)_+ = \max(s, 0)$ and $(s)_- = -\min(s, 0)$, define the weighted squared loss

$$\ell_\Gamma(z) := \frac{1}{2}(\Gamma(z)_-^2 + (z)_+^2).$$

Theorem 3. *Let Assumptions 2, 5, 6 hold. If*

$$\mathbb{E}[Y(A_{1:T})^2 \prod_{t=t^\star+1}^{T} \rho_t^2 \mid A_{t^\star} = a_{t^\star}, H_{t^\star}] < \infty$$

a.s., then $\eta^\star(H_{t^\star}; a_{t^\star})$ is lower bounded a.s. by the unique solution

$$\kappa^\star(H_{t^\star}; a_{t^\star}) = \underset{f(H_{t^\star})}{\operatorname{argmin}} \quad \mathbb{E}\left[\frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star})} \times \ell_\Gamma\left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - f(H_{t^\star})\right)\right].$$

See Section B.3.2 for the proof. From Theorem 3, our final lower bound on $\mathbb{E}[Y(\bar{A}_{1:T})]$ is given by

$$\mathbb{E}\Big[\prod_{t=1}^{t^\star-1} \rho_t \sum_{a_{t^\star}} \bar{\pi}_{t^\star}(a_{t^\star} | \bar{H}_{t^\star}(A_{1:t^\star-1})) \times (1 - \pi_{t^\star}(a_{t^\star} | H_{t^\star}))\kappa^\star(H_{t^\star}; a_{t^\star}) + \pi_{t^\star}(A_{t^\star} | H_{t^\star}) Y(A_{1:T}) \prod_{t=1}^{T} \rho_t\Big]. \tag{3.5}$$

Our approach yields a loss minimization problem for each possible action, where the dimension of this supervised learning problem is that of the observed history $H_{t^\star}$. If confounding occurs very late in a decision process sequence, the space of histories can be very large and this may incur a significant computational cost. However If confounding occurs early, the space of possible histories is small and this learning problem becomes easier. This is the scenario for the domains we consider in our experiments. Compared to the non-sequential setting studied by Yadlowsky et al. [2018], the sequential nature of our problem requires carefully adjusting for future actions, which shows up as the product of importance weights inside the loss minimization problem. In the special case when the last decision is confounded ($t^\star = T$), our loss minimization formulation reduces to the non-sequential result due to Yadlowsky et al. [2018]. The optimality results from their work would then carry over to the method suggested here. However, when confounding occurs in any other

decision besides the last, the relaxation from η^* to κ^* makes the bounds feasible to compute, yet loose.

In cases where there is low, yet sufficient, overlap, weighted importance sampling (WIS) can dramatically reduce variance, at the cost of increased bias, with respect to the usual IS estimator. While our approach uses the IS to adjust for the differences between the behavior and evaluation policy, adjusting the bound in (Equation 3.5) to use WIS, instead, is straightforward. Altering the importance reweighting inside the loss function for $\kappa^\star$ to be normalized, like WIS, warrants further investigation.

Going beyond the single-decision confounding model (Equation 3.2) appears challenging both computationally and statistically. Under the general confounding model (Equation 3.1), we can formulate an optimization problem similar to that in Proposition 3 over multiple likelihood ratios corresponding to each confounded decision. Due to the multiplicative structure of the likelihood ratios, this is a nonconvex optimization problem, and convex duality does not apply. It is unclear how to develop statistically and computationally tractable reformulations of this problem akin to our loss minimization procedure, and we leave it as a topic of future research.

Consistency We now show that an empirical approximation to our loss minimization problem yields a consistent estimate of $\kappa^\star(\cdot)$. We require the standard overlap assumption that requires ρ_t be uniformly bounded for all t, that is, actions cannot be too rare under the behavior policy relative to the evaluation policy.

Since it is not feasible to optimize over the class of all functions $f(H_{t^\star})$, we consider a parameterization $f_\theta(H_{t^\star})$ where $\theta \in \mathbb{R}^d$. We provide provable guarantees in the simplified setting where $\theta \mapsto f_\theta$ is linear, so that the loss minimization problem is convex. That is, we assume that f_θ is represented by a finite linear combination of some arbitrary basis functions of $H_{t^\star}$. As long as the parameterization is well-specified so that $\kappa^\star(H_{t^\star}; a_{t^\star}) = f_{\theta^\star}(H_{t^\star})$ for some $\theta^\star \in \Theta$, an empirical plug-in solution converges to $\kappa^\star$ as the number of samples n grows to infinity. We let $\Theta \subseteq \mathbb{R}^d$ be our model space; our theorem allows $\Theta = \mathbb{R}^d$.

In the below result, $\widehat{\pi}_t(a_t \mid H_t)$ is a consistent estimator of $\pi_t(a_t \mid H_t)$ trained on a separate i.i.d. dataset $\mathcal{D}_n$; such estimators can be trained using sample splitting and standard supervised learning methods. Define the set S_ϵ of ϵ-approximate optimizers of the empirical plug-in problem

$$\min_{f(H_{t^\star})} \widehat{\mathbb{E}}_n \left[\frac{\mathbf{1}\left\{A_{t^\star} = a_{t^\star}\right\}}{\widehat{\pi}_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \widehat{\rho}_t - f(H_{t^\star}) \right) \right]$$

where $\widehat{\mathbb{E}}_n$ is the empirical distribution of the separate data, and

$$\widehat{\rho}_t := \frac{\bar{\pi}(A_t \mid \bar{H}_t(A_{1:t-1}))}{\widehat{\pi}_t(A_t \mid H_t(A_{1:t-1}))}.$$

We assume we observe i.i.d. episodes, and that episodes (unit) do not effect one another, so the observed cumulative reward is the potential outcome at the observed action sequence, $Y(A_{1:T})$.

See Section B.3.3 for the proof of the below result.

Theorem 4. *Let Assumptions 2, 5, 6 hold, and let there be a* $C \in (1, \infty)$ *s.t.* $\forall t$, $\rho_t \leq C$ *a.s.. Let* $\theta \mapsto f_\theta$ *be linear such that* $f_{\theta^\star}(\cdot) = \kappa^\star(\cdot, a_{t^\star})$ *for some unique* $\theta^\star \in \mathbb{R}^d$. *Let* $\mathbb{E}|Y(A_{1:T})|^4 < \infty$, *and* $\mathbb{E}[|f_\theta(H_{t^\star})|^4] < \infty$ *for all* $\theta \in \Theta$. *If for all* t, $\widehat{\pi}_t(\cdot|\cdot) \to \pi_t(\cdot|\cdot)$ *pointwise a.s.,* $\widehat{\rho}_t \leq 2C$, *and* $(2C)^{-1} \leq \widehat{\pi}_{t^\star}(a_{t^\star}|H_{t^\star}) \leq 1$ *a.s., then* $\liminf_{n\to\infty} dist(\theta^\star, S_{\varepsilon_n}) \xrightarrow{p} 0$ $\forall \varepsilon_n \downarrow 0$.

Hence, under the hypothesis of Theorem 4, a plug-in estimator is consistent for the lower bound (Equation 3.5).

3.6 Experiments

We now illustrate how our approach can generate meaningful certificates of robustness to unobserved confounding in realistic scenarios. We consider selecting evaluation policies using OPE methods, such as comparing the expected performance of a new policy to an existing policy. We empirically validate our method in sequential OPE problems where confounding is primarily an issue in only a single decision. Since counterfactual outcomes are only known in simulations, we focus on simulated healthcare examples motivated by a real OPE application: management of sepsis patients. Since this application uses discrete state space, we present another example with continuous-valued states: developmental interventions for autistic children, where using standard OPE methods result in wrong identification of the optimal policy. Our results demonstrate scalability of our loss minimization approach in both discrete and continuous settings, as well as short and medium horizons ($5 \sim 10$). We observe that beyond 10 time steps, overlap becomes a problem, and statistical estimation becomes challenging.

3.6.1 Managing sepsis for ICU patients

As outlined in Section 3.2, automated policies hold much promise in management of sepsis in ICU patients. However, ICU observational data about sepsis patients may often lack information about important confounders, such as important unrecorded comorbidities that affect a clinician's initial decision whether to administer antibiotics. In subsequent time steps, we assume the (highly-trained) clinical care team follows standard protocols based on vitals signs and lab measurements, and hence their subsequent decisions are unconfounded. On the sepsis simulator developed by Oberst and Sontag [2019], we illustrate how such confounders can bias OPE methods, and demonstrate that our worst-case approach can allow reliable selection of candidate policies under confounding.

We consider a scenario where automated policies have been proposed using existing medical knowledge, and we wish to evaluate their benefits relative to the current standard of care. We

evaluate three different policies, all of which only differ in their initial prescription of antibiotics, and otherwise act optimally. The first policy, *without antibiotics (WO)*, does not administer antibiotics initially, whereas the second policy, *with antibiotics (W)*, always administers antibiotics initially. For our last policy, we follow Oberst and Sontag [2019] and use the *optimal* policy learned by running policy iteration on this simulator—naturally this procedure does not have confounding. We stress that our first two policies are identical to the optimal policy after the initial time step. The true performance of the with antibiotics (W) and optimal policy is quite similar, and better than the without antibiotics (WO) policy (see Figure 3.1).

Simulator Oberst and Sontag [2019]'s simulator state space consists of a binary indicator for diabetes, and four vital signs {heart rate, blood pressure, oxygen concentration and glucose level} that take values in a subset of {very high, high, normal, low, very low}; size of the state space is $|\mathcal{S}_t| = 1440$. There are three binary treatment options {antibiotics, vasopressors, and mechanical ventilation} ($|\mathcal{A}_t| = 2^3$). In our experiments, simulation continues either until at most $T = 5$ (horizon) time steps, death (reward -1), or discharge (reward +1). Patients are discharged when all vital signs are in the normal range without treatment. Patients die if at least three vitals are out of the normal range. We refer the reader to `https://github.com/clinicalml/gumbel-max-scm` for details regarding the simulator.

The optimal policy Recall that we assume that the decisions are made near-optimally. To learn the optimal policy, we generate 2000 samples for each transition and constructed the transition matrix $P(s, a, s')$ and the reward matrix $R(s, a, s')$ of the MDP. Similar to Oberst and Sontag [2019] we used policy iteration to learn the optimal policy. We create a near-optimal (soft optimal) policy by having the policy take a random action with probability 0.05, and the optimal action with probability 0.95. The value function (for the optimal policy) was computed using value iteration. The horizon is $T = 5$ and the discount factor $\gamma = 0.99$, which results in soft optimal policy having an average value (over the possible distribution of state states) of 0.14.

Confounding We injected confounding in the first decision of this simulation by defining two different policies: *with antibiotics* and *without antibiotics*. *with antibiotics* which is identical to the soft optimal policy except that the probability mass of actions without antibiotics is moved to the corresponding action with antibiotics. For example, if the probability of the action a_1 =(antibiotics on, vasopressors off, ventilation on) in the soft optimal policy is p_1, and a_1' =(antibiotics off, vasopressors off, ventilation on) is p_1', then in the *with antibiotics* a_1 has probability $p_1 + p_1'$ and a_1' has probability zero in this new policy. The "without antibiotics" does the opposite: moves probability mass of actions with antibiotics to the corresponding action without antibiotics. In our confounding scenario, for healthy patients we administer antibiotics (i.e. follow the "with antibiotics") policy with a higher probability (w.p. $\frac{\sqrt{\Gamma}}{1+\sqrt{\Gamma}}$). For unhealthy patients, we administer antibiotics with a

lower probability (w.p. $\frac{1}{1+\sqrt{\Gamma}}$).

Concretely, to compute the transition from a state conditional on an action, we do inverse transform sampling: we generate a uniform random variable U_t on $[0, 1]$, and use this to index into the transition probability distribution for the next state, sorted by the states' value function and current reward. This coupling ensures that if U_t is large, then the next state will have a high value, and if U_t is small, then the next state will have a low value. The hidden variable U used for confounding in the first decision is $U = \sum_{t=1}^{T} U_t$, which serves as a surrogate for the health of patient, because the larger U is, the more likely the patient is to have improving state values. We choose a threshold u_0, and if $U > u_0$, the behavior policy follows the action with antibiotics, and if $U \leq u_0$, the behavior policy follows the action without antibiotics, thus introducing confounding.

After the first decision, the behaviour policy is a mixture of two policies: 85% the soft optimal policy and 15% of a sub-optimal policy that is similar to the soft optimal but the vasopressors action is flipped. For example, if probability of the action a_1 =(antibiotics on, vasopressors off, ventilation on) is p_1, and a_1' =(antibiotics on, vasopressors on, ventilation on) in p_1' in the soft optimal policy, then the sub-optimal has probability p_1' and p_1 for action a_1 and a_1', respectively.

Loss minimization Since the state and action space are discrete, we learn the tabular value $\kappa(s, a)$ for each state action pair separately to minimize the empirical loss. Additionally, in order to compute the upper bound of both ours and the naive method, we compute the negative of the lower bound on the negative of return (cost).

Behaviour Policy We estimate the behaviour policies from the data in two parts: the first time step and time steps $t = 2$ through $t = 5$. By the assumptions stated above, each of these policies depends only on the previous state, and we learn the tabular probability of each state action pair $\pi_t(a|s)$ separately.

Results We first consider when our approach happens to use the same confounding degree as what is present in the simulator, $\Gamma = \Gamma^\star$. Figure 3.1 plots the value of the three evaluation decision policies estimated using the data generated with $\Gamma^\star = 2.0$, which is a fairly small amount of confounding. Confounding leads standard OPE methods that assume sequential ignorability for the behavior policy to underestimate the peformance of the without antibiotics (WO) policy, and overestimate the performance of the with antibiotics (W) and optimal policies. This inflates the expected benefit of the W and optimal policies compared to the WO policy.

The naive approach (Equation 3.3) results in very wide estimated intervals over the potential policy performance, and therefore cannot be used to reliably infer the superiority of W and optimal policy over WO even when $\Gamma = 2.0$. On the other hand, our proposed method certifies the robustness of the benefit of immediately administering antibiotics; our lower bounds on the performance of the W and optimal policies are better than the upper bound on the performance under the WO policy.

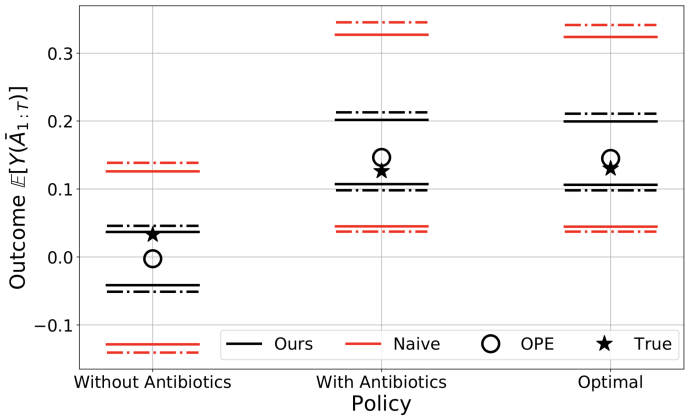


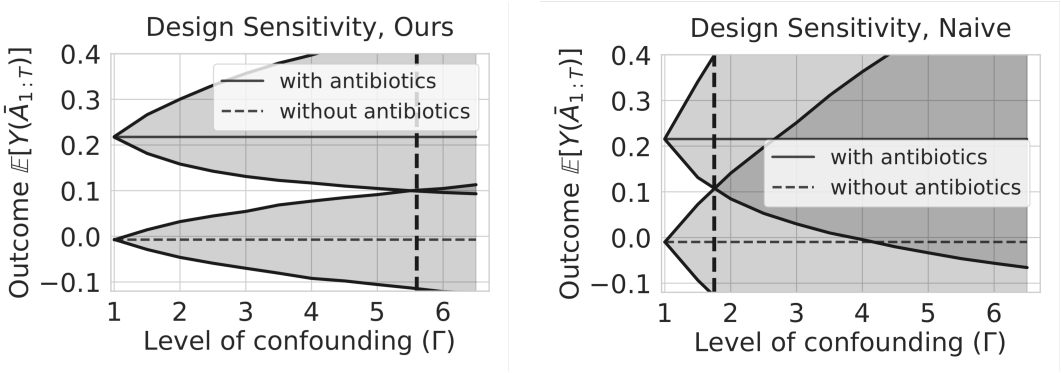
Figure 3.1: data was generated using $\Gamma^\star = 2.0$. Each policies' true value is shown with a start and a standard OPE estimate (ignoring confounding) is shown with an empty circle. Black lines show the estimated upper and lower bound on policy performance using our approach and red lines correspond to the naive approach, both using $\Gamma = 2.0$. Dashed lines represents 95% quantile.

We next consider a much larger amount of confounding, generating the observational data with $\Gamma^\star = 5.0$. To explore the design sensitivity of our method and our naïve lower bound approach, we use a range ofΓ values in our method. Figure 3.2 (a) and (b) shows that for our method, the lower bound on the performance of the W policy meets the upper bound on that of the WO policy at $\Gamma = 5.6$. In other words, our approach can reliably estimate that the W policy is better than the WO policy up to assuming an amount of confounding determined by $\Gamma = 5.6$ when the true $\Gamma^\star = 5.0$. In contrast, our proposed naive bound (Equation 3.3) has a a design sensitivity of $\Gamma = 1.75$, meaning the bounds quickly fails to be informative far below the true amount of data confounding. Our method allow concluding that the W policy is superior to the WO policy even when a substantial amount of unobserved confounding exists in the initial decision. We present another design sensitivity experiment, with $\Gamma^\star = 1.0$. Figure 3.3 (a,b) shows design sensitivity of our method (1.7) versus the naive method (1.23).

3.6.2 Communication interventions for minimally verbal children with autism

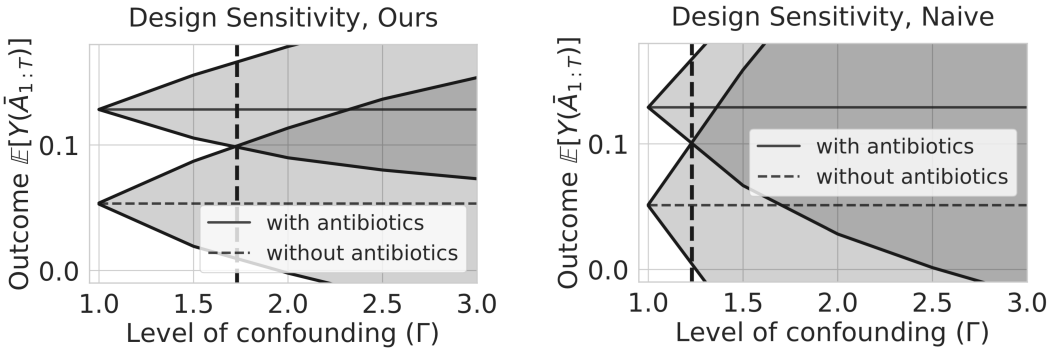
In this section, we consider another motivating scenario from healthcare, but one which naturally involves continuous variables to demonstrate that our approach is also able to compute reasonable lower bounds for such a case, while using function approximation.

Minimally verbal children represent 25-30% of children with autism, and often have poor prognosis in terms of social functioning [Rutter et al., 1967, Anderson et al., 2009]. We are interested in comparing non-adaptive versus adaptive approaches that aim to improve spoken communication, measured by the number of speech utterances. We introduce confounding using a simulator for



(a) Design sensitivity: ours (b) Design sensitivity: naive

Figure 3.2: Panels (a) and bc) plot design sensitivity. Data was generated with $\Gamma^\star = 5$. Estimated lower and upper bound of two policies (with and without antibiotics) under (a) our approach with design sensitivity 5.6, and (b) naive approach with design sensitivity 1.75.



(a) Design sensitivity: ours (b) Design sensitivity: naive

Figure 3.3: Sepsis simulator design sensitivity. Data generation process with level of confounding $\Gamma^\star = 1.0$. Estimated lower and upper bound of two policies (with and without antibiotics) under (a) our approach with sensitivity 1.7 (b) naive approach with sensitivity 1.23.

autistic children developed by Lu et al. [2016], which models the data from a (real) sequential, multiple assignment, randomized trial (SMART) [Kasari et al., 2014]. Despite their randomized trial, Kasari et al. [2014] note that very few randomized trials of these interventions exist, and the number of individuals in these trials tends to be small. It is therefore reasonable to think that in similar settings it would be beneficial to use existing off-policy data to evaluate new intervention protocols.

In the simulator there are two developmental interventions (actions): behavioral language interventions (BLI) delivered by a therapist, and an augmented/alternative communication (AAC) approach implemented with a speech generation device. There are two decision points in the data generation process: week 0 and week 12. Number of speech utterances are measured at week 0, 12, 24 and 36: note that the action / intervention applied at week 12 persists from week 12 to the end

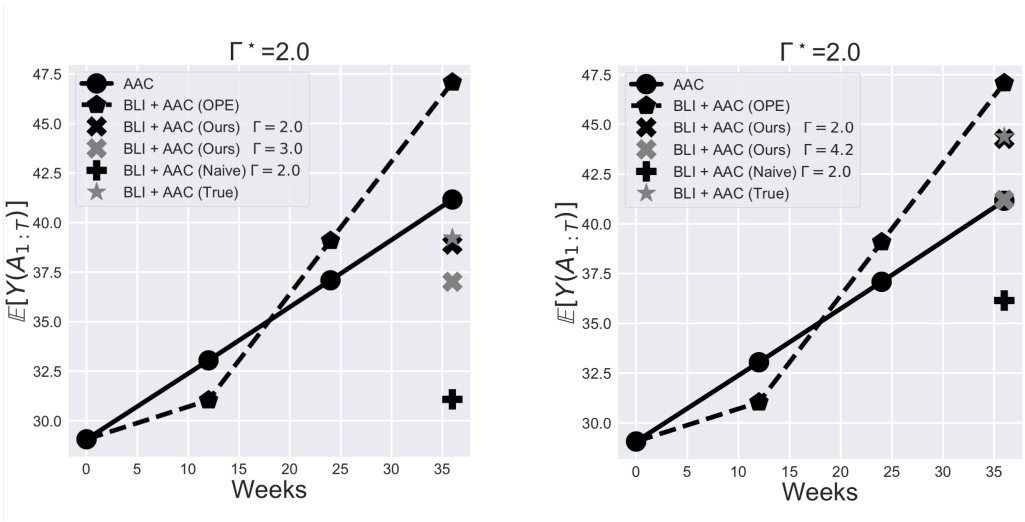
of the process, which means this is a 2 time step decision problem. Here the outcome is modeled as a continuous variable representing the average number of speech utterances for a given patient.

We consider a scenario where participants were recruited and randomly assigned to the two treatment options initially (i.e., $A_1 \sim \text{Unif}(\{\text{BLI}, \text{AAC}\})$), and a recourse action is taken after a follow-up visit after 12 weeks. Depending on the progress of patients at Week 12, the clinician decides whether to switch to AAC devices for children who started with BLI. Since this intervention requires a specialized device—whose supply is limited—it is likely that the clinicians assign AAC devices for whom it has a higher chance of being effective. Such subjective assessments are likely based on the their interaction with patients that contain partial, noisy information about the final outcome, which are often not recorded properly. Therefore, while there is confounding in the second decision ($t^\star = 2$), its influence may be appropriately bounded (i.e., Assumption 5 is plausible). To simulate confounding, we expand the simulator to create variables that partially influences the effectiveness of switching from BLI to AAC, and use knowledge of this to alter the behavior policy decisions at Week 12. The resulting confounding satisfies our model of bounded confounding (Assumption 5).

In our evaluations, we compare an adaptive policy (BLI + AAC) that starts with BLI, and augments BLI with AAC at week 12 if the patient is a slow responder, against a non-adaptive policy that uses AAC through the whole treatment. We simulate two different settings where the effect of switching to the AAC treatment varies; our simulation parameters are within the suggested range of Lu et al. [2016]'s recommendations based on the SMART trial data. We note that OPE estimates for the non-adaptive policy (AAC) is unbiased since observations for this outcome are unconfounded. Our loss minimization for computing the lower bound using $\kappa(a_{t^\star}, H_{t^\star})$ is done using a 4 layers neural network with Relu activations, we use backpropogation with AdamOptimizer and weighted squared loss given in Theorem 3. We use logistic regression to estimate the behavior policy, note that this is the marginalized behavior policy since the latent confounder is unobserved.

Details of experiments In the autism experiments, our data generation process (simulator) is adopted from Lu et al. [2016, Appendix B]. Each individual has a set of covariates X, consisting of six mean-centered features: {`age`, `gender`, `indicator of African American`, `indicator of Caucasian`, `indicator of Hispanic`, `indicator of Asian`}. The Autism SMART trial Kasari et al. [2014] simulator specifies a set of 300 individuals: to obtain a sample size N, we sample with replacement from this set. For details on the simulator, we refer to Appendix B of Lu et al. [2016]. At the first timestep there are two actions available $A_1 \in \{-1, 1\}$, where $A_1 = 1$ denote BLI, and $A_1 = -1$ denote AAC. At the second timestep there are three actions $A_2 \in \{-1, 0, 1\}$, where $A_2 = 1$ denote assigning intensified BLI to slow responders, $A_2 = -1$ denote assigning AAC to slow responder and $A_2 = 0$ denote continuing with the same action for fast responders.

Confounding The original simulator did not have confounding. We now describe how we introduce confounding in this setting. Lu et al. [2016, Appendix B] specifies the effect of the second



(a) Case I (b) Case II

Figure 3.4: Autism simulation. Outcome of two different policies, confounded adaptive policy (BLI+AAC) and un-confounded non-adaptive policy (AAI). Data generation process with the level of confounding $\Gamma^\star = 2.0$.

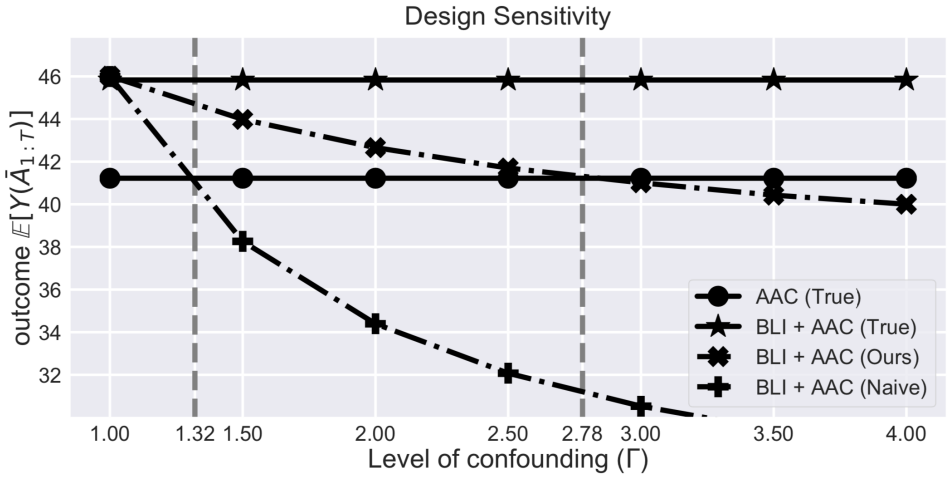


Figure 3.5: Autism simulation design sensitivity. Data generation process with the level of confounding $\Gamma^\star = 1.0$. True value of adaptive (BLI+AAC) and non-adaptive (AAC) policies along with estimated lower bound on outcome using our and naive approach with sensitivity 2.28 and naive approach with sensitivity 1.28

action (whether to augment BLI with AAC) on the reward outcome Y as follows:

$$Y = \eta_{31}^T X + \eta_{22} Y_0 + \eta_{33}^T A_1 + \eta_{34} Y_{12} - 2\theta(1-R)(A_1+1)A_2 + \epsilon.$$

A_1 is either -1 or 1. Therefore the final term (outside of the noise ϵ) is non-zero only when $A_1 = 1$, and we can interpret θ as the effect size of the adaptive policy (which always takes $A_1 = 1$); for exact definition of the effect size refer to Lu et al. [2016]. For those more familiar with the RL literature, it is related to the advantage function. In the original paper, Figure 7 in Lu et al. [2016] were generated using 4 different values of θ. The parameters used in these simulations are in the range reported by Lu et al. [2016].

We introduce confounding by varying θ (thereby impacting the potential outcome) and then altering the behavioral treatment decisions according to the knowledge of that θ. More precisely, given a θ_0, for each individual, we randomly set $\theta_0 + \sigma_\theta$ or $\theta_0 - \sigma_\theta$. The second action is 1 with probability $\frac{\sqrt{\Gamma}}{1+\sqrt{\Gamma}}$ if $\theta \geq \theta_0$ and 1 with probability $\frac{1}{1+\sqrt{\Gamma}}$ if $\theta \leq \theta_0$. In our experiments, we take $\sigma_\theta = 5$.

Loss minimization To estimate $\kappa(H_{t^\star}; a_{t^\star})$ in the loss minimization problem, we used a neural network with 3 hidden layers of size {`128, 128, 128, 64`} with `Relu` activations, followed by a single linear output layer. We initialize the layers with Xavier initialization and used the Adam optimizer with learning rate 10^{-3}. The input H_t is 10-dimensional consisting of 6 covariates, indicator of slow responder, initial action A_1, number of speech utterances after the initial action, and an interaction term between A_1 and the slow responder indicator.

Behaviour Policy We use logistic regression to estimate the behaviour policy from the observed data: note that this is not the true behavior policy, because that depends on the (latent) confounding. Different models were fit for the first and second time steps. For the first timestep the learned model is $\pi_1(A_1|H_1)$, where H_1 contains the observed X (6 covariates), and $A_1 \in \{-1, 1\}$. For the estimated behavior policy in the second timestep $\pi_2(A_2|H_2)$, H_2 includes X (6 covariates), the action A_1, indicator of slow responder, the interaction term between A_1 and the indicator, and the number of speech utterances after the initial action.

Results In Case I, we define the parameters such that the adaptive policy (BLI+AAC) is worse the non-adaptive policy (AAC) (lower true outcome / performance). As shown in Figure 3.4 (a), standard OPE approach overestimates the outcome of the adaptive policy even given a mild level of confounding $\Gamma^\star = 2.0$, and would incorrectly suggest the BLI+AAC policy outperforms the AAC policy. On the other hand, our lower bounds on the adaptive policy computed using $\Gamma = 2$ (recall the true confounding amount is unknown to our approach) suggest the OPE estimates may be biased enough to affect conclusions; the observed advantages of the adaptive policy may be attributed solely to unobserved confounding, even under reasonable values of confounding ($\Gamma = 2$).

In Case II, we change the parameters so that the BLI+AAC policy is better than the AAC policy, and again use a true amount of confounding of $\Gamma^\star = 2.0$ in the data generation process. Standard OPE estimates again overestimate the outcome for the BLI+AAC policy (Figure 3.4(b)). The naive

lower bound results in a conservative lower bound that would again indicate no conclusions can be drawn about the relative performance of BLI+AAC versus AAC. However, our method can certify the superiority of the BLI+AAC policy when the level of confounding used in the computation of the lower bounds is up to $\Gamma = 4.2$, thereby providing a case where our approach can provide useful certificates of benefit of a new decision policy under non-trivial levels of confounding.

Figure 3.5 plots the design sensitivity of our method against the naive approach (Equation 3.3), when there is in fact no confounding in the data generation process ($\Gamma^\star = 1$). Compared to the naive approach (design sensitivity is $\Gamma = 1.32$), our method allows certifying robustness of the finding—that the adaptive policy is advantageous—up to realistic levels of confounding (design sensitivity is $\Gamma = 2.78$).

3.7 Related Work

The dynamic treatment regime literature [Robins, 1986, Murphy, 2003] addressed many early questions around using observational data for sequential decision making, and developed a rich set of methods adapted for epidemiological questions. The reinforcement learning (RL) community is increasingly interested in developing theory and methods for the related problem of batch RL across a broad set of applications, because of new models and data availability (see e.g. [Thomas et al., 2019, Liu et al., 2018b, Le et al., 2019, Thomas et al., 2015, Komorowski et al., 2018, Hanna et al., 2017, Gottesman et al., 2019b]).

The majority of OPE methods for batch reinforcement learning rely on sequential ignorability (though often unstated). There is an extensive body of work for off-policy policy evaluation and optimization under this assumption, including doubly robust methods [Jiang and Li, 2015, Thomas and Brunskill, 2016] and recent work that provides semiparametric efficiency bounds [Kallus and Uehara, 2020]; often the behavior policy is assumed to be known. Notably, Liu et al. [2018b] highlights how estimation error in the behavioral policy can bias value estimates, and Nie et al. [2019], Hanna et al. [2019] provides OPE estimators based on an estimator of the behavior policy. When sequential ignorability doesn't hold, the expected cumulative rewards under an evaluation policy cannot be identified from observable data. All of the above estimators are biased in the presence of unobserved confounding, since neither the outcome model nor the importance sampling weights can correct for the effect of the unobserved confounder.

The do-calculus and its sequential backdoor criterion on the associated directed acyclic graph [Pearl, 2009] gives identification results for OPE. Like sequential ignorability, this precludes the existence of unobserved confounding. Hence, methods that assume the sequential backdoor criterion will be biased in their presence.

We study the effects of unobserved confounding on OPE in sequential decision making problems, deriving bounds on the performance of the evaluation policy when sequential ignorability is relaxed.

For problems where only one decision is made, a variety of methods developed in the econometrics, statistics, and epidemiology literature estimate bounds on treatment effects and expected rewards. Manski [1990] developed bounds that only assume bounded rewards, though they are too conservative to identify whether one action is superior to another. Then, Manski [1990] and other works posit models that bound the effect of unobserved confounding on the outcome [Robins et al., 2000, Brumback et al., 2004], or—like ours—on the actions taken by the behavior policy [Cornfield et al., 1959, Rosenbaum and Rubin, 1983, Imbens, 2003]. Recent work studied approaches that can apply to heterogeneous treatment effects [Yadlowsky et al., 2018, Kallus et al., 2018], policy evaluation [Jung et al., 2018], and policy optimization [Kallus and Zhou, 2018].

Tennenholtz et al. [2020] studied OPE for partially observable Markov decision processes (POMDPs) and developed an identification strategy based on the independence structure of POMDP, similar to single decision work of Miao et al. [2018]. We do not assume the existence of such variables or independence structures and seek to develop a lower bound on OPE. This also distinguishes our work from prior work which focuses on an algorithmic, scalable approach for when a single, time-invariant confounder is present, and which does not seek to present bounds on the OPE [Lu et al., 2018].

In sequential settings, Zhang and Bareinboim [2019] derived partial identification bounds on policy performance with limited restrictions on the influence of the unobserved confounder on observed decisions, much like the single decision work of Manski [1990], which they use to guide online RL algorithms. Unfortunately, these bounds can be too conservative to guide selection of policies. Robins et al. [2000], Robins [2004], Brumback et al. [2004] instead posit a model for how confounding in each time step affects the outcome of interest and derive bounds under this model. Their work is motivated by potential confounding in the effects of dynamic treatment regimes for HIV therapy on CD4 counts in HIV-positive men. Our work is complementary to these in that we instead assume limited influence of the unobserved confounder on the behavior policy's actions.

Yadlowsky et al. [2018] takes a similar approach as ours to bound the effect of confounding on treatment effects when there is only one action taken. Our approach allows for comparing sequences of actions derived according to an evaluation policy, by adjusting for the way actions in all time steps depend on the current states and history, and effect future states and rewards. One notable challenge that only occurs in sequential problems is adjusting for actions that occur after the confounded decision at time $t^\star$; these actions depend on the confounded decision through the history generated. A natural approach is to individually bound the potential outcomes $\mathbb{E}[Y(\bar{A}_{1:t^\star-1}, a_{t^\star:T})]$ for all $a_{t^\star:T}$, where each bound is given by a loss minimization problem. Under this approach—which is analogous to that of Yadlowsky et al. [2018]—computing a lower bound to $\mathbb{E}[Y(\bar{A}_{1:T})]$ requires $\prod_{t=t^\star}^{T} |\mathcal{A}_t|$ loss minimization problems, making it statistically and computationally intractable when $t^\star$ is small (e.g. $t^\star = 1$ in our sepsis example). Instead, we consider averaged outcomes $\mathbb{E}[Y(\bar{A}_{1:t^\star-1}, a_{t^\star}, \bar{A}_{t^\star+1:T})]$ in Theorem 3, which allows us to obtain a lower bound on $\mathbb{E}[Y(\bar{A}_{1:T})]$ by solving $|\mathcal{A}_{t^\star}|$ loss minimization problems.

3.8 Discussion

We proposed methods for analyzing the sensitivity of OPE methods to unobserved confounding in sequential decision making problems. We demonstrated how our approach can certify robustness of OPE in some settings, or raise concerns about its validity based on sensitivity to unobserved confounding. Our loss minimization method allows computing worst-case bounds over our bounded unobserved confounding model, while adjusting for observed features via importance sampling.

As a consequence, our estimators face the same challenges that standard importance-sampling-based OPE methods face: high variance when there is little overlap between the evaluation and behavior policy. In our experiments, importance sampling was effective since we ensured that there was sufficient overlap and focused on shorter horizons. In other settings, lack of overlap poses fundamental difficulties in off-policy evaluation, beyond issues with confounding, as others have also noted [Gottesman et al., 2019a]. While stationary importance sampling (SIS) can reduce variance, rewards under stationary distributions (should they exist) are not appropriate for the problems studied in this paper; SIS [Hallak and Mannor, 2017, Liu et al., 2018a, Xie et al., 2019] nevertheless still suffers high variance when there is a lack of overlap. Fujimoto et al. [2019], Kumar et al. [2019] suggest some promising algorithmic approaches for only considering policies with sufficient overlap: while more work is needed, policies generated by these approaches would be more amenable to OPE, and should improve the statistical properties of our method.

It is natural to consider extending our single-decision confounding model to settings where a handful of decisions (say 2-5) are affected by unobserved confounding. Worst-case bounds on $\mathbb{E}[Y(\bar{A}_{1:T})]$ under such extensions require solving optimization problems involving products of likelihood ratios defined over different confounded time periods. Since these problems are nonconvex, they require new approaches than our developments which heavily depends on convex duality.

Chapter 4

Identification of Subgroups With Similar Benefits in Off-Policy Policy Evaluation

Off-policy policy evaluation methods for sequential decision making can be used to help identify if a proposed decision policy is better than a current baseline policy before deployment. However, a new decision policy may be better than a baseline policy for some individuals but not others. This has motivated a push towards personalization and accurate per-state estimates of heterogeneous treatment effects (HTEs). Given the limited data present in many important applications, individual predictions can come at a cost to accuracy and confidence in such predictions. We develop a method to balance the need for personalization with confident predictions by identifying subgroups where it is possible to confidently estimate the expected difference in a new decision policy relative to a baseline. We propose a novel loss function that accounts for the uncertainty during the subgroup partitioning phase. In experiments, we show that our method can be used to form accurate predictions of HTEs where other methods struggle.

4.1 Introduction

Recent advances in technology and regulations around them have enabled the collection of an unprecedented amount of data of past decisions and outcomes in different domains such as health care, recommendation systems, and education. This offers a unique opportunity to learn better decision-making policies using observational data. Off-policy policy evaluation (OPE) is concerned with estimating the value of a proposed policy (*evaluation policy*) using the data collected under a different policy (*behavior policy*). Estimating the value of an evaluation policy before deployment is

essential, especially when interacting with the environment is expensive, risky, or unethical, such as in health care [Gottesman et al., 2019a]. Fortunately, the reinforcement learning (RL) community has developed different methods and theories focused on OPE e.g. [Jiang and Li, 2015, Thomas and Brunskill, 2016, Kallus and Uehara, 2020].

OPE has been used extensively in the literature to demonstrate the superiority of a proposed evaluation policy relative to the baseline (behaviour) policy e.g. [Komorowski et al., 2018]; however, the evaluation policy may be better than the behaviour policy for some individuals but not others. Hence, only looking at the estimated value of the evaluation policy before deployment, may be misleading. In the non-sequential setting, a growing literature has focused on personalization and estimation of heterogeneous treatment effect (HTE), the individual-level differences in potential outcomes under the proposed evaluation policy versus the behaviour policy [Athey et al., 2019]. These methods often aim to learn a parametric or non-parametric function to predict HTE for each individual [Athey and Imbens, 2016, Nie and Wager, 2017].

However, due to the limited availability of data and long horizon, the goal of personalization for each individual can be unrealistic, especially in sequential settings. In practice, individual predictions of treatment effect in these settings can be inaccurate and highly uncertain, providing no actionable information. In this paper, we aim to provide actionable information to domain experts. Specifically, we ask " *What subgroups of individuals can we confidently predict that will be significantly benefited or harmed by adopting the evaluation policy?*". Asking this question instead of "What is the treatment effect for each individual?" allows us to group individuals that have similar treatment effects together, pool the data, and make predictions that are both more accurate and confident.

We build upon a recursive partitioning algorithm proposed by Athey and Imbens [2016] for the non-sequential setting which, similar to classification and regression trees (CART) [Breiman et al., 1984], greedily minimize a loss function. Estimation of the loss function with this method can be too noisy and often results in over-splitting, yielding too many subgroups, and inaccurate or uncertain prediction, when applied to sequential settings. One key challenge in estimating the loss function is estimating the variance of the estimator. Sample variance estimates are unstable and, as the horizon increases, become even more unreliable.

To achieve our goal of identifying subgroups with significant treatment effects, we make two contributions. First, we use a proxy of the variance of treatment effect estimator that can be efficiently computed and is stable given the available amount of data. This mitigates the problem of over-splitting. Second, we incorporate a regularization term that incentivizes recursive partitioning to find subgroups with significant treatment effects. Our interest to find subgroups with significant benefit or harm as a result of adopting the evaluation policy requires domain expert's information to characterize what a significant benefit or harm means. For example, a clinician may consider an increase of 10% in survival rate significant, so only subgroups with a confident prediction of 10% decrease or increases in survival rate will provide actionable information. The regularization term

allows us to incorporate such information into the loss function. Combining these two additions, we propose a new loss function that can be efficiently computed and hence be used in recursive partitioning to achieve our goal.

On a simulated example of sepsis management [Oberst and Sontag, 2019] we show how our proposed method can be used to find subgroups with significant treatment effect, providing more accurate and confident predictions. Additionally, we apply our method to the sepsis cohort of MIMIC III dataset [Johnson et al., 2016], illustrating how our method may be useful in identifying subgroups in which a new decision policy may be beneficial or harmful relative to the standard approach. Our proposed approach can be used as a tool to help inform policy deployment and could help guiding hypothesis formation for future experimentation.

4.2 Related Work

Estimating the value of a new decision policy arise in many different applications, such as personalized medicine [Obermeyer and Emanuel, 2016], bandits [Dudík et al., 2011] and sequential decision makings [Thomas and Brunskill, 2016]. The RL community has developed different methods and theories for off-policy policy evaluation (OPE) in sequential setting. These methods mostly fall into different categories: importance sampling [Precup, 2000], model based and doubly robust methods [Dudík et al., 2011, Thomas and Brunskill, 2016, Jiang and Li, 2015]. These methods can be used along with our algorithm to estimate group treatment effect for a particular group; however, these methods do not offer a way to perform partitioning.

In non-sequential settings, a growing number of literature seek to estimate heterogeneous treatment effect (HTE) using different approaches. Imai and Ratkovic [2014] uses the LASSO to estimates the effect of treatments, Shalit et al. [2017] uses neural networks and offers generalization bound for individual treatment effect (ITE). Nie and Wager [2017] proposed two step estimation procedure using double machine learning and orthogonal moments [Chernozhukov et al., 2018] that can be applied on observational data to infer HTEs, and recently, Lee et al. [2020] suggests a robust partitioning algorithm by inducing homogeneity in groups. However, these methods were developed for non-sequential settings and naïvely applying them to sequential setting will result in predictions with low accuracy and high uncertainty. Our work draws close parallel to methods using recursive partitioning to estimate HTEs [Athey and Imbens, 2016, Athey et al., 2019], but those works suffer from over-splitting the feature space in sequential setting due to noisy estimation of the loss function. We propose a different loss function that can be better estimated given the lack of data and incorporate domain expert knowledge.

4.3 Setting and Background

We consider episodic stochastic decision processes with a finite action space $\mathcal{A}$, continuous state space $\mathcal{X} \in \mathbb{R}^M$, reward function $R : \mathcal{X} \times \mathcal{A} \to [0, R_{max}]$ and discount factor $\gamma \in [0, 1]$. A policy π maps the state space to a probability distribution over the action space, and we assume each episode lasts at most H steps.

A set of trajectories $\mathcal{T} = \{\tau_1, \ldots, \tau_N\}$ is provided. Each trajectory τ_i consists of a state x_t, action a_t and observed reward r_t at step t, $\tau_i = \{x_0^i, a_0^i, r_0^i, \ldots, x_H^i\}$. Actions are generated by following a known behaviour policy π_b, $a_t \sim \pi_b(s_t)$. We denote the evaluation policy by π_e.

Rényi divergence For $\alpha \geq 0$ the Rényi divergence for two distribution P and Q is defined by [Cortes et al., 2010]

$$D_\alpha(P||Q) = \frac{1}{\alpha - 1} \log_2 \sum_x Q(x) \left(\frac{P(x)}{Q(x)} \right)^{\alpha - 1}$$

We denote the exponential in base 2 by

$$d_\alpha(P||Q) = 2^{D_\alpha(P||Q)}$$

The importance weights of two distributions P and Q can be defined as $w(x) = P(x)/Q(x)$. Moments of the importance weight can be represented by the Rényi divergence, that is

$$\begin{aligned} \mathbb{E}_{x \sim Q}[w(x)^2] &= d_2(P||Q) \\ \mathbb{V}_{x \sim Q}[w(x)] &= d_2(P||Q) - 1 \end{aligned}$$

Effective Sample Size The effective sample size (ESS) [Kong, 1992] is often used for diagnosis of IS estimators, ESS is defined as

$$\begin{aligned} ESS(P||Q) &= \frac{N}{1 + \mathbb{V}_{x \sim Q}[w(x)]} \\ &= \frac{N}{d_2(P||Q)} \end{aligned}$$

where N is the number of samples drawn to estimate the importance weights. The following display is a common estimator of this quantity based on the importance weights [Owen, 2013].

$$\widehat{ESS}(P||Q) = \frac{\left(\sum_{i=1}^N w_i \right)^2}{\sum_{i=1}^N w_i^2}$$

4.4 Framework for Subgroup Identification

Our focus is to robustly quantify the expected benefit or cost of switching from a behavior policy to a proposed evaluation policy on subsets of the population. To do so it is helpful to extend the standard notion of the treatment effect to the (sequential decision) policy treatment effect. We define the individual treatment effect $t(x; \pi_e, \pi_b)$ for a possible initial state x as

$$t(x; \pi_e, \pi_b) = \mathbb{E}_{\pi_e}\left[\sum_{t=0}^{H} \gamma^t r_t | x_0 = x\right] - \mathbb{E}_{\pi_b}\left[\sum_{t=0}^{H} \gamma^t r_t | x_0 = x\right]. \tag{4.1}$$

Before we introduce our definition of group treatment effects, we first define a partitioning over the state space by $L = \{l_1, \ldots, l_M\} \in \Pi$, such that $\bigcup_{i=1}^{M} l_i = \mathcal{X}$ and $\forall i, j : \quad l_i \bigcap l_j = \emptyset$. Define the partition function $l(x; L) = l_i$ such that $x \in l_i$. Given a partitioning L, partition-value function for a policy π can be defined as:

$$v(x; L, \pi) = \mathop{\mathbb{E}}_{x' \sim \mathcal{X} a \sim \pi(.|x')}\left[\sum_{t=0}^{H} \gamma^t r_t | x_0 = x', x' \in l(x; L)\right] \tag{4.2}$$

Using this function we can define group treatment effect, similar to the individual treatment effect as,

$$T(x; L, \pi_b, \pi_e) = v(x; L, \pi_e) - v(x; L, \pi_b) \tag{4.3}$$

note that group treatment effect is constant within every l_i, and we refer to each l_i as a group. With little abuse of notation we denote the individual treatment effect by $t(x)$ and group treatment effect by $T(x; L)$ and interchangeably use group and subgroup.

4.4.1 Group treatment effect estimator

Given a partition L, a set of trajectories $\mathcal{T}$, the behaviour policy π_b and an evaluation policy π_e the following estimator defines the group treatment effect estimator for an initial state x over a dataset $\mathcal{D} = \{(x_0, \rho_0, g_0), \ldots, (x_N, \rho_N, g_N)\}$,

$$\hat{T}(x; L) = \frac{1}{|\{x_i | x_i \in l(x; L)\}|} \sum_{i | x_i \in l(x;L)} (\rho_i g_i - g_i) \tag{4.4}$$

Where, $x_i = x_0^i$ is the initial state of a trajectory τ_i, g_i is the discounted return $g_i = \sum_{t=0}^{H} \gamma^t r_t^i$ and ρ_i is the importance sampling ratio

$$\rho_i = \prod_{t=0}^{H} \frac{\pi_e(a_t^i | x_t^i)}{\pi_b(a_t^i | x_t^i)}$$

It is straightforward to show that $\hat{T}(x;L)$ is an unbiased estimator of $T(x;L)$ in every group.

The aim of this chapter is to find groups L (subgroups) that we can accurately predict the group treatment effect for them. Following much of the literature [Athey and Imbens, 2016, Thomas and Brunskill, 2016] we focus on the MSE criteria to rank different estimators defined by different partitioning; however, as explained later, we modify this loss in multiple ways to account for our goal.

$$MSE(\hat{T};L) = \mathop{\mathbb{E}}_{x\sim\mathcal{X}}\left[\left(t(x) - \hat{T}(x;L)\right)^2\right]$$

Note that MSE loss is infeasible to compute, as we do not observe individual treatment effect $t(x)$. We form the adjusted MSE (AMSE) as

$$AMSE(\hat{T};L) = \mathop{\mathbb{E}}_{x\sim\mathcal{X}}\left[\left(t(x) - \hat{T}(x;L)\right)^2 - t(x)^2\right] \tag{4.5}$$

Adjusted MSE and MSE impose the same ranking among different partitioning as $\mathop{\mathbb{E}}_{x\sim\mathcal{X}}[t(x)^2]$ is independent from the partitioning. Note that adjusted MSE, similar to MSE cannot be computed; however, we now show that it is equivalent to an expectation over quantities that can be estimated from data,

Theorem 5. *For a given partition $L \in \Pi$, let $T(x;L)$ be the group treatment effect defined in equation 4.3, $t(x)$ be the individual treatment effect as defined in equation 4.1 and $\hat{T}(x;L)$ an unbiased estimator of $T(x;L)$. The following equality holds for the adjusted MSE.*

$$AMSE(\hat{T};L) = -\mathbb{E}_{x\sim\mathcal{X}}\left[\hat{T}^2(x;L)\right] + 2\ \mathbb{E}_{x\sim\mathcal{X}}\left[\mathbb{V}\left[\hat{T}(x;L)\right]\right]$$

Where $\mathbb{V}[\hat{T}(x;L)]$ is the variance of the estimator $\hat{T}(x;L)$.

Proof. We start by decomposing the adjusted MSE by adding and subtracting $T(x;L)$,

$$\begin{aligned}
AMSE(\hat{T};L) &= -\mathbb{E}_{x\sim\mathcal{X}}\left[\left(t(x) - \hat{T}(x;L)\right)^2 - t(x)^2\right] \\
&= \mathbb{E}_{x\sim\mathcal{X}}\Big[\underbrace{(t(x) - T(x;L))^2 - t(x)^2}_{(i)} \\
&\quad + \underbrace{(T(x;L) - \hat{T}(x;L))^2}_{ii} \\
&\quad + \underbrace{2(t(x) - T(x;L))(T(x;L) - \hat{T}(x;L))}_{(iii)}\Big]
\end{aligned}$$

Now we look at each part separately, for part (i),

$$\mathbb{E}_{x\sim\mathcal{X}}\left[(t(x) - T(x;L))^2 - t(x)^2\right] \tag{4.6}$$

$$= \mathbb{E}_{x\sim\mathcal{X}}\left[T(x;L)^2 - 2t(x)T(x;L)\right] \tag{4.7}$$

$$\overset{①}{=} \sum_{l_i\in L} P(l_i)T(x;l_i)^2 - 2\sum_{l_i\in L} P(l_i)T(x;l_i) \tag{4.8}$$

$$= -\sum_{l_i\in L} P(l_i)T(x;l_i)^2 \tag{4.9}$$

$$\overset{②}{=} -\mathbb{E}_{x\sim\mathbb{X}}[T(x;L)^2] \tag{4.10}$$

Where in ① we expand the expectation over each group of the partition $L = \{l_1, \ldots, l_M\}$. Note that, $T(x;l_i) = T(x;L)$ such that $x \in l_i$, by definition, $T(x;l_i)$ is constant for all $x \in l_i$. In ② we used the fact that $\mathbb{E}_{x\in l_i}[t(x)] = T(x;l_i)$.

Now, consider the variance of $\hat{T}(x;l_i)$ for group $l_i \in L$,

$$\begin{aligned}\mathbb{V}\left[\hat{T}(x;l)\right] &= \mathop{\mathbb{E}}_{x\in l_i}[\hat{T}^2(x;l_i)] - \left[\mathop{\mathbb{E}}_{x\in l_i}\hat{T}(x;l_i)\right]^2 \\ &= \mathop{\mathbb{E}}_{x\in l_i}[\hat{T}^2(x;l_i)] - T(x;l_i)^2 \end{aligned} \tag{4.11}$$

Which follows by $\hat{T}(x;l_i)$ being an unbiased estimator of $T(x;l_i)$. Taking the expectation over the feature space and substituting equation 4.11 into 4.6,

$$\begin{aligned}\mathbb{E}_{x\sim\mathcal{X}}[(t(x) - T(x;L))^2 - t(x)^2] &= -\sum_{l_i\in L} P(l_i)T(x;l_i)^2 \\ &= \sum_{l_i\in L} P(l_i)\left[\mathbb{V}\left[\hat{T}(x;l_i)\right] - \mathbb{E}_{x\sim l_i}\left[\hat{T}^2(x;l_i)\right]\right] \\ &= \mathbb{E}_{x\sim\mathbb{X}}\left[\mathbb{V}\left[\hat{T}(x;l)\right]\right] - \mathbb{E}_{x\sim\mathbb{X}}\left[\hat{T}^2(x;l)\right]\end{aligned}$$

Now we consider the second part part (ii),

$$\begin{aligned}&\mathbb{E}_{x\sim\mathbb{X}}\left[\left(T(x;L) - \hat{T}(x;L)\right)^2\right] \\ &\quad= \sum_{l_i} P(l_i)\mathbb{E}_{x\in l_i}\left[\left(T(x;l_i) - \hat{T}(x;l_i)\right)^2\right] \\ &\quad= \sum_{l_i} P(l_i)\mathbb{E}_{x\in l_i}\left[\left(\mathbb{E}_{x\in l_i}[\hat{T}(x;l_i)] - \hat{T}(x;l_i)\right)^2\right] \\ &\quad= \sum_{l_i} P(l_i)\mathbb{V}\left[\hat{T}(x;l_i)\right] \\ &\quad= \mathbb{E}_{x\sim\mathbb{X}}\left[\mathbb{V}\left[\hat{T}(x;L)\right]\right]\end{aligned}$$

Where the third line follows by $\hat{T}(x;l_i)$ being an unbiased estimator of $T(x;l_i)$.

Looking at the last term (iii),

$$\begin{aligned}
\mathbb{E}_{x\sim\mathcal{X}}\left[\left(t(x)\hat{T}(x;L)\right)\right] &= \sum_{l_i\in L} P(l_i)\mathbb{E}_{x\in l_i}\left[t(x)\hat{T}(x;l_i)\right] \\
&= \sum_{l_i\in L} P(l_i)\hat{T}(x;l_i)\mathbb{E}_{x\in l_i}\left[t(x)\right] \\
&= \sum_{l_i\in L} P(l_i)\hat{T}(x;l_i)T(x;l_i) \\
&= \mathbb{E}_{x\in\mathcal{X}}\left[\hat{T}(x;l_i)T(x;l_i)\right]
\end{aligned}$$

Which implies $\mathbb{E}_{x\sim\mathcal{X}}\left[(t(x)-T(x;L))\hat{T}(x;L)\right]=0$. As a result,

$$\begin{aligned}
(iii) &= 2\mathbb{E}_{x\in\mathcal{X}}\left[(t(x)-T(x;L))\left(T(x;L)-\hat{T}(x;L)\right)\right] \\
&= 2\mathbb{E}_{x\in\mathcal{X}}\left[(t(x)-T(x;L))\left(T(x;L)\right)\right] \\
&\quad - 2\mathbb{E}_{x\in\mathcal{X}}\left[(t(x)-T(x;L))\left(\hat{T}(x;L)\right)\right] = 0
\end{aligned}$$

Putting the results together, proves the result. □

The results of theorem 5 suggest an estimatable quantity that can be used to select among different potential partitions. More precisely, given a dataset $\mathcal{D}$ the empirical adjusted MSE can be written as,

$$EMSE(\hat{T};L) = -\frac{1}{N}\sum_{i=1}^{N}\hat{T}^2(x_i;L) + \frac{2}{N}\sum_{i=1}^{N}\mathbb{V}\left[\hat{T}(x_i;L)\right] \tag{4.12}$$

Where $\mathbb{V}[\hat{T}(x_i;L)]$ is the variance of the estimator $\hat{T}(x_i;L)$ in the subgroup l_i s.t. $l_i = l(x_i;L)$. We next describe an algorithm to construct a good partition that minimizes the above loss, as well as how we alter the above loss to further our goal of being able to robustly estimate group treatment effects.

4.5 Algorithm for Subgroup Identification

In this section we first assume access to a loss function $\mathcal{L}(L)$ and describe the recursive partitioning algorithm to minimize it. Further we discuss the modifications we apply to the empirical adjusted MSE in section 4.5.2 to obtain the loss function $\mathcal{L}(L)$.

4.5.1 Algorithm

In order to partition the feature space to different subgroups we minimize a loss function $\mathcal{L}(L)$ with recursive partitioning, $\min_{L\in\Pi}\ \mathcal{L}(L)$. First in the partitioning phase, similar to classification and regression tree (CART) [Breiman et al., 1984], we build a tree by greedily splitting the feature space to minimize the loss function, and stop splitting further when there is no such split that results in the reduction of the loss function (partitioning phase), we call this a treatment effect tree.

After building the treatment effect tree, each leaf l_i is a group and we can form an estimate of the group treatment effect by equation 4.4 (estimation phase). Note that in the estimation phase, different OPE methods such as model based and doubly robust [Thomas and Brunskill, 2016, Liu et al., 2018b] can be used to form the prediction. In this work, we use the same estimator in the partitioning and estimation phase and mainly focus on developing a loss function to be used in the partitioning phase. Additionally, we compute confidence intervals around our estimation by bootstrapping [Efron and Tibshirani, 1994].

4.5.2 Loss Function

One way to estimate the empirical adjusted MSE in equation 4.12 is by substituting the variance term with the sample variance of the estimator. For example for the estimator $\hat{T}(x;L)$ defined in equation 4.4 the sample variance $\widehat{\mathbb{V}}[.]$ can be estimated by,

$$\widehat{\mathbb{V}}[\hat{T}(x_i;L)] = \frac{1}{|l_i|}\sum_{i|x_i\in l_i}\Big(g_i(\rho_i-1)-\hat{T}(x_i;L)\Big)^2 \tag{4.13}$$

where, $l_i = l(x_i;L)$. By substituting the sample variance in equation 4.13 into equation 4.12, we obtain a loss function that can be easily computed by data. This is similar to the loss proposed by Athey and Imbens [2016] in the non-sequential setting.

However, estimation of the sample variance may be very noisy due to limited data, particularly in our sequential setting. A mis-estimation of the variance may result in an avoidable undesirable split in partitioning phase that would have not happened given a better estimate of the variance. Indeed, over-splitting is a common failure mode of using this loss function as we demonstrate in our experiments. We now consider alternate estimates to the sample variance for use in our target loss function.

Variance Estimation. To mitigate the issue of over-splitting we modify the loss function by a proxy of the variance term which can be computed efficiently. Our derivation is similar to a derivation for the variance of OPE presented in Metelli et al. [2018] with a few minor differences.

First note that the variance of the treatment effect estimator $\hat{T} = \frac{1}{N}\sum_i(\rho_i-1)g_i$ can be upper bounded by the variance of the importance sampling weights. Since $\mathbb{V}[\hat{T}] \le \mathbb{E}[\hat{T}^2]$

$$\mathbb{V}[\hat{T}] \leq \frac{\|g\|_\infty^2}{N^2}\mathbb{E}\left[\sum_i(\rho_i - 1)^2\right]$$
$$= \frac{1}{N}\|g\|_\infty^2 \mathbb{V}[\rho],$$

where the last equality follows by observing that $\mathbb{E}[\rho] = 1$. As noted by Metelli et al. [2018], we can use the Rényi divergence $D_\alpha(P_e||P_b)$ to represent moments of the importance weights of two distributions P_e and P_b, $w(x) = P_e(x)/P_b(x)$. Denote the exponential in base 2 by $d_\alpha(P_e||P_b) = 2^{D_\alpha(P_e||P_b)}$ and let P_e and P_b be the distribution of the trajectories introduced by the evaluation and behaviour policy. The variance of the treatment effect estimator can be written as

$$\mathbb{V}[\hat{T}] \leq \|g\|_\infty^2 \left(\frac{d_2(P_e||P_b)}{N} - \frac{1}{N}\right)$$

This expression can be related to the effective sample size of the original dataset given the evaluation policy

$$\mathbb{V}[\hat{T}] \leq ||g||_\infty^2 \left(\frac{1}{ESS} - \frac{1}{N}\right) \tag{4.14}$$

Note that in the special case of behaviour policy being the same as the evaluation policy, this bound evaluates to zero. We denote the RHS of equation 4.14 by $\mathbb{V}_u[.]$.

In our work, we use $\mathbb{V}_u[.]$ in each leaf as a proxy of variance of the estimator in the leaf. That is,

$$\mathbb{V}_u[\hat{T}(x_i; L)] = ||g(x)||_\infty^2 \left(\frac{1}{ESS(l_i)} - \frac{1}{|l_i|}\right), \tag{4.15}$$

where we use the common ESS Owen [2013] estimate $ESS(l_i)$ by,

$$\widehat{ESS}(l_i) = \frac{(\sum_j \rho_j)^2}{\sum_j \rho_j^2}$$

where the sum is over samples inside the group i, $\{j|x_j \in l_i\}$. $\mathbb{V}_u[.]$ can be computed efficiently and the conservative variance estimation using $\mathbb{V}_u[.]$ avoids the problem of variance underestimation. We will see experimentally our proposed approach can result in more stable partitioning.

Note that another approach to get a better estimate of the variance could be to leverage bootstrapping; however, using such a procedure is not feasible in the partitioning phase due to it's prohibitive high computational cost, since it has to be done for every evaluation of the loss function.

Regularization In many applications, actionable information needs to satisfy certain conditions. For example, a clinician may consider a knowledge of group treatment effect useful, if we can guarantee with high probability that the treatment effect is α bounded away from zero. Our above

loss function which is focused on minimizing the mean squared error would not necessarily identify these practically significant subgroups.

Therefore we now introduce a regularization term into our loss function to encourage finding such partitions where some subgroups have treatment effects that are bounded away from zero. In order to do so we use Cantelli's inequality to derive a lower bound on the estimator defined in equation 4.4. While this is a weaker bound than Bernstein, this allows us to avoid assuming we have access to an upper bound on the importance weights.

We do assume that the function $\hat{T}[x; L] : \mathcal{X} \to \mathbb{R}$, which is an expectation over states, is a bounded function ($||\hat{T}(x; L)||_\infty < \infty$). We start by writing Cantelli's inequality applied to the random variable $\hat{T}(x; L)$,

$$\mathbb{P}\left(\hat{T}(x; L) - \mathbb{E}[\hat{T}(x; L)] \geq \lambda\right) \leq \frac{1}{1 + \frac{\lambda^2}{\mathbb{V}[\hat{T}(x;L)]}}$$

Assigning δ to the right hand side and considering the complementary event, we have with probability $1 - \delta$,

$$\mathbb{E}[\hat{T}(x; L)] \geq \hat{T}(x; L) - \sqrt{\frac{1-\delta}{\delta}\mathbb{V}[\hat{T}(x; L)]} \tag{4.16}$$

With Equation 4.16 we define the (margin α) regularization term

$$\mathcal{R}(x_i; L, \alpha) = \max\left\{0, \alpha - \left(|\hat{T}(x_i; L)| - c\sqrt{\mathbb{V}_u[\hat{T}(x_i; L)]}\right)\right\} \tag{4.17}$$

Note that we used $\mathbb{V}_u[.]$ instead of $\mathbb{V}[.]$ in equation 4.17 to avoid issues arising from under estimation of the variance. Although we can obtain c by setting a specific value of δ, we view this as a tuning parameter for regularization. Recall that equation 4.17 is only used in the recursive partitioning process and the final bounds on the group treatment effect are obtained by bootstrapping in the estimation phase.

In a similar fashion, other types of regularizations depending on domain expert's input can be used in partitioning phase. For example, a domain expert may be interested to take more risks if the predicted group treatment effect is larger. This type of information can be incorporated as,

$$\mathcal{R}(x_i; L, \alpha) = \max\left\{0, \alpha - \frac{|\hat{T}(x_i; L)|}{c\sqrt{\mathbb{V}_u[\hat{T}(x_i; L)]}}\right\}$$

Loss Function By combining the regularization term and using the proxy variance, we obtain our final loss function.

$$\mathcal{L}(L) = -\frac{1}{N}\sum_{i=1}^{N}\hat{T}^2(x_i; L) + \frac{2}{N}\sum_{i=1}^{N}\mathbb{V}_u\left[\hat{T}(x_i; L)\right] + \frac{C}{N}\sum_{i=1}^{N}\mathcal{R}(x_i; L, \alpha), \tag{4.18}$$

where C is the regularization constant. This loss is minimized using recursive partitioning. We call our algorithm GIOPE, group identification in off-policy policy evaluation. Note in Theorem 5 we relied on $\hat{T}(x; L)$ be an unbiased estimate of $T(x; L)$. To accomplish this with our chosen estimator for $T(x; l)$ we use independent set of samples for partitioning phase and the estimation phase. The importance of sample splitting to avoid overfitting during off policy estimation is well studied (e.g. [Craig et al., 2020, Athey and Imbens, 2016]).

4.6 Experiments

We illustrate how our approach allows us to partition the feature space into subgroups such that we can make confident and accurate predictions of the group treatment effect. We empirically evaluate our method in sequential decision making settings, compare to the baseline and perform ablation analysis to show the benefit of each modification we have proposed.

We start by a simple toy example to illustrate the benefits of our method, later we evaluate our method on a simulated health care example, management of sepsis patients [Oberst and Sontag, 2019]. We compare to a strong baseline, causal forests [Athey et al., 2019] that was developed for non-sequential setting. Additionally, we use freely available MIMIC III dataset of ICU patients [Johnson et al., 2016], and focus on the sepsis cohort Komorowski et al. [2018] to show how our method can be used with real world data.

4.6.1 Toy MDP

We consider a simple Markov decision process (MDP) with the state space $x \in [0, 1]$, discrete action space $a \in \{-1, 0, 1\}$ and the reward function is defined as,

$$r(x) = 1 - |x - 0.5|$$

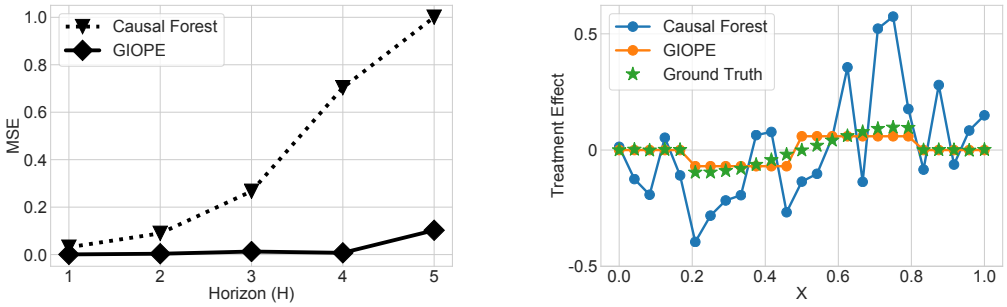
The transition dynamic is specified by,

$$x_{t+1} = clip(x_t + \kappa \times a_t + \epsilon, 0, 1)$$

where the function $clip(x, a, b)$, clips the value of x between a and b, $\kappa = 0.2$ and $\epsilon \sim \mathcal{N}(0, 0.05)$. Each episode lasts H steps. Intuitively, action 1 takes the agent to the right, -1 to the left and 0 same location with some gaussian noise. If the agent hits the boundary, the action has no effect on the position.

The behaviour policy, takes action with the following probabilities

$$\begin{cases} x < 0.2: \ \pi_b(-1) = 0.25, \ \pi_b(0) = 0.25, \ \pi_b(1) = 0.5 \\ x \geq 0.2: \ \pi_b(-1) = 0.5, \ \pi_b(0) = 0.25, \ \pi_b(1) = 0.25 \end{cases}$$



(a) (b)

Figure 4.1: Toy MDP. (a) Mean squared error of treatment effect prediction for our method and causal forest(CF). (b) True and predicted treatment effect for different values of x for our method and causal forest.

And the evaluation policy,

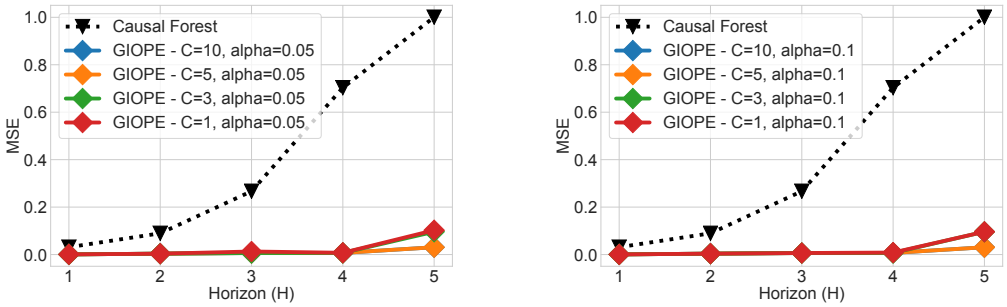
$$\begin{cases} x > 0.8: \ \pi_e(-1) = 0.5, \ \pi_e(0) = 0.25, \ \pi_e(1) = 0.25 \\ x \leq 0.8: \ \pi_e(-1) = 0.25, \ \pi_e(0) = 0.25, \ \pi_e(1) = 0.5 \end{cases}$$

We generated 50000 trajectories with the behaviour policy for horizons $\{1, 2, 3, 4, 5\}$ and averaged all results over 10 runs. First we look at the mean squared error of the treatment effect prediction on 25 equally spaced points in $[0, 1]$. Figure 4.1 (a) compares the MSE between our method (with regularization constant $C = 1.0$ and margin $\alpha = 0.05$) with causal forest (CF). GIOPE shows smaller MSE and as the horizon increases the benefit is more apparent. Figure 4.1 (b) shows the predicted value of the treatment effect for our method and causal forest for horizon $H = 4$ along with the true treatment effect for different values of x. This illustrates the reason of performance gap shown in figure 4.1 (a), our method partitions the state space and makes the same prediction for each subgroup that results in more accurate predictions, whereas causal forests over-splits and compute different values of the treatment effect for every value of x which are often inaccurate.

Figure 4.2 compares the mean squared error of our method versus the causal forests for different range of hyper-parameters. Panel (a) shows the results for margin $\alpha = 0.05$ and values of regularization constant $C = \{1, 3, 5, 10\}$ and panel (b) shows the results for margin $\alpha = 0.1$. As shown, regularization has small effects on the results and the results reported in the main text holds for a large range of hyper-parameters.

4.6.2 Sepsis Simulation

There has been growing number of literature that seek to learn an automated policy to manage septic patients in the ICU, reader may find a short review in Gottesman et al. [2019a]. However,



(a) $\alpha = 0.05$ (b) $\alpha = 0.1$

Figure 4.2: Toy MDP. (a) regularization margin $\alpha = 0.05$, (b) regularization margin $\alpha = 0.1$

newly suggested decision policies may be beneficial for some subgroups of patients while harmful to others. We use the sepsis simulator developed in Oberst and Sontag [2019] to show case this scenario and evaluate our models in detecting such subgroups.

Simulator In this simulator each patient is described by four vital signs {heart rate, blood pressure, oxygen concentration and glucose level} and a binary indicator of diabetes, that take values in a subset of {very high, high, normal, low, very low}, that results in a state space of size $|S| = 1440$. In each step the agent can take an action to put the patient on or off of treatment options, {antibiotics, vasopressors, and mechanical ventilation}, so that the action space has cardinality $|A| = 2^3$. Each episode run until the horizon H which incurs the reward of -1 upon death, +1 upon discharge and 0 otherwise. We use a discount factor of $\gamma = 0.99$ across all experiments, and all reported results are averaged over 15 different runs.

Data Generation In order to form the behaviour and the evaluation policy we assume both policies act nearly optimal with some modifications. We perform policy iteration to find the (deterministic) optimal policy for this environment and soften the policy by subtracting 0.1 probability from the optimal action and equally distributing it among other actions, we call this policy π_{st}. We assume the behaviour policy π_b is similar to π_{st} except it has has 15% less chance of using the mechanical ventilator. On the other hand, the evaluation policy π_e is similar to π_{st} but has 20% less chance of using the vasopressor. Notice that, the evaluation policy utilized the mechanical ventilator more and vassopressor less than the behaviour policy. Regardless of the horizon, the evaluation policy achieves better expected discounted return than the behaviour policy. However, there are subgroups of individuals, for example diabetics, that will worse off by using the evaluation policy. We generate 50000 trajectories using the behaviour policy for different horizons for different horizons {5, 7, 9, 11, 13}.

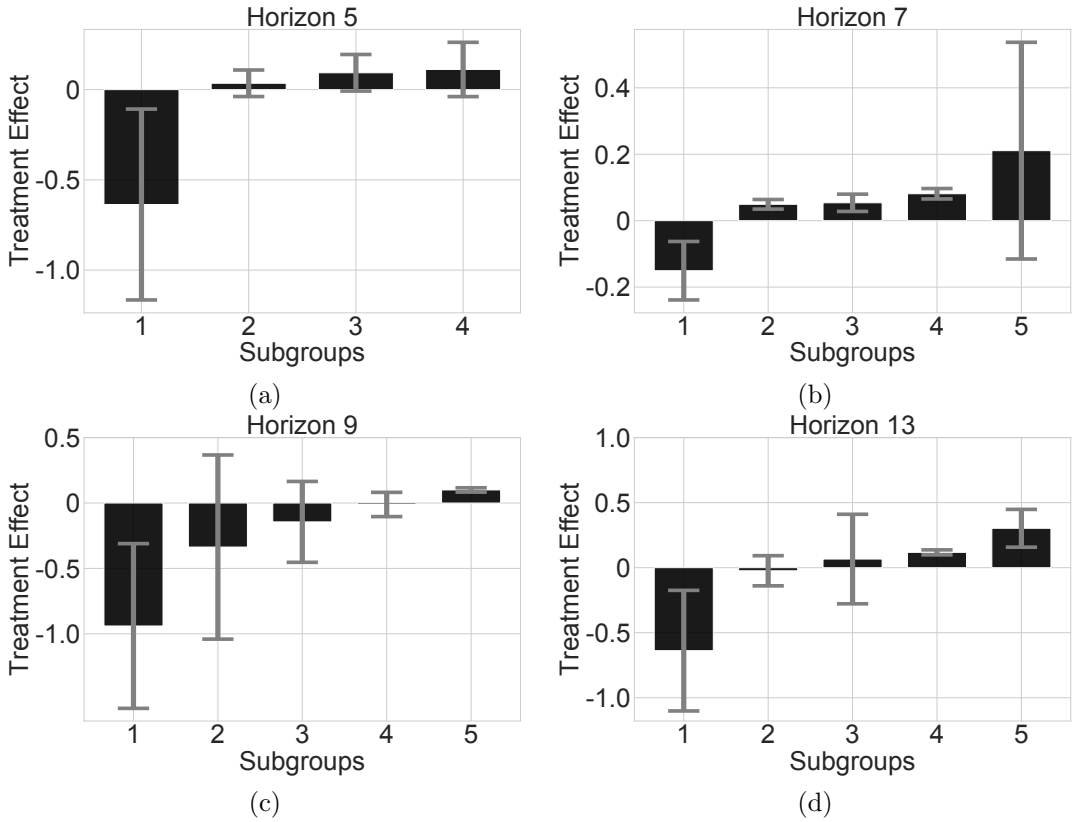


Figure 4.3: Identified groups. (a) Horizon 5,(b) Horizon 7, (c) 9 and (d) 13.

Hyperparameters We used the following set of hyper-parameters. Regularization constant $C = 5.0$, regularization margin $\alpha = 0.05$, regularization confidence value $\delta = 0.4$, maximum depth of the tree $d = \infty$ and minimum number of samples in each leaf 50.

Identified Subgroups First, we qualitatively describes the groups discovered by our method. Figure 4.3 show the subgroups and their group treatment effect for one run of horizons 5, 7, 9, and 13. One can see that in each setting, our methods can identify subgroups with significant negative treatment effect, group 1 in all settings, which all of them represents group with diabetes and elevated heart rate. This qualitative analysis unfortunately cannot be done for methods like causal forest as they are not designed to yield distinctive subgroups.

Comparison First we look at the mean squared error computed on the individual level. In order to compute this value, for each individual in the test set that consists of $n = 20000$ samples from the same distribution as the training set, we sample 30 different trajectories using the evaluation policy and the behaviour policy to compute the true treatment effect for each individual. Mean squared

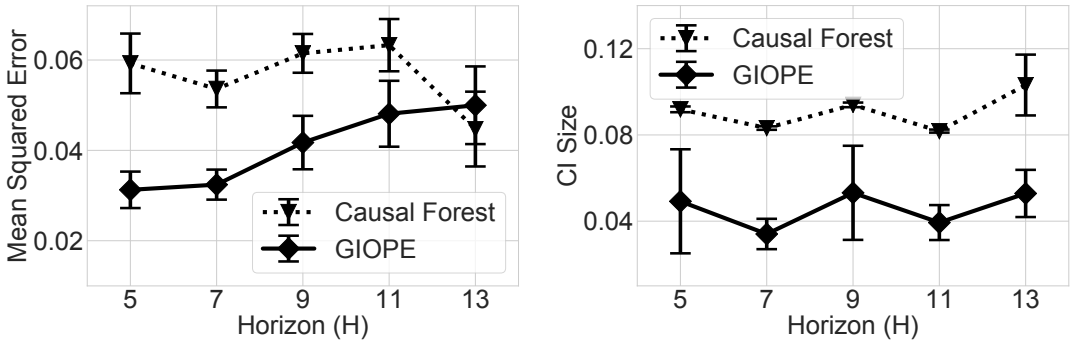


Figure 4.4: Sepsis simulator, comparison with causal forest (CF). (a) Mean squared error of prediction. (b) Average size of the 95% confidence intervals (CI)

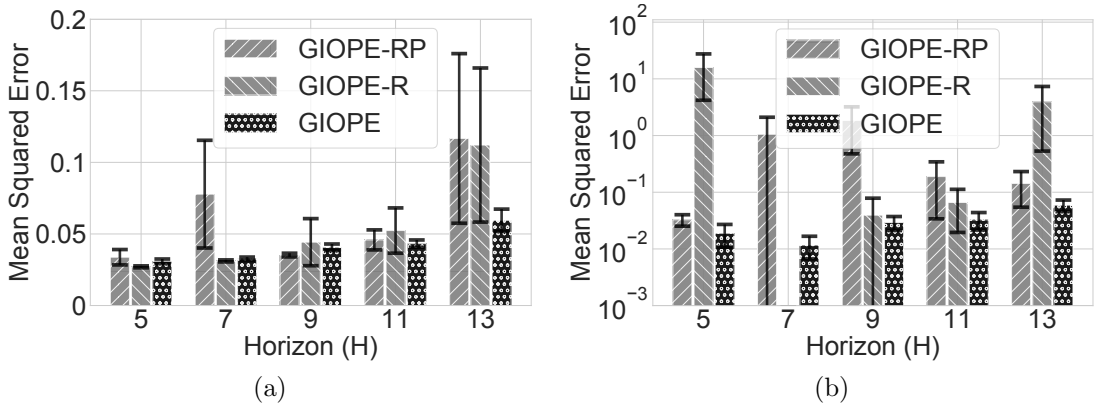


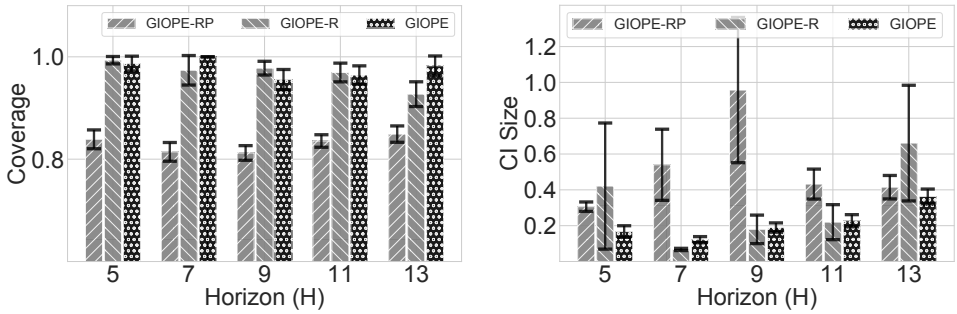
Figure 4.5: Ablation study. (a) Mean squared error computed on individual level. (b) Group mean squared error

errors is then simply

$$\frac{1}{n}\sum_{i=1}^{n}(t(x_i) - \hat{t}(x_i))^2$$

Figure 4.4 shows the mean squared error of prediction made by our method versus causal forest (CF). As shown in figure 4.4 (a), our method outperforms the baseline but as horizon increases both models struggles to generate valid results. Panel (b) of figure 4.4 shows the average size of the 95% confidence intervals. This highlights one of the main benefits of our method, that is more accurate prediction along with tighter confidence intervals.

Ablation Study In order to showcase the benefit of each modification that we proposed, we perform ablation study on the sepsis simulator. We compare three different methods with 15 different runs.



(a) (b)

Figure 4.6: 95% Coverage of Our method (*GIOPE*), Our method without regularization term (*GIOPE-R*) and Our method without regularization and proxy variance (*GIOPE-RP*). Results account for 15 different runs. (a) Percentage of groups that the true group treatment effect is covered by the 95% confidence interval. (b) Average size of confidence intervals.

.

1. *GIOPE*: Using the loss function presented in equation 4.18. In all experiments, the value of regularization is set to $C = 5.0$ and the margin $\alpha = 0.05$. We found that changing this regularization value has little effect on the results as we presented in the supplementary materials.

2. *GIOPE - Regularization (GIOPE-R)*: Using the loss function in equation 4.12 with the suggested proxy variance in equation 4.15.

3. *GIOPE - Regularization and Proxy Variance (GIOPE-RP)*: This method uses the loss function presented in equation 4.12 with the sample variance estimate. Note that this basic version is similar to the loss function proposed in Athey and Imbens [2016].

First we look at mean squared error computed on the individual level. As shown in Figure 4.5 (a) our method shows significant benefits compared to GIOPE-R and GIOPE-RP. This comes with an important observation that our method also shows more stability as the performance does not fluctuates as much across different horizons as well as having smaller standard errors. Note that, our method do not optimize for this objective and the individual mean squared error is best minimized with the sample variance in the limit of infinite data, the benefit comes as an externality of avoiding to predict each individual separately.

Next we look at mean squared error in group treatment effect. That is, for a groups i, denote the prediction of the group treatment effect by $\hat{g}_i$ and the true group treatment effect by g_i, then the group MSE is defined as

$$\frac{1}{G}\sum_{i=1}^{G}(g_i - \hat{g}_i)^2$$

where G is the total number of groups. Figure 4.5 (b) shows the MSE in group treatment effect as we increase the horizon. Similar to individual MSE, our method obtains lower MSE and displays

more stability across different horizons. This stability is mainly due to avoiding to over split. For example, average number of discovered groups in GIOPE-RP method for horizon 13 is 26 whereas for other GIOPE-R is 5 and GIOPE is 4.

Finally we look at coverage. Figure 4.6, panel (a) shows the coverage of 95% confidence intervals of the true group treatment effect for different methods and horizons. Methods that use variance proxy instead of sample variance show consistently more coverage. Figure 4.6 panel (b) shows the average size of the confidence interval for each group treatment effect prediction. This indicates that using the upper bound along with regularization (GIOPE) yields more coverage while offering tighter confidence intervals. This observation highlights the main benefit of using regularization along with proxy variance that allows us to discover groups that we can more accurately and confidently predict their treatment effect. Smaller standard error of confidence intervals size highlights the stability of GIOPE across different runs.

In order to evaluate the effect of hyper-parameters, we perform the ablation study for two different values of regularization confidence interval $\delta = 0.1$ and $\delta = 0.4$ and two different values of regularization constant $C = 5.0$ and $C = 1.0$. Figure 4.7 (a) shows the result for mean squared error and (b) for group mean squared error. As shown, the effect of regularization is small, and the same results as in the main text can be obtained with different range of hyper-parameters. Similarly, figure 4.7 (c) shows the coverage of the 95% confidence interval and (d) is the average size of CI. The results obtained in the main text holds with different value of hyper-parameters.

4.6.3 ICU data - MIMIC III

To show how our method can be used on a real data set, we use a cohort of 14971 septic patients in the freely accessible MIMIC III dataset [Johnson et al., 2016]. Prior work Komorowski et al. [2018] used off policy learning and proposed a new decision policy that might provide improved patient outcomes on average. Our training set consist of 14971 individuals, with 8442 male and 6529 female. The mortality rate in our cohort is 18.4%. The feature space is of size 44 consist of the following values: `gender`, `re_admission`, `mechvent`, `age`, `Weight_kg`, `GCS`, `HR`, `SysBP`, `MeanBP`, `DiaBP`, `RR`, `Temp_C`, `FiO2_1`, `Potassium`, `Sodium`, `Chloride`, `Glucose`, `Magnesium`, `Calcium`, `Hb`, `WBC_count`, `Platelets_count`, `PTT`, `PT`, `Arterial_pH`, `paO2`, `paCO2`, `Arterial_BE`, `Arterial_lactate`, `HCO3`, `Shock_Index`, `Shock_Index`, `PaO2_FiO2`, `cumulated_balance`, `SOFA`, `SIRS`, `SpO2`, `BUN`, `Creatinine`, `SGOT`, `SGPT`, `Total_bili`, `INR`, `output_total`, `output_4hourly`.

In order to estimate the behaviour policy, we use KNN with $k = 100$ on the test set, we use l_2 distance with uniform weights across different features to measure the distance. If an action was not taken among all 100 nearest neighbours, we assign the probability 0.01 to the action. We used IV fluid and mechanical ventilation for actions and used 20% quantile to discretize the action space into 25 actions.

For the evaluation policy, we used a similar method as the behaviour policy on a random subset

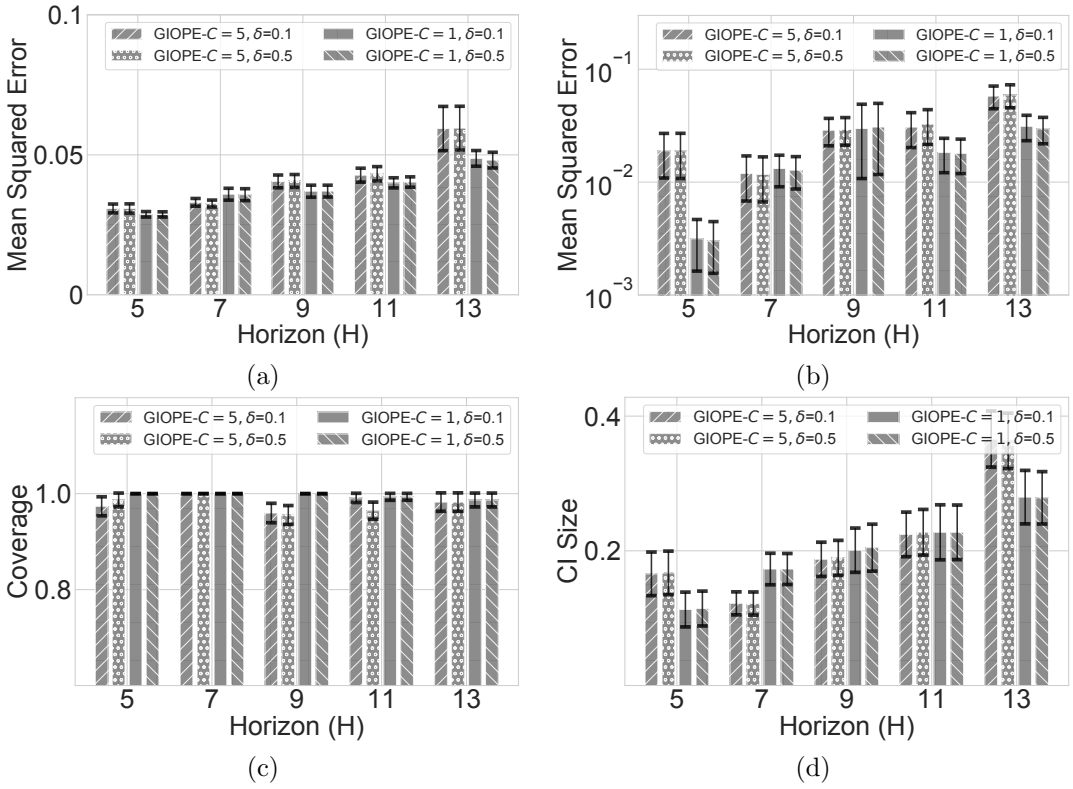


Figure 4.7: Ablation study, results of GIOPE for four different values of parameters. (a) Mean squared error (b) group mean squared error, (c) 95% confidence interval coverage and (d) average size of confidence intervals

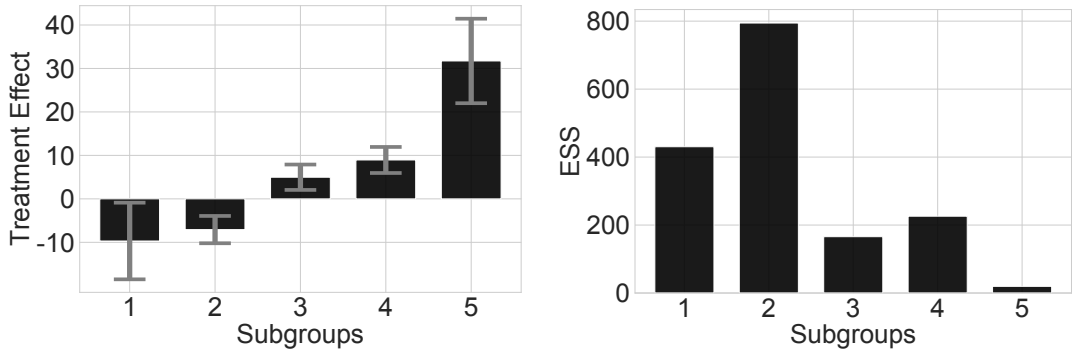
of training set (20% of the training data). We only used the following features to estimate the distance for the evaluation policy,

{`HR`, `SysBP`, `Temp_C`, `Sodium`, `Chloride`, `Glucose`, `Calcium`, `paO2`, `Arterial_BE`, `SOFA`, `SIRS`, `Creatinine`}

Simiarly, if an action was not taken among all 100 nearest neighbours, we assign the probability 0.01 to the action. We used the following set of hyper-parameters: regularization constant $C = 100.0$, regularization margin $\alpha = 0.0$, regularization confidence value $c = 2.0$, maximum depth of the tree $d = \infty$ and minimum number of samples in each leaf 1000.

Using weighted importance sampling the estimated value of the decision policy is 65.33 with effective sample size of 146.8 which suggest an increase of 2.43 on the survival chance compared to the behaviour policy. Here we take this decision policy and estimate its impact on different potential subgroups.

In Figure 4.8 we present the five groups produced by our algorithm along with their estimated group treatment effect (which is the difference between the baseline clinician policy and the decision



(a) (b)

Figure 4.8: MIMIC III dataset. Although positive treatment effect is predicted by weighted importance sampling on the full cohort, groups 1 and 2 will like be harmed by the evaluation policy. (a) Estimated treatment effect for each subgroup, (b) Effective sample size of weighted importance sampling for each subgroup

policy) and effective sample size of weighted importance sampling in each subgroup. While some of patients fall into subgroup 3, 4 and 5, there are a number of patients that may experience no benefit or even a potential negative treatment effect from the proposed new treatment policy (groups 2 and 1). This highlights how our method may be useful in identifying subgroups in which a new decision policy may be beneficial or harmful relative to the standard approach.

We caveat the results in this section by noting that using IS based methods on real world datasets, and the MIMIC III dataset in particular is very susceptible to noise induced by the small effective sample size of the cohort [Gottesman et al., 2019a]. Furthermore, our method is susceptible to this source of noise twice, as IS based estimators are used both in the partitioning phase and the estimation phase. However, despite their high susceptibility to noise, IS methods are often applied to the MIMIC III dataset for their theoretical properties, but their results for real data should be interpreted with caution. In our experiment we intentionally designed the decision policy close to the behaviour policy to avoid issues arising from small effective sample size.

4.7 Summary and Conclusion

In this chapter, we proposed a novel method to partition the feature space, enabling us to find subgroups that we can accurately and confidently predict the group treatment effect for them. Our approach is in contrast with previous methods that estimate individual-level treatment effects, yielding uncertain and less accurate predictions. We do so, by proposing a novel loss function that utilizes; 1. A proxy on the variance estimator that is easy to compute and stable; 2. A regularization term that incentivizes the discovery of groups with significant treatment effects and allows us to integrate domain expert's input into partitioning algorithm. We further evaluate our

method on both simulated domains and real-world data.

Our method can leverage the existing data to raise caution when necessary about a possible negative effect of the newly suggested decision policy on some subgroups. Additionally, results from our method when applied to observational data can help to design multi-stage randomized trials that are powered toward detecting harm or benefit of the evaluation policy compared to the baseline policy for specific subgroups.

There are multiple immediate avenues of interesting research questions. Integrating other methods of off-policy policy evaluation in the partitioning phase, for example doubly robust or model-based method [Thomas and Brunskill, 2016]. This requires obtaining a stable estimate of the variance of the estimator that can be efficiently computed. We suggested using recursive partitioning to minimize the loss function; however, it's possible that these greedy algorithms fail to find the optimal solution; thus, an immediate question is how can we apply better optimization techniques. Finally, we considered partitions with respect to the initial state, an interesting avenue of research is extending our work to incorporate transition dynamics in the partitioning.

Chapter 5

Conclusion

5.1 Future Research Possibilities

We now briefly outline some of the immediate avenues of research opportunities based on the research conducted in this dissertation.

Safe Exploration for Safe Policy Learning: The algorithm described in chapter 2 proposes a scalable and efficient way to conduct sample efficient learning while learning a safe policy. However, another important question is how to achieve this goal while performing safe exploration. For example, consider an automated policy to regulate the blood glucose level of type 1 diabetes patients, as described in section 2.6. The automated policy has to avoid areas of hypo/ hyperglycemia not only when deploying the optimal policy, but also in the process of learning such policy. There has been literature concerning safe exploration in MDP [Moldovan and Abbeel, 2012, Berkenkamp et al., 2017]. While most of the literature focuses on safe exploration for risk-neutral objectives such as expected value, it would be interesting to extend this research to safe exploration for risk-sensitive objectives such as CVaR.

Subgroup Identification in Off-Policy Policy Evaluation using Transition Dynamics: In Chapter 4 we developed an algorithm to identify subgroups with similar treatment effects in off-policy policy evaluation where we considered subgroups based on the initial state. Another interesting question is how to identify subgroups based on the transition dynamics. For example, consider the problem of managing sepsis patients presented in section 4.6. Our approach in Chapter 4 looks at patients upon admission and identifies subgroups that would benefit or hurt based on the new decision policy. Although this approach may yield clinically relevant subgroups, and there has been some evidence supporting this way of grouping for different sepsis phenotypes [Seymour et al., 2019], it is interesting to consider subgroups based on the development of their clinical conditions.

For example, patients that react differently/similarly to various medications may be part of the same subgroup suggesting partitioning based on transition dynamics rather than initial states.

Model Misspecification in Model Based Off-Policy Policy Evaluation: In Chapter 3 and 4, we discussed overlap as one of the main limitations of off-policy policy evaluation (OPE). Small overlap hinders our ability to conduct statistical inference and results in high variance estimates. One way to make progress in OPE is to construct a model of the environment given the collected data and perform OPE using the model. In model-based OPE, we will encounter the same problem, that is the evaluation policy may explore part of the state-action space that is under-explored by the behavior policy. To mitigate this issue there has been recent interest in using pessimism to discourage the RL agent to explore the part of the state-action space that we are highly uncertain about [Yu et al., 2020, Liu et al., 2020].

Even with pessimism, learning a good model is important. We can use general function approximation techniques to build the model. However we need a large amount of data to train them, and in many real-world applications, we do not have access to this amount of data. Alternatively, we can learn a model from the space of limited capacity models that we can train using the amount of data available. However, by doing so, we will probably encounter model misspecification, that is the true model of the world does not fall into our model class. An interesting avenue of research is how to account for model misspecification in OPE. How to detect them? and How to develop bounds on OPE under model misspecification?

5.2 Summary of Contributions

We started this work by outlining recent advances of Reinforcement Learning (RL) in applications such as playing games and robotics. We noted that RL has achieved human or superhuman level performances in those applications. However, in some real-world applications such as education and health care those performances are not matched. We then outlined some of the main challenges that arise when we try to apply RL in real-world applications. This dissertation is a step toward a more robust RL that helps us unleash the potentials of RL to high-stakes real-world applications.

In Chapter 2, we considered online RL, where an agent is allowed to interact with the environment. We argued that in many real-world applications, we may be interested in the distribution of the outcome instead of the expected value, especially we may want to avoid catastrophic outcomes. That motivated us to consider risk-sensitive objectives such as CVaR (Conditional Value at Risk) instead of risk-neutral objectives such as expected value.

In many real-world applications number of interactions with the environment is limited, so we focused on the sample efficiency of our algorithm. We developed a sample efficient algorithm to learn a

CVaR-optimal policy. To do so, we built upon recent advances of distributional Reinforcement Learning [Bellemare et al., 2017] to learn the distribution of the outcome. To encourage sample efficient exploration, we integrated the principle of Optimism in the Face of Uncertainty (OFU) [Brafman and Tennenholtz, 2002] in distributional RL by introducing a novel operator, distributional optimistic operator. We evaluated the performance of our method on two simulated health domains: automation of insulin injection in type 1 diabetes patients, and managing treatment of HIV patients. We showed that our method outperforms the baseline in those domains.

In Chapter 3 and 4 we considered offline RL, where the agent is not allowed to interact with the environment, so we need to rely on the historically collected data. We considered the problem of off-policy policy evaluation (OPE) where we are interested in evaluating a new decision policy (evaluation policy) from the data collected by another policy (behavior policy). We tackled two main problems that arise while doing OPE in real-world applications.

In Chapter 3, we considered the problem of unobserved confounding. Does the RL agent have access to all the information that was used by the behavior policy to make a decision? For example, the clinician may use visual observations or conversations with the patients that are not recorded in the electronic health record but affect their decision-making. We showed that under no assumption unobserved confounding can make OPE arbitrary biased. Then we introduced the bounded confounding model, to bound the effect of unobserved confounding on the behavior policy.

Under the bounded confounding model, if confounding happens at every time step, OPE can be exponentially biased (in the horizon). We argued that in many practical applications, confounding happens in only one-time step. Under these assumptions, we developed bound on OPE to assess its robustness to the unobserved confounding. We evaluated our method in two simulates health domains: management of sepsis patients, and treatment policy for autistic children. We showed that our method can be used to certify or cast doubts on the value of the evaluation policy under unobserved confounding.

In Chapter 4, we considered heterogeneity in OPE. An evaluation policy is often evaluated with respect to its expected outcome; however, the policy may affect different subgroups of the population differently. We developed an algorithm to automatically identify those subgroups that will hurt or benefit from adopting the new decision policy instead of using the behavior policy.

To conclude, while much more research remains to be done, this dissertation takes a step towards more robust reinforcement learning that can be applied to real-world applications.

Bibliography

Carlo Acerbi and Dirk Tasche. On the coherence of expected shortfall. *Journal of Banking & Finance*, 26(7):1487–1503, 2002.

Alekh Agarwal, Sarah Bird, Markus Cozowicz, Luong Hoang, John Langford, Stephen Lee, Jiaji Li, Dan Melamed, Gal Oshri, Oswaldo Ribas, et al. Making contextual decisions with low technical debt. *arXiv:1606.03966*, 2016.

Ilge Akkaya, Marcin Andrychowicz, Maciek Chociej, Mateusz Litwin, Bob McGrew, Arthur Petron, Alex Paino, Matthias Plappert, Glenn Powell, Raphael Ribas, et al. Solving rubik's cube with a robot hand. *arXiv preprint arXiv:1910.07113*, 2019.

Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, and Guoliang Chen. Deep speech 2: end-to-end speech recognition in English and Mandarin. pages 173–182, 2016.

Deborah K. Anderson, Rosalind S. Oti, Catherine Lord, and Kathleen Welch. Patterns of growth in adaptive social abilities among children with autism spectrum disorders. *Journal of Abnormal Child Psychology*, 37(7):1019–1034, Oct 2009. ISSN 1573-2835. doi: 10.1007/s10802-009-9326-0. URL `https://doi.org/10.1007/s10802-009-9326-0`.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical finance*, 9(3):203–228, 1999.

Thomas Åstebro and Samir Elhedhli. The effectiveness of simple decision heuristics: Forecasting commercial success for early-stage ventures. *Management Science*, 52(3):395–409, 2006.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Susan Athey and Stefan Wager. Estimating treatment effects with causal forests: An application. *arXiv preprint arXiv:1902.07409*, 2019.

Susan Athey, Julie Tibshirani, Stefan Wager, et al. Generalized random forests. *Annals of Statistics*, 47(2):1148–1178, 2019.

Meysam Bastani. Model-free intelligent diabetes management using machine learning. 2014.

Marc Bellemare, Sriram Srinivasan, Georg Ostrovski, Tom Schaul, David Saxton, and Remi Munos. Unifying count-based exploration and intrinsic motivation. In *Advances in Neural Information Processing Systems*, pages 1471–1479, 2016.

Marc G Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pages 449–458. PMLR, 2017.

Felix Berkenkamp, Matteo Turchetta, Angela Schoellig, and Andreas Krause. Safe model-based reinforcement learning with stability guarantees. In *Advances in neural information processing systems*, pages 908–918, 2017.

Alberto Bietti, Alekh Agarwal, and John Langford. A contextual bandit bake-off. *arXiv preprint arXiv:1802.04064*, 2018.

Su L. Blodgett, Lisa Green, and Brendan O'Connor. Demographic dialectal variation in social media: A case study of African-American English. pages 1119–1130, 2016.

Ronen I Brafman and Moshe Tennenholtz. R-max-a general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3(Oct):213–231, 2002.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

Andrew J. Brent. Meta-analysis of time to antimicrobial therapy in sepsis: Confounding as well as bias. *Critical Care Medicine*, 45(2), 2017.

David B Brown. Large deviations bounds for estimating conditional value-at-risk. *Operations Research Letters*, 35(6):722–730, 2007.

Babette A. Brumback, Miguel A. Hernán, Sebastien J. P. A. Haneuse, and James M. Robins. Sensitivity analyses for unmeasured confounding assuming a marginal structural model for repeated measures. *Statistics in Medicine*, 23(5):749–767, 2004.

Sébastien Bubeck, Nicolo Cesa-Bianchi, et al. Regret analysis of stochastic and nonstochastic multi-armed bandit problems. *Foundations and Trends® in Machine Learning*, 5(1):1–122, 2012.

Asaf Cassel, Shie Mannor, and Assaf Zeevi. A general approach to multi-armed bandits under risk criteria. In *COLT*, 2018.

Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G. Bellemare. Dopamine: A Research Framework for Deep Reinforcement Learning. 2018. URL `http://arxiv.org/abs/1812.06110`.

Bibhas Chakraborty and Susan A Murphy. Dynamic treatment regimes. *Annual review of statistics and its application*, 1:447–464, 2014.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

Yinlam Chow. *Risk-sensitive and data-driven sequential decision making*. PhD thesis, PhD thesis, Institute of Computational and Mathematical Engineering . . . , 2017.

Yinlam Chow and Mohammad Ghavamzadeh. Algorithms for cvar optimization in mdps. *arXiv preprint arXiv:1406.3339*, 2014.

Yinlam Chow, Aviv Tamar, Shie Mannor, and Marco Pavone. Risk-sensitive and robust decision-making: a cvar optimization approach. *arXiv preprint arXiv:1506.02188*, 2015.

Yinlam Chow, Mohammad Ghavamzadeh, Lucas Janson, and Marco Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *The Journal of Machine Learning Research*, 18(1):6070–6120, 2017.

William Clarke and Boris Kovatchev. Statistical tools to analyze continuous glucose monitor data. *Diabetes technology & therapeutics*, 11(S1):S–45, 2009.

Jerome Cornfield, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder. Smoking and lung cancer: Recent evidence and a discussion of some questions. *Journal of the National Cancer Institute*, 22(1):173–203, 1959.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. In *Nips*, volume 10, pages 442–450. Citeseer, 2010.

Erin Craig, Donald A Redelmeier, and Robert J Tibshirani. Finding and assessing treatment effect sweet spots in clinical trial data. *arXiv preprint arXiv:2011.10157*, 2020.

Will Dabney, Georg Ostrovski, David Silver, and Rémi Munos. Implicit quantile networks for distributional reinforcement learning. In *International conference on machine learning*, pages 1096–1105. PMLR, 2018.

Christoph Dann, Lihong Li, Wei Wei, and Emma Brunskill. Policy certificates: Towards accountable reinforcement learning. *arXiv preprint arXiv:1811.03056*, 2018.

Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. Extraneous factors in judicial decisions. *Proceedings of the National Academy of Sciences*, 108(17):6889–6892, 2011.

Erick Delage and Shie Mannor. Percentile optimization for markov decision processes with parameter uncertainty. *Operations research*, 58(1):203–213, 2010.

Mandeep K Dhami. Psychological models of professional decision making. *Psychological Science*, 14 (2):175–180, 2003.

Nat Dilokthanakul and Murray Shanahan. Deep reinforcement learning with risk-seeking exploration. In *International Conference on Simulation of Adaptive Behavior*, pages 201–211. Springer, 2018.

Laurent Dinh, Jascha Sohl-Dickstein, and Samy Bengio. Density estimation using real nvp. *arXiv preprint arXiv:1605.08803*, 2016.

Miroslav Dudík, John Langford, and Lihong Li. Doubly robust policy evaluation and learning. *arXiv preprint arXiv:1103.4601*, 2011.

Aryeh Dvoretzky, Jack Kiefer, Jacob Wolfowitz, et al. Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *The Annals of Mathematical Statistics*, 27(3):642–669, 1956.

Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

Damien Ernst, Guy-Bart Stan, Jorge Goncalves, and Louis Wehenkel. Clinical data based optimal sti strategies for hiv: a reinforcement learning approach. In *Proceedings of the 45th IEEE Conference on Decision and Control*, pages 667–672. IEEE, 2006.

Scott Fujimoto, David Meger, and Doina Precup. Off-policy deep reinforcement learning without exploration. 2019.

Joseph Futoma, Anthony Lin, Mark Sendak, Armando Bedoya, Meredith Clement, Cara O'Brien, and Katherine Heller. Learning to treat sepsis with multi-output gaussian process deep recurrent q-networks, 2018. URL `https://openreview.net/forum?id=SyxCqGbRZ`.

Nicolas Galichet, Michele Sebag, and Olivier Teytaud. Exploration vs exploitation vs safety: Risk-aware multi-armed bandits. In *Asian Conference on Machine Learning*, pages 245–260, 2013.

Alborz Geramifard, Christoph Dann, Robert H Klein, William Dabney, and Jonathan P How. Rlpy: a value-function-based reinforcement learning framework for education and research. *Journal of Machine Learning Research*, 16(46):1573–1578, 2015.

Omer Gottesman, Fredrik Johansson, Matthieu Komorowski, Aldo Faisal, David Sontag, Finale Doshi-Velez, and Leo Anthony Celi. Guidelines for reinforcement learning in healthcare. *Nature medicine*, 25(1):16–18, 2019a.

Omer Gottesman, Yao Liu, Scott Sussex, Emma Brunskill, and Finale Doshi-Velez. Combining parametric and nonparametric models for off-policy evaluation. In *International Conference on Machine Learning*, pages 2366–2375, 2019b.

Alex Graves, Marc G Bellemare, Jacob Menick, Remi Munos, and Koray Kavukcuoglu. Automated curriculum learning for neural networks. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1311–1320. JMLR. org, 2017.

Patrick J Grother, George W Quinn, and P Jonathon Phillips. Report on the evaluation of 2d still-image face recognition algorithms. *NIST Interagency/Internal Reports (NISTIR)*, 7709, 2010.

Assaf Hallak and Shie Mannor. Consistent on-line off-policy evaluation. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 1372–1383. JMLR. org, 2017.

Josiah Hanna, Scott Niekum, and Peter Stone. Importance sampling policy evaluation with an estimated behavior policy. In *Proceedings of the 36th International Conference on Machine Learning (ICML)*, June 2019.

Josiah P Hanna, Peter Stone, and Scott Niekum. Bootstrapping with models: Confidence intervals for off-policy evaluation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. 2018.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

MA Hernán and JM Robins. *Causal Inference: What If*. Boca Raton: Chapman & Hall/CRC, 2020.

Dirk Hovy and Anders Søgaard. Tagging performance correlates with author age. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (Short Papers)*, volume 2, pages 483–488, 2015.

Michael D. Howell and Andrew M. Davis. Management of Sepsis and Septic Shock. *JAMA*, 317(8): 847–848, 02 2017. ISSN 0098-7484. doi: 10.1001/jama.2017.0131. URL `https://doi.org/10.1001/jama.2017.0131`.

Tien-Chung Hu, F Moricz, and R Taylor. Strong laws of large numbers for arrays of rowwise independent random variables. *Acta Mathematica Hungarica*, 54(1-2):153–162, 1989.

Kosuke Imai and Marc Ratkovic. Covariate balancing propensity score. *Journal of the Royal Statistical Society: Series B: Statistical Methodology*, pages 243–263, 2014.

Guildo W Imbens. Sensitivity to exogeneity assumptions in program evaluation. *American Economic Review*, 93(2):126–132, 2003.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for reinforcement learning. 2010.

Nan Jiang and Lihong Li. Doubly robust off-policy value evaluation for reinforcement learning. *arXiv preprint arXiv:1511.03722*, 2015.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3:160035, 2016.

Jongbin Jung, Ravi Shroff, Avi Feller, and Sharad Goel. Algorithmic decision making in the presence of unmeasured confounding. *arXiv:1805.01868 [stat.ME]*, 2018.

Nathan Kallus and Masatoshi Uehara. Double reinforcement learning for efficient off-policy evaluation in markov decision processes. *Journal of Machine Learning Research*, 21(167):1–63, 2020.

Nathan Kallus and Angela Zhou. Confounding-robust policy improvement. In *Advances in Neural Information Processing Systems*, pages 9269–9279, 2018.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Interval estimation of individual-level causal effects under unobserved confounding. *arXiv preprint arXiv:1810.02894*, 2018.

Connie Kasari, Ann Kaiser, Kelly Goods, Jennifer Nietfeld, Pamela Mathy, Rebecca Landa, Susan Murphy, and Daniel Almirall. Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 53(6):635–646, 2014.

Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being optimistic to be conservative: Quickly learning a cvar policy. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4436–4443, 2020.

Alan J King and Roger JB Wets. Epi-consistency of convex stochastic programs. *Stochastics and Stochastic Reports*, 34(1-2):83–92, 1991.

Ravi Kumar Kolla, Krishna Jagannathan, et al. Risk-aware multi-armed bandits using conditional value-at-risk. *arXiv preprint arXiv:1901.00997*, 2019.

Torsten Koller, Felix Berkenkamp, Matteo Turchetta, and Andreas Krause. Learning-based model predictive control for safe exploration. In *2018 IEEE Conference on Decision and Control (CDC)*, pages 6059–6066. IEEE, 2018.

Matthieu Komorowski, Leo A Celi, Omar Badawi, Anthony C Gordon, and A Aldo Faisal. The artificial intelligence clinician learns optimal treatment strategies for sepsis in intensive care. *Nature medicine*, 24(11):1716–1720, 2018.

Augustine Kong. A note on importance sampling using standardized weights. *University of Chicago, Dept. of Statistics, Tech. Rep*, 348, 1992.

Aviral Kumar, Justin Fu, Matthew Soh, George Tucker, and Sergey Levine. Stabilizing off-policy q-learning via bootstrapping error reduction. In *Advances in Neural Information Processing Systems*, pages 11761–11771, 2019.

Tor Lattimore and Csaba Szepesvári. Bandit algorithms. *preprint*, 2018.

Hoang Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *International Conference on Machine Learning*, pages 3703–3712, 2019.

Hyun-Suk Lee, Yao Zhang, William Zame, Cong Shen, Jang-Won Lee, and Mihaela van der Schaar. Robust recursive partitioning for heterogeneous treatment effects with uncertainty quantification. *arXiv preprint arXiv:2006.07917*, 2020.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*, 2016.

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*, pages 661–670, 2010.

Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. pages 5356–5366, 2018a.

Yao Liu, Omer Gottesman, Aniruddh Raghu, Matthieu Komorowski, Aldo A Faisal, Finale Doshi-Velez, and Emma Brunskill. Representation balancing mdps for off-policy policy evaluation. In *Advances in Neural Information Processing Systems*, pages 2644–2653, 2018b.

Yao Liu, Adith Swaminathan, Alekh Agarwal, and Emma Brunskill. Provably good batch reinforcement learning without great exploration. *arXiv preprint arXiv:2007.08202*, 2020.

Chaochao Lu, Bernhard Schölkopf, and José Miguel Hernández-Lobato. Deconfounding reinforcement learning in observational settings. *arXiv preprint arXiv:1812.10576*, 2018.

Xi Lu, Inbal Nahum-Shani, Connie Kasari, Kevin G Lynch, David W Oslin, William E Pelham, Gregory Fabiano, and Daniel Almirall. Comparing dynamic treatment regimes using repeated-measures outcomes: modeling considerations in smart studies. *Statistics in medicine*, 35(10): 1595–1615, 2016.

David Luenberger. *Optimization by Vector Space Methods*. Wiley, 1969.

Ashique Rupam Mahmood, Hado Van Hasselt, and Richard S Sutton. Weighted importance sampling for off-policy learning with linear function approximation. In *NIPS*, pages 3014–3022, 2014.

Chiara Dalla Man, Francesco Micheletto, Dayu Lv, Marc Breton, Boris Kovatchev, and Claudio Cobelli. The uva/padova type 1 diabetes simulator: new features. *Journal of diabetes science and technology*, 8(1):26–34, 2014.

Travis Mandel, Yun-En Liu, Sergey Levine, Emma Brunskill, and Zoran Popovic. Offline policy evaluation across representations with applications to educational games. In *AAMAS*, pages 1077–1084, 2014.

Charles F Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80(2):319–323, 1990.

Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In *International Conference on Machine Learning*, pages 4424–4434, 2019.

Clement J McDonald. Medical heuristics: the silent adjudicators of clinical practice. *Annals of Internal Medicine*, 124(1_Part_1):56–62, 1996.

Alberto Maria Metelli, Matteo Papini, Francesco Faccio, and Marcello Restelli. Policy optimization via importance sampling. 2018.

Wang Miao, Zhi Geng, and Eric J Tchetgen Tchetgen. Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika*, 105(4):987–993, 2018.

Oliver Mihatsch and Ralph Neuneier. Risk-sensitive reinforcement learning. *Machine learning*, 49 (2):267–290, 2002.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Thomas M Moerland, Joost Broekens, and Catholijn M Jonker. The potential of the return distribution for exploration in rl. *arXiv preprint arXiv:1806.04242*, 2018.

Teodor Mihai Moldovan and Pieter Abbeel. Safe exploration in markov decision processes. *arXiv preprint arXiv:1205.4810*, 2012.

Susan A Murphy. Optimal dynamic treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 65(2):331–355, 2003.

Hongseok Namkoong, Ramtin Keramati, Steve Yadlowsky, and Emma Brunskill. Off-policy policy evaluation for sequential decisions under unobserved confounding. *arXiv preprint arXiv:2003.05623*, 2020.

Xinkun Nie and Stefan Wager. Quasi-oracle estimation of heterogeneous treatment effects. *arXiv preprint arXiv:1712.04912*, 2017.

Xinkun Nie, Emma Brunskill, and Stefan Wager. Learning when-to-treat policies. *arXiv preprint arXiv:1905.09751*, 2019.

Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *The New England journal of medicine*, 375(13):1216, 2016.

Michael Oberst and David Sontag. Counterfactual off-policy evaluation with Gumbel-max structural causal models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 4881–4890, Long Beach, California, USA, 09–15 Jun 2019. PMLR. URL `http://proceedings.mlr.press/v97/oberst19a.html`.

Georg Ostrovski, Marc G Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2721–2730. JMLR. org, 2017.

Art B. Owen. *Monte Carlo theory, methods and examples.* 2013.

Judea Pearl. *Causality.* Cambridge University Press, 2009.

Doina Precup. Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series*, page 80, 2000.

Aniruddh Raghu, Matthieu Komorowski, Leo Anthony Celi, Peter Szolovits, and Marzyeh Ghassemi. Continuous state-space models for optimal sepsis treatment-a deep reinforcement learning approach. *arXiv preprint arXiv:1705.08422*, 2017.

Andrew Rhodes, Laura E. Evans, Waleed Alhazzani, et al. Surviving sepsis campaign: International guidelines for management of sepsis and septic shock: 2016. *Intensive Care Medicine*, 43(3): 304–377, 2017.

James Robins. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical modelling*, 7 (9-12):1393–1512, 1986.

James M Robins. Optimal structural nested models for optimal sequential decisions. In *Proceedings of the Second Seattle Symposium in Biostatistics*, pages 189–326. Springer, 2004.

James M. Robins, Andrea Rotnitzky, and Daniel O. Scharfstein. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In M. Elizabeth Halloran and Donald Berry, editors, *Statistical Models in Epidemiology, the Environment, and Clinical Trials*, pages 1–94, New York, NY, 2000. Springer New York. ISBN 978-1-4612-1284-3.

R. T. Rockafellar and R. J. B. Wets. *Variational Analysis*. Springer, New York, 1998.

R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

Paul R Rosenbaum. Observational studies. In *Observational studies*, pages 1–17. Springer, 2002.

Paul R Rosenbaum. *Design of Observational Studies*, volume 10. Springer, 2010.

Paul R Rosenbaum. Sensitivity analysis in observational studies. *Wiley StatsRef: Statistics Reference Online*, 2014.

Paul R Rosenbaum and Donald B Rubin. Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(2):212–218, 1983.

Mark Rowland, Marc G Bellemare, Will Dabney, Rémi Munos, and Yee Whye Teh. An analysis of categorical distributional reinforcement learning. *arXiv preprint arXiv:1802.08163*, 2018.

Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. 100(469): 322–331, 2005.

Michael Rutter, David Greenfeld, and Linda Lockyer. A five to fifteen year follow-up study of infantile psychosis: Ii. social and behavioural outcome. *British Journal of Psychiatry*, 113(504): 1183–1199, 1967. doi: 10.1192/bjp.113.504.1183.

Amir Sani, Alessandro Lazaric, and Rémi Munos. Risk-aversion in multi-armed bandits. In *Advances in Neural Information Processing Systems*, pages 3275–3283, 2012.

Piotr Sapiezynski, Valentin Kassarnig, and Christo Wilson. Academic performance prediction in a gender-imbalanced environment. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, volume 1, pages 48–51, 2017.

Makoto Sato, Hajime Kimura, and Shibenobu Kobayashi. Td algorithm for the variance of return and mean-variance reinforcement learning. *Transactions of the Japanese Society for Artificial Intelligence*, 16(3):353–362, 2001.

Christopher W. Seymour, Foster Gesten, Hallie C. Prescott, Marcus E. Friedrich, Theodore J. Iwashyna, Gary S. Phillips, Stanley Lemeshow, Tiffany Osborn, Kathleen M. Terry, and Mitchell M. Levy. Time to treatment and mortality during mandated emergency care for sepsis. *New England Journal of Medicine*, 376(23):2235–2244, 2017. doi: 10.1056/NEJMoa1703058. URL `https://doi.org/10.1056/NEJMoa1703058`. PMID: 28528569.

Christopher W Seymour, Jason N Kennedy, Shu Wang, Chung-Chou H Chang, Corrine F Elliott, Zhongying Xu, Scott Berry, Gilles Clermont, Gregory Cooper, Hernando Gomez, et al. Derivation, validation, and potential treatment implications of novel clinical phenotypes for sepsis. *Jama*, 321 (20):2003–2017, 2019.

Uri Shalit, Fredrik D Johansson, and David Sontag. Estimating individual treatment effect: generalization bounds and algorithms. In *International Conference on Machine Learning*, pages 3076–3085. PMLR, 2017.

Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

Weiwei Shen, Jun Wang, Yu-Gang Jiang, and Hongyuan Zha. Portfolio choices with orthogonal bandit learning. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, Laurent Sifre, Dharshan Kumaran, Thore Graepel, et al. A general reinforcement learning algorithm that masters chess, shogi, and go through self-play. *Science*, 362(6419):1140–1144, 2018.

Sarah A Sterling, W Ryan Miller, Jason Pryor, Michael A Puskarich, and Alan E Jones. The impact of timing of antibiotics on outcomes in severe sepsis and septic shock: a systematic review and meta-analysis. *Critical care medicine*, 43(9):1907, 2015.

Alexander L Strehl and Michael L Littman. An analysis of model-based interval estimation for markov decision processes. *Journal of Computer and System Sciences*, 74(8):1309–1331, 2008.

Aviv Tamar and Shie Mannor. Variance adjusted actor critic algorithms. *arXiv preprint arXiv:1310.3697*, 2013.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems*, pages 1468–1476, 2015a.

Aviv Tamar, Yonatan Glassner, and Shie Mannor. Optimizing the cvar via sampling. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015b.

Rrachel Tatman. Gender and dialect bias in YouTube's automatic captions. In *First Workshop on Ethics in Natural Langauge Processing*, volume 1, pages 53–59, 2017.

Guy Tennenholtz, Shie Mannor, and Uri Shalit. Off-policy evaluation in partially observable environments. 2020.

Philip Thomas and Emma Brunskill. Data-efficient off-policy policy evaluation for reinforcement learning. In *International Conference on Machine Learning*, pages 2139–2148, 2016.

Philip Thomas and Erik Learned-Miller. Concentration inequalities for conditional value at risk. In *International Conference on Machine Learning*, pages 6225–6233, 2019.

Philip S Thomas, Georgios Theocharous, and Mohammad Ghavamzadeh. High-confidence off-policy evaluation. In *In Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

Philip S Thomas, Bruno Castro da Silva, Andrew G Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. Preventing undesirable behavior of intelligent machines. *Science*, 366(6468):999–1004, 2019.

Amos Tversky and Daniel Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and uncertainty*, 5(4):297–323, 1992.

Sattar Vakili and Qing Zhao. Mean-variance and value at risk in multi-armed bandit problems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1330–1335. IEEE, 2015.

Sattar Vakili and Qing Zhao. Risk-averse multi-armed bandit problems under mean-variance measure. *IEEE Journal of Selected Topics in Signal Processing*, 10(6):1093–1111, 2016.

Emiliano A Valdez. On tail conditional variance and tail covariances. *UNSW Actuarial Studies, Sydney*, 2004.

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.

Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in starcraft ii using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

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.

John White. *Bandit algorithms for website optimization*. " O'Reilly Media, Inc.", 2012.

Markus Wübben and Florian v Wangenheim. Instant customer base analysis: Managerial heuristics often "get it right". *Journal of Marketing*, 72(3):82–93, 2008.

Tengyang Xie, Yifei Ma, and Yu-Xiang Wang. Towards optimal off-policy evaluation for reinforcement learning with marginalized importance sampling. pages 9665–9675, 2019.

Yu Xie, Jennie E Brand, and Ben Jann. Estimating heterogeneous treatment effects with observational data. *Sociological methodology*, 42(1):314–347, 2012.

Steve Yadlowsky, Hongseok Namkoong, Sanjay Basu, John Duchi, and Lu Tian. Bounds on the conditional and average treatment effect with unobserved confounding factors. *arXiv preprint arXiv:1808.09521*, 2018.

Tianhe Yu, Garrett Thomas, Lantao Yu, Stefano Ermon, James Zou, Sergey Levine, Chelsea Finn, and Tengyu Ma. Mopo: Model-based offline policy optimization. *arXiv preprint arXiv:2005.13239*, 2020.

Junzhe Zhang and Elias Bareinboim. Near-optimal reinforcement learning in dynamic treatment regimes. 2019.

Shushang Zhu and Masao Fukushima. Worst-case conditional value-at-risk with application to robust portfolio management. *Operations research*, 57(5):1155–1168, 2009.

Alexander Zimin, Rasmus Ibsen-Jensen, and Krishnendu Chatterjee. Generalized risk-aversion in stochastic multi-armed bandits. *arXiv preprint arXiv:1405.0833*, 2014.

Appendix A

Supplementary Materials for Chapter 2

In this chapter we provide technical proofs for materials presented in chapter 2. First we start by proving some helper lemmas.

A.1 Proof of basic lemmas

Before we give the proof of our main results, we give a set of essentially standard lemmas that we build on in the rest of the chapter.

Lemma 4. *Let $Z, \tilde{Z} \in \mathcal{Z}$ be such that $F_{\tilde{Z}(s,a)} \leq F_{Z(s,a)}$ for all (s,a). Assume finitely many states $|\mathcal{S}| < \infty$ and actions $|\mathcal{A}| < \infty$. Let $\hat{R}(s,a)$ and $\hat{P}(s,a)$ be the empirical reward distributions and transition probabilities. Assume further that*

$$\tilde{c}(s,a) \geq \|F_{\hat{R}(s,a)} - F_{R(s,a)}\|_\infty + \sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s,a) - P(s'|s,a)\right) \sum_{a' \in \mathcal{A}} \pi(a'|s') F_{\gamma Z(s',a') + R(s,a)}(x) \tag{A.1}$$

where $\tilde{c}(s,a) = \frac{c}{\sqrt{n(s,a)}}$ is the shift in the optimism operator O_c. Then $F_{O_c \hat{\mathcal{T}}^\pi \tilde{Z}(s,a)} \leq F_{\mathcal{T}^\pi Z(s,a)}$ for all (s,a). Note that it is sufficient to pick $\tilde{c}(s,a) \geq \|F_{\hat{R}(s,a)} - F_{R(s,a)}\|_\infty + \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1$ to ensure the Assumption in Eq. (Equation A.1).

Lemma 4. We start with some basic identities which follow directly from the definition of CDFs

$$F_{P^\pi Z(s,a)}(x) = \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} P(s'|s,a)\pi(a'|s') F_{Z(s',a')}(x) \tag{A.2}$$

$$F_{\gamma Z(s,a)}(x) = F_{Z(s,a)}(x/\gamma) \tag{A.3}$$

$$F_{R(s,a)+Z(s,a)}(x) = \int F_{Z(s,a)}(x-y)dF_{R(s,a)}(y) = \int F_{R(s,a)}(x-y)dF_{Z(s,a)}(y) \tag{A.4}$$

where the integrals are Lebesgue-Stieltjes integrals and understood to be taken over $[V_{\min}, V_{\max}]$. We omit the limits in the following to unclutter notation. These identities allow us to derive expressions for $F_{\mathcal{T}^\pi Z(s,a)}$ and $F_{O_c\hat{\mathcal{T}}^\pi Z'(s,a)}$:

$$\begin{aligned}
F_{\mathcal{T}^\pi Z(s,a)}(x) &= F_{R(s,a)+\gamma P^\pi Z(s,a)}(x) && \text{(A.5)}\\
&= \int \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} P(s'|s,a)\pi(a'|s')F_{\gamma Z(s',a')}(x-y)\,dF_{R(s,a)}(y) && \text{(A.6)}\\
&= \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} P(s'|s,a)\pi(a'|s')\int F_{\gamma Z(s',a')}(x-y)\,dF_{R(s,a)}(y) && \text{(A.7)}\\
&= \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} P(s'|s,a)\pi(a'|s')F_{\gamma Z(s',a')+R(s,a)}(x) && \text{(A.8)}
\end{aligned}$$

$$\begin{aligned}
F_{O_c\hat{\mathcal{T}}^\pi\tilde{Z}(s,a)}(x) &= F_{O_c(\hat{R}(s,a)+\gamma\hat{P}^\pi\tilde{Z}(s,a))}(x) && \text{(A.9)}\\
&= 0 \vee \left(\int \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \hat{P}(s'|s,a)\pi(a'|s')F_{\gamma\tilde{Z}(s',a')}(x-y)\,dF_{\hat{R}(s,a)}(y) - \tilde{c}(s,a)\right) && \text{(A.10)}\\
&= 0 \vee \left(\sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \hat{P}(s'|s,a)\pi(a'|s')F_{\gamma\tilde{Z}(s',a')+\hat{R}(s,a)}(x) - \tilde{c}(s,a)\right) && \text{(A.11)}
\end{aligned}$$

Here we exchanged the finite sum with the integral by linearity of integrals. Using these identities, we will show that $F_{O_c\hat{\mathcal{T}}^\pi\tilde{Z}(s,a)}(x) - F_{\mathcal{T}^\pi Z(s,a)}(x) \leq 0$ for all x. Consider any fixed x and first the case where the max in $F_{O_c\hat{\mathcal{T}}^\pi\tilde{Z}(s,a)}(x)$ is attained by 0. In this case $F_{O_c\hat{\mathcal{T}}^\pi\tilde{Z}(s,a)}(x) - F_{\mathcal{T}^\pi Z(s,a)}(x) = -F_{\mathcal{T}^\pi Z(s,a)}(x) \leq 0$ because CDFs take values in $[0,1]$. For the second case, we combine (Equation A.8) and (Equation A.11) to write

$$\begin{aligned}
&F_{O_c\hat{\mathcal{T}}^\pi\tilde{Z}(s,a)}(x) - F_{\mathcal{T}^\pi Z(s,a)}(x)\\
&\quad= \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \hat{P}(s'|s,a)\pi(a'|s')F_{\gamma\tilde{Z}(s',a')+\hat{R}(s,a)}(x) - \tilde{c}(s,a)\\
&\quad- \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} P(s'|s,a)\pi(a'|s')F_{\gamma Z(s',a')+R(s,a)}(x)\\
&\quad= -\tilde{c}(s,a)\\
&\quad+ \sum_{s'\in\mathcal{S},a'\in\mathcal{A}} \hat{P}(s'|s,a)\pi(a'|s')\left(F_{\gamma\tilde{Z}(s',a')+R(s,a)}(x) - F_{\gamma Z(s',a')+R(s,a)}(x)\right) && \text{(A.12)}
\end{aligned}$$

$$+ \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \hat{P}(s'|s,a)\pi(a'|s') \left(F_{\gamma \tilde{Z}(s',a') + \hat{R}(s,a)}(x) - F_{\gamma \tilde{Z}(s',a') + R(s,a)}(x) \right) \tag{A.13}$$

$$+ \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \pi(a'|s') \left(\hat{P}(s'|s,a) - P(s'|s,a) \right) F_{\gamma Z(s',a') + R(s,a)}(x). \tag{A.14}$$

In the following, we consider each of the terms (Equation A.12)–(Equation A.14) separately. Let us start with (Equation A.12) and bound

$$\begin{aligned} &F_{\gamma \tilde{Z}(s',a') + R(s,a)}(x) - F_{\gamma Z(s',a') + R(s,a)}(x) \\ &= \int \left[F_{\tilde{Z}(s,a)} \left(\frac{x-y}{\gamma} \right) - F_{Z(s,a)} \left(\frac{x-y}{\gamma} \right) \right] dF_{R(s,a)}(y) \leq \int 0 dF_{R(s,a)}(y) = 0 \end{aligned} \tag{A.15}$$

where the identity follows from the basic identities in Eq. (Equation A.2) and the inequality from the assumption that $F_{\tilde{Z}(s,a)} \leq F_{Z(s,a)}$ for all (s,a). Hence, the term in Eq. (Equation A.12) is always non-positive. Moving on to the term in Eq. (Equation A.13) which we bound with similar tools as

$$F_{\gamma \tilde{Z}(s',a') + \hat{R}(s,a)}(x) - F_{\gamma \tilde{Z}(s',a') + R(s,a)}(x) \tag{A.16}$$

$$= \int \left[F_{\hat{R}(s,a)}(x-y) - F_{R(s,a)}(x-y) \right] dF_{\gamma \tilde{Z}(s,a)}(y) \tag{A.17}$$

$$\leq \int \left| F_{\hat{R}(s,a)}(x-y) - F_{R(s,a)}(x-y) \right| dF_{\gamma \tilde{Z}(s,a)}(y) \tag{A.18}$$

$$\leq \sup_z \left| F_{\hat{R}(s,a)}(z) - F_{R(s,a)}(z) \right| \int dF_{\gamma \tilde{Z}(s,a)}(y) \tag{A.19}$$

$$= \left\| F_{\hat{R}(s,a)} - F_{R(s,a)} \right\|_\infty. \tag{A.20}$$

This yields that Eq. (Equation A.13) is bounded by this as well

$$\sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} \hat{P}(s'|s,a)\pi(a'|s') \left\| F_{\hat{R}(s,a)} - F_{R(s,a)} \right\|_\infty = \left\| F_{\hat{R}(s,a)} - F_{R(s,a)} \right\|_\infty. \tag{A.21}$$

Finally, the last term from Eq. (Equation A.14) can be bounded as follows

$$\sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s,a) - P(s'|s,a) \right) \sum_{a' \in \mathcal{A}} \pi(a'|s') F_{\gamma Z(s',a') + R(s,a)}(x) \tag{A.22}$$

$$\leq \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \left\| \sum_{a' \in \mathcal{A}} \pi(a'|\cdot) F_{\gamma Z(s',a') + R(s,a)}(x) \right\|_\infty \tag{A.23}$$

$$\leq \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \left\| \sum_{a' \in \mathcal{A}} \pi(a'|\cdot) \times 1 \right\|_\infty \tag{A.24}$$

$$\leq \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \tag{A.25}$$

Combining the individual bounds for each of the terms (Equation A.12)–(Equation A.14), we end

up with

$$
\begin{aligned}
& F_{O_c \hat{\mathcal{T}}^\pi \tilde{Z}(s,a)}(x) - F_{\mathcal{T}^\pi Z(s,a)}(x) && \text{(A.26)}\\
\leq & -\tilde{c}(s,a) + \|F_{\hat{R}(s,a)} - F_{R(s,a)}\|_\infty \\
& + \sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s,a) - P(s'|s,a)\right) \sum_{a' \in \mathcal{A}} \pi(a'|s') F_{\gamma Z(s',a') + R(s,a)}(x) && \text{(A.27)}\\
\leq & -\tilde{c}(s,a) + \|F_{\hat{R}(s,a)} - F_{R(s,a)}\|_\infty + \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1, && \text{(A.28)}
\end{aligned}
$$

which is non-positive as long as $\tilde{c}(s,a) \geq \|F_{\hat{R}(s,a)} - F_{R(s,a)}\|_\infty + \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1$ which completes the proof. □

Lemma 5. *Let F be a CDF of a bounded non-negative random variable and $\nu \in \mathbb{R}$ be arbitrary. Then $\mathbb{E}_F[(\nu - X)^+] = \int_0^\nu F(y)dy$. Hence, one can write the conditional value at risk of a variable $X \sim F$ for any CDF F with $F(0) = 0$ as*

$$
\mathrm{CVaR}_\alpha(F) = \sup_\nu \left\{ \frac{1}{\alpha} \int_0^\nu (\alpha - F(y))dy \right\}. \tag{A.29}
$$

Lemma 5. We rewrite $\mathbb{E}_F[(\nu - X)^+]$ as follows

$$
\begin{aligned}
\mathbb{E}_F[(\nu - X)^+] =& \mathbb{E}_F[(\nu - X)\mathbf{1}\{X \leq \nu\}] = \nu F(\nu) - \mathbb{E}_F[X\mathbf{1}\{X \leq \nu\}] && \text{(A.30)}\\
\overset{①}{=}& \nu F(\nu) - \mathbb{E}_F\left[\mathbf{1}\{X \leq \nu\} \int_0^\infty \mathbf{1}\{X > y\}dy\right] && \text{(A.31)}\\
\overset{②}{=}& \nu F(\nu) - \int_0^\infty \mathbb{P}_F\left[y < X \leq \nu\right] dy && \text{(A.32)}\\
=& \nu F(\nu) - \int_0^\nu (F(\nu) - F(y))dy = \int_0^\nu F(y)dy && \text{(A.33)}
\end{aligned}
$$

where ① follows from $a = \int_0^a dx = \int_0^\infty \mathbf{1}\{a > x\}dx$ which holds for any $a \geq 0$ and ② uses Tonelli's theorem to exchange the two integrals. Plugging this identity into

$$
\nu - \frac{1}{\alpha}\mathbb{E}_F[(\nu - X)^+] = \frac{1}{\alpha}\left(\nu\alpha - \int_0^\nu F(y)dy\right) = \frac{1}{\alpha}\int_0^\nu (\alpha - F(y))dy \tag{A.34}
$$

and taking the sup over ν gives the desired result. □

Lemma 6. *Let F and G be the CDFs of two non-negative random variables and let ν_F, ν_G be a maximizing value of ν in the definition of $\mathrm{CVaR}_\alpha(F)$ and $\mathrm{CVaR}_\alpha(G)$ in Equation 2.1 respectively. Then:*

$$
|\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G)| \leq \frac{1}{\alpha}\int_0^{\max\{F^{-1}(\alpha), G^{-1}(\alpha)\}} |G(y) - F(y)|dy \tag{A.35}
$$

$$\leq \frac{\max\{F^{-1}(\alpha), G^{-1}(\alpha)\}}{\alpha} \sup_x |F(x) - G(x)| \tag{A.36}$$

Lemma 6. Assume w.l.o.g. that $\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G) \geq 0$. Denote by ν_F any maximizing value of ν in the definition of $\mathrm{CVaR}_\alpha(F)$. By Lemma 4.2 and Equation (4.9) in Acerbi and Tasche [2002], a possible value of ν_F is $F^{-1}(\alpha)$. Then we can write the differences in CVaR as

$$\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G) \leq \nu_F - \alpha^{-1}\mathbb{E}_F[(\nu_F - X)^+] - (\nu_F - \alpha^{-1}\mathbb{E}_G[(\nu_F - X)^+]) \tag{A.37}$$

$$= \frac{1}{\alpha}(\mathbb{E}_G[(\nu_F - X)^+] - \mathbb{E}_F[(\nu_F - X)^+]). \tag{A.38}$$

Using Lemma 5 in Equation (Equation A.38) gives

$$\mathrm{CVaR}_\alpha(F) - \mathrm{CVaR}_\alpha(G) \leq \frac{1}{\alpha}\left(\int_0^{\nu_F} G(y)dy - \int_0^{\nu_F} F(y)dy\right) \tag{A.39}$$

$$\leq \frac{1}{\alpha}\int_0^{\nu_F} |G(y) - F(y)|dy \leq \frac{\nu_F}{\alpha} \sup_y |F(y) - G(y)|. \tag{A.40}$$

We can in full analogy upper-bound $\mathrm{CVaR}_\alpha(G) - \mathrm{CVaR}_\alpha(F)$ and arrive at the desired statement. □

Lemma 7. *Let G and F be CDFs of non-negative random variables so that $\forall x \geq 0: \quad F(x) \geq G(x)$. Then for any $\alpha \in [0, 1]$, we have $\mathrm{CVaR}_\alpha(F) \leq \mathrm{CVaR}_\alpha(G)$.*

Proof. Consider now the following difference

$$\frac{1}{\alpha}\int_0^{\nu} (\alpha - G(y))dy - \frac{1}{\alpha}\int_0^{\nu} (\alpha - F(y))dy = \frac{1}{\alpha}\int_0^{\nu} (F(y) - G(y))dy \geq 0. \tag{A.41}$$

By Lemma 5, we have that

$$\mathrm{CVaR}_\alpha(G) - \mathrm{CVaR}_\alpha(F) \tag{A.42}$$

$$= \sup_\nu \left\{\frac{1}{\alpha}\int_0^{\nu} (\alpha - G(y))dy\right\} - \sup_\nu \left\{\frac{1}{\alpha}\int_0^{\nu} (\alpha - F(y))dy\right\}. \tag{A.43}$$

Let ν_F denote a value of ν that achieves the supremum in $\frac{1}{\alpha}\int_0^{\nu}(\alpha - F(y))dy$ (which exists by Lemma 4.2 and Equation (4.9) in Acerbi and Tasche [2002]). Then

$$\mathrm{CVaR}_\alpha(G) - \mathrm{CVaR}_\alpha(F) \geq \frac{1}{\alpha}\int_0^{\nu_F} (\alpha - G(y))dy - \frac{1}{\alpha}\int_0^{\nu_F} (\alpha - F(y))dy \geq 0. \tag{A.44}$$

□

Lemma 8. *Consider a sequence of CDFs $\{F_n\}$ on $x \in [V_{\min}, V_{\max}]$ with $V_{\min} \geq 0$ that converges in ℓ_2 distance to F_O as $n \to \infty$. Then $\mathrm{CVaR}_\alpha(F_n) \to \mathrm{CVaR}_\alpha(F_O)$ as $n \to \infty$.*

Proof. Consider the sequence of Wasserstein d_1 distance between the CDFs:

$$d_1\left(F_n, F_{Z_O}\right) = \int_{V_{\min}}^{V_{\max}} |F_n(x) - F_O(x)| dx \tag{A.45}$$

$$\leq \left(\int_{V_{\min}}^{V_{\max}} |F_n(x) - F_O(x)|^2 dx\right)^{1/2} \left(\int_{V_{\min}}^{V_{\max}} 1 dx\right)^{1/2} \tag{A.46}$$

$$= \sqrt{V_{\max} - V_{\min}}\ell_2(F_n, F_O) \tag{A.47}$$

Where the second line follows by Hölder's inequality. The right hand side goes to 0 as $n \to \infty$, which implies convergence in Wasserstein d_1 distance (p=1). Finally, using Lemma 6, convergence of CVaR follows. □

Lemma 9 (Difference in CVaR)**.** *Let F be the CDF of a random variable bounded by $[0, U]$ and $\hat{F}$ be the empirical CDF obtained by n, i.i.d samples drawn from F. Let $\epsilon > 0$ and $\mathcal{G} = \left\{\sup_x |F(x) - \hat{F}(x)| \leq \epsilon\right\}$ be the event that the empirical CDF is uniformly ϵ-close to the true CDF F. Define $\tilde{F}(x) = 0 \vee (\hat{F}(x) - \epsilon \mathbf{1}\{x \in [0, U]\})$. Then in event $\mathcal{G}$ the following inequality holds*

$$|\operatorname{CVaR}_\alpha(F) - \operatorname{CVaR}_\alpha(\tilde{F})| \leq \frac{2\tilde{F}^{-1}(\alpha)\epsilon}{\alpha}. \tag{A.48}$$

Proof. By Lemma 6, the triangle-inequality and the definition of $\mathcal{G}$ and $\tilde{F}$

$$|\operatorname{CVaR}_\alpha(F) - \operatorname{CVaR}_\alpha(\tilde{F})| \leq \frac{\tilde{F}^{-1}(\alpha)}{\alpha} \sup_x |F(x) - \tilde{F}(x)| \tag{A.49}$$

$$\leq \frac{\tilde{F}^{-1}(\alpha)}{\alpha} \sup_x |F(x) - \hat{F}(x)| + \frac{\tilde{F}^{-1}(\alpha)}{\alpha} \sup_x |\hat{F}(x) - \tilde{F}(x)| \tag{A.50}$$

$$\leq \frac{2\tilde{F}^{-1}(\alpha)\epsilon}{\alpha}. \tag{A.51}$$

□

Lemma 10 (Down-shift is optimistic for CVaR)**.** *Let F be the CDF of a random variable bounded by $[0, U]$ and $\hat{F}$ be the empirical CDF obtained by n, i.i.d samples drawn from F. Let $\epsilon > 0$ and $\mathcal{G} = \left\{\sup_x |F(x) - \hat{F}(x)| \leq \epsilon\right\}$ be the event that the empirical CDF is uniformly ϵ-close to the true CDF F. Define $\tilde{F}(x) = 0 \vee (\hat{F}(x) - \epsilon \mathbf{1}\{x \in [0, U]\})$. Then in event $\mathcal{G}$ the following inequality holds*

$$\operatorname{CVaR}_\alpha(F) \leq \operatorname{CVaR}_\alpha(\tilde{F}) \tag{A.52}$$

Proof. By construction of $\tilde{F}$, we have $\tilde{F}(x) \leq F(x)$ for all x on $\mathcal{G}$ and hence the statement follows by Lemma 7. □

A.2 Proof of Theorem 1

Theorem (1). *Let the shift parameter in the optimistic operator be sufficiently large which is* $c = O\left(\ln(|\mathcal{S}||\mathcal{A}|/\delta)\right)$. *Then with probability at least* $1-\delta$, *the iterates* $\mathrm{CVaR}_\alpha((O_c\hat{\mathcal{T}}^\pi)^m Z_0)$ *converges for any risk level* α *and initial* $Z_0 \in \mathcal{Z}$ *to an optimistic estimate of the policy's conditional value at risk. That is, with probability at least* $1-\delta$,

$$\forall s,a : \mathrm{CVaR}_\alpha((O_c\hat{\mathcal{T}}^\pi)^\infty Z_0(s,a)) \geq \mathrm{CVaR}_\alpha(Z_\pi(s,a)).$$

Theorem 1. By Lemma 7 and Lemma 3 by Bellemare et al. [2017], we know that $Z_{i+1} \leftarrow O_c\hat{\mathcal{T}}^\pi Z_i$ converges to a unique fixed-point Z_∞, independent of the initial Z_0. Hence, without loss of generality, we can choose $Z_0 = Z_\pi$.

We proceed by first showing how our result will follow under a particular definition of c, and then show what that definition is. Assume that we have obtained a value for c that satisfies the assumption of Lemma 4, and let $\tilde{Z} = Z_i$ and $Z = Z_\pi$. Then Lemma 4 implies that if $F_{Z_i(s,a)} \leq F_{Z_\pi(s,a)}$ for all (s,a), then also $F_{Z_{i+1}(s,a)} \leq F_{Z_\pi(s,a)}$ for all (s,a). Thus, $F_{Z_\infty(s,a)} \leq F_{Z_\pi(s,a)}$ for all (s,a). Finally, we can use Lemma 7 to obtain the desired result of our proof statement, $\mathrm{CVaR}_\alpha(F_{Z_\infty(s,a)}) \geq \mathrm{CVaR}_\alpha(F_{Z_\pi(s,a)})$ for all (s,a).

Going back, we use concentration inequalities to determine the value of c that ensures the required condition in Lemma 4 (expressed in Eq. (Equation A.1)). The DKW-inequality which give us that for any (s,a) with probability at least $1-\delta$

$$\|F_{R(s,a)} - F_{\hat{R}(s,a)}\|_\infty \leq \sqrt{\frac{1}{2n(s,a)}\ln\frac{2}{\delta}}. \tag{A.53}$$

Further, the inequality by Weissman et al. [2003] gives that

$$\|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1 \leq \sqrt{\frac{2|\mathcal{S}|}{n(s,a)}\ln\frac{2}{\delta}}. \tag{A.54}$$

Combining both with a union bound over all $|\mathcal{S}\times\mathcal{A}|$ state-action pairs, we get that it is sufficient to choose

$$c = \sqrt{(1+4|\mathcal{S}|)\ln(4|\mathcal{S}||\mathcal{A}|/\delta)} \geq \sqrt{2|\mathcal{S}|\ln(4|\mathcal{S}||\mathcal{A}|/\delta)} + \sqrt{\ln(4|\mathcal{S}||\mathcal{A}|/\delta)}/2 \tag{A.55}$$

to ensure that $\tilde{c}(s,a) \geq \|F_{\hat{R}(s,a)} - F_{R(s,a)}\|_\infty + \|\hat{P}(\cdot|s,a) - P(\cdot|s,a)\|_1$ allowing us to apply Lemma 4.

However, we can improve this result by removing the polynomial dependency of c on the number of states $|\mathcal{S}|$ as follows. Consider a fixed (s,a) and denote

$$v(s', x) := \sum_{a' \in \mathcal{A}} \pi(a'|s') F_{\gamma Z(s',a')+R(s,a)}(x)$$

where we set $Z = Z_\pi$. Our goal is to derive a concentration bound on

$$\sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s, a) - P(s'|s, a) \right) v(s', x)$$

that is tighter than the bound derived from $\|\hat{P}(\cdot|s, a) - P(\cdot|s, a)\|_1$. Note that v is not a random quantity and hence $\sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s, a) - P(s'|s, a) \right) v(s', x)$ is a normalized sum of independent random variables for any x. To deal with the continuous variable x which prevents us from applying a union bound over x directly, we use a covering argument. Let $K \in \mathbb{N}$ be arbitrary and consider the discretization set

$$\bar{\mathcal{X}} = \{x \in \mathbb{R} \mid \exists k \in [K],\ \exists s' \in \mathcal{S},\ \forall x' < x,\ v(s', x') < k/K \leq v(s', x)\}. \tag{A.56}$$

Define $\bar{v}(s', x) = v(s', \max\{x' \in \bar{\mathcal{X}} \ : \ x' \leq x\})$ as the discretization of v at the discretization points in $\bar{\mathcal{X}}$. This construction ensures that the discretization error is uniformly bounded by $1/K$, that is, $|\bar{v}(s, x) - v(s, x)| \leq 1/K$ holds for all $s \in \mathcal{S}$ and $x \in [V_{\min}, V_{\max}]$. Hence, we can bound for all $x \in [V_{\min}, V_{\max}]$

$$\sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s, a) - P(s'|s, a) \right) v(s', x) \tag{A.57}$$

$$= \sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s, a) - P(s'|s, a) \right) \bar{v}(s', x) \tag{A.58}$$

$$+ \sum_{s' \in \mathcal{S}} \left(\hat{P}(s'|s, a) - P(s'|s, a) \right) (v(s', x) - \bar{v}(s', x)) \tag{A.59}$$

$$\overset{①}{\leq} \sqrt{\frac{1}{2n(s, a)} \ln \frac{|\bar{\mathcal{X}}|}{\delta}} + \|\hat{P}(\cdot|s, a) - P(\cdot|s, a)\|_1 \|v(\cdot, x) - \bar{v}(\cdot, x)\|_\infty \tag{A.60}$$

$$\overset{②}{\leq} \sqrt{\frac{1}{2n(s, a)} \ln \frac{|\bar{\mathcal{X}}|}{\delta}} + \frac{1}{K} \sqrt{\frac{2|\mathcal{S}|}{n(s, a)} \ln \frac{2}{\delta}} \tag{A.61}$$

where in ①, we applied Hoeffding's inequality to the first term in combination with a union bound over $\bar{\mathcal{X}}$ as $\bar{v}(\cdot, x)$ can only take $|\bar{\mathcal{X}}|$ values in $\mathbb{R}^{|\mathcal{S}|}$. The second term was bounded with Hölder's inequality and in ② the concentration inequality from Eq. (Equation A.54) was used. Combining this bound with Eq. (Equation A.53) by applying a union bound over all states and actions, we get that picking

$$c \geq \sqrt{\frac{1}{2} \ln \frac{3|\mathcal{S}||\mathcal{A}|}{\delta}} + \sqrt{\frac{1}{2} \ln \frac{3|\mathcal{S}||\mathcal{A}||\mathcal{X}|}{\delta}} + \sqrt{\frac{2|\mathcal{S}|}{K^2} \ln \frac{6|\mathcal{S}||\mathcal{A}|}{\delta}} \tag{A.62}$$

is sufficient to apply Lemma 4. Since $v(s', \cdot)$ is non-decreasing, the size of the discretization set is at most $|\tilde{\mathcal{X}}| \leq |\mathcal{S}|K$ and by picking $K = \sqrt{|\mathcal{S}|}$, we see that $c = O\left(\ln(|\mathcal{S}||\mathcal{A}|/\delta)\right)$ is sufficient.

□

A.3 CVaR-Bandit

In this section we prove technical theorems presented in section 2.7.2 and present a new Bandit algorithm, CVaR-Brown.

A.3.1 Proof of Theorem 2

Theorem (2). *[DKW-UCB regret bound] Consider DKW-UCB on a stochastic k-armed bandit problem with bounded rewards in range $[0, U]$. For any horizon n, if $\delta = 1/n^2$ then the expected CVaR-regret after the nth timestep is bounded by*

$$R_n^\alpha \leq \sum_{i=1}^{K} \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta_i^\alpha} + 3\sum_{i=1}^{K} \Delta_i^\alpha$$

Additionally the CVaR-regret after the nth timsestep is also bounded by,

$$R_n^\alpha \leq 4\sqrt{nk\ln(\sqrt{2}n)}\frac{U}{\alpha} + 3\sum_{i}^{k} \Delta_i^\alpha$$

Proof. Our proof closely follows the proof of UCB from Lattimore and Szepesvári [2018]. Let c_i^α denote the CVaR of arm i and $\hat{F}_{i,t}$ denote the empirical CDF of the ith arm before timestep t. Define $\tilde{c}_i^\alpha(t)$ as

$$\tilde{c}_i^\alpha(t) = \text{CVaR}_\alpha(\tilde{F}_{i,t})$$

Where $\tilde{F}_{i,t}$ is defined as follows,

$$\tilde{F}_{i,t}(x) = \left(\hat{F}_{i,t} - \sqrt{\frac{\ln(2/\delta)}{2T_i(t)}}\mathbf{1}\{x \in [0, U]\}\right)^+$$

$$\epsilon_i(t) = \frac{U}{\alpha}\sqrt{\frac{2\ln(2/\delta)}{T_i(t)}}$$

First observe that CVaR decomposes as $R_n^\alpha = \sum_{i=1}^{K} \Delta_i^\alpha \mathbb{E}[T_i(n)]$. We want to bound $\mathbb{E}[T_i(n)]$ for each suboptimal arm i. Without loss of generality we assume arm 1 is the optimal arm. Define the "good event" G_i as:

$$G_i = \{c_1^\alpha \leq \min_{t\in[n]} \tilde{c}_1^\alpha(t)\} \cap \bigcup_{i\in[K]} \{\tilde{c}_i^\alpha(u_i) \leq c_1^\alpha\}$$

We chose $u_i \in [n]$ later. Following Lattimore and Szepesvári [2018] we can show by contradiction that if G_i then $T_i(n) \leq u_i$. First, since $T_i(n) \leq n$ we can write:

$$\mathbb{E}[T_i(n)] = \mathbb{E}[T_i(n)\mathbb{I}\{G_i\}] + \mathbb{E}[T_i(n)\mathbb{I}\{G_i^c\}] \leq u_i + \mathbb{P}(G_i^c)n \tag{A.63}$$

Suppose $T_i(n) > u_i$ on event G_i, that means arm i was played more than u_i times over n rounds and so there must be a round $t \in [n]$ where $T_i(t-1) = u_i$ and $A_t = i$.

$$\begin{aligned}
\tilde{c}_i^\alpha &= \text{CVaR}_\alpha \left(\hat{F}_{i,t-1} - \sqrt{\frac{\ln(2/\delta)}{2T_i(t-1)}} \right) \\
&= \text{CVaR}_\alpha \left(\hat{F}_{i,u_i} - \sqrt{\frac{\ln(2/\delta)}{2u_i}} \right) \\
&< c_1^\alpha \\
&< \tilde{c}_1^\alpha(t-1)
\end{aligned}$$

Where the second line follows by $T_i(t-1) = u_i$ and the third and the fourth follows by the definition of event G_i. Hence $A_t = \arg\max_j \tilde{c}_j^\alpha \neq i$, which is a contradiction, so when G_i occurs $T_i(n) \leq u_i$. It is left to show the probably of the complement of the good event is low. Consider G_i^c

$$G_i^c = \{c_1^\alpha > \min_{t \in [n]} \tilde{c}_1^\alpha(t)\} \cup \{\tilde{c}_i^\alpha(u_i) > c_1^\alpha\} \tag{A.64}$$

and let us first consider the probability of the first part

$$\mathbb{P}\left(c_1^\alpha > \min_{t \in [n]} \tilde{c}_1^\alpha(t) \right) = \mathbb{P}\left(\exists t \in [n] \ : \ c_1^\alpha > \tilde{c}_1^\alpha(t) \right) \tag{A.65}$$

and using optimism as shown in Lemma 10 with $\epsilon = \sqrt{\frac{\ln(2/\delta)}{2T_1(t)}}$ we can upper bound this probability as

$$\leq \mathbb{P}\left(\exists t \in [n] \ : \ \sup_x |\hat{F}_{1,t}(x) - F_1(x)| > \sqrt{\frac{\ln(2/\delta)}{2T_1(t)}} \right) \tag{A.66}$$

and combining a union bound over the first $n \geq T_1(t)$ samples of arm 1 with the Dvoretzky-Kiefer-Wolfowitz inequality, we further bound this as

$$\leq n\delta. \tag{A.67}$$

For the second term of the failure event in Equation (Equation A.64), recall $\Delta_i^\alpha = c_1^\alpha - c_i^\alpha$, and we chose u_i such that $\Delta_i^\alpha \geq \epsilon_i(u_i)$

$$\mathbb{P}(\tilde{c}_i^\alpha(u_i) > c_1^\alpha) = \mathbb{P}(\tilde{c}_i^\alpha(u_i) - c_i^\alpha > \Delta_i^\alpha) \leq \mathbb{P}(\tilde{c}_i^\alpha(u_i) - c_i^\alpha > \epsilon_i(u_i)) \tag{A.68}$$

applying Lemma 9 and let t_i be the round at which arm i was observed the u_ith time

$$\leq \mathbb{P}\left(\sup_x |\hat{F}_{i,t_i}(x) - F_i(x)| > \sqrt{\frac{\ln(2/\delta)}{2u_i}}\right) \leq \delta \tag{A.69}$$

where the final bound follows from the Dvoretzky-Kiefer-Wolfowitz inequality. Hence, the probability of the failure event is bounded as $\mathbb{P}(G_i^c) \leq (n+1)\delta$. Substituting this bound into (Equation A.63):

$$\mathbb{E}[T_i(n)] \leq u_i + n(n+1)\delta \tag{A.70}$$

It remains to determine u_i which can be chosen as the first integer that satisfies $\Delta_i^\alpha \geq \epsilon_i(u_i)$:

$$u_i = \left\lceil \frac{2\ln(2/\delta)U^2}{\alpha^2 {\Delta_i^\alpha}^2} \right\rceil$$

Substituting into (Equation A.70), and choosing $\delta = \frac{1}{n^2}$:

$$\mathbb{E}[T_i(n)] \leq \left\lceil \frac{2\log(2n^2))U^2}{\alpha^2 {\Delta_i^\alpha}^2} \right\rceil + 2 \leq 3 + \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 {\Delta_i^\alpha}^2}$$

Substituting this into CVaR-regret decomposition, we get the desired bound

$$R_n^\alpha = \sum_{i=1}^k \Delta_i^\alpha \mathbb{E}[T_i(n)] \leq \sum_{i=1}^K \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta_i^\alpha} + 3\sum_{i=1}^K \Delta_i^\alpha$$

One can also prove a sublinear regret bound that does not depend on the reciprocal of the gaps.

$$\begin{aligned}
R_n^\alpha = \sum_{i=1}^k \Delta_i^\alpha \mathbb{E}[T_i(n)] &= \sum_{i:\Delta_i^\alpha < \Delta} \Delta_i^\alpha \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i^\alpha \geq \Delta} \Delta_i^\alpha \mathbb{E}[T_i(n)] \\
&\leq n\Delta + \sum_{i:\Delta_i^\alpha \geq \Delta} \Delta_i^\alpha \mathbb{E}[T_i(n)] \\
&\leq n\Delta + \sum_{i:\Delta_i^\alpha \geq \Delta} \left(3\Delta_i^\alpha + \frac{4\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta_i^\alpha}\right) \\
&\leq n\Delta + \frac{4k\ln(\sqrt{2}n)U^2}{\alpha^2 \Delta} + \sum_{i=1}^k 3\Delta_i^\alpha \\
&\leq 4\sqrt{nk\ln(\sqrt{2}n)}\frac{U}{\alpha} + 3\sum_{i=1}^k \Delta_i^\alpha \\
&\leq 4\frac{U}{\alpha}\sqrt{nk\ln(\sqrt{2}n)} + 3kU
\end{aligned}$$

Where the first inequality follows by $\sum_{i:\Delta_i^\alpha < \Delta} T_i(n) \leq n$, and the last line follows by choosing

Algorithm 3: Brown-UCB for MABs

Input: Risk level α, reward range U, max # of pulls n
Output: Series of actions $A_1, A_2, \dots A_n$.

Choose each arm once.
Initialize $t = 1$, Set $\delta = 1/n^2$.
for $t = 1, \dots, n$ **do**
  for $i = 1, \dots, k$ **do**
    $\text{UCB}_i^{\text{Brown}}(t) = \text{CVaR}_\alpha(\hat{F}_{i,t}) + U\sqrt{\frac{5\log(3/\delta)}{\alpha T_i(t)}}$;
  Play action $A_t = \text{argmax}_i \, \text{UCB}_i^\alpha(t)$;
  Update empirical CDF of arm A_i;

$\Delta = \frac{U}{\alpha}\sqrt{\frac{4k\ln(\sqrt{2}n)}{n}}$. □

A.3.2 Brown-UCB

The Brown-UCB algorithm presented in section 2.7.4 uses the upper confidence bound presented in Brown [2007]. Similar to Cassel et al. [2018] we compute the empirical CVaR of each arm and add an optimism bonus to build an upper confidence bound.

$$UCB_i^{\text{Brown}}(t) = \text{CVaR}_\alpha(\hat{F}_{i,t}) + U\sqrt{\frac{5\log(3/\delta)}{\alpha T_i(t)}}$$

we use $\delta = 1/n^2$. Algorithm 3 describes the algorithm.

Theorem 6 (Brown-UCB regret bound)**.** *Consider Brown-UCB on a stochastic k-armed bandit problem with bounded rewards in range $[0, U]$. For any horizon n, if $\delta = 1/n^2$ then the expected CVaR-regret after the nth timestep is bounded by*

$$R_n^\alpha \leq \sum_{i=1}^{K} \frac{40\ln(\sqrt{3}n)U^2}{\alpha\Delta_i^\alpha} + 3\sum_{i=1}^{K}\Delta_i^\alpha$$

Additionally the CVaR-regret after the nth timsestep is also bounded by,

$$R_n^\alpha \leq 4\sqrt{\frac{10kn\ln(\sqrt{3}n)}{\alpha}}U + 3KU$$

Proof. Our proof style mimics theorem A.3.1. Let c_i^α denote the CVaR of arm i, and $\hat{c}_i^\alpha(t)$ be the empirical CVaR of the ith arm before timestep t. First, we observe that the CVaR-regret decomposes as $R_n^\alpha = \sum_{i=1}^{k}\Delta_i^\alpha\mathbb{E}[T_i(n)]$. The general strategy we use is to bound $\mathbb{E}[T_i(n)]$ for each suboptimal arm i. To do this, we define the "good" event

$$G_i = \{c_1^\alpha < \min_{t\in[n]} \mathrm{UCB}_1^\alpha(t)\} \cap \{\mathrm{UCB}_i^\alpha(u_i) < c_1^\alpha\} \tag{A.71}$$

where $u_i \in [n]$ is a constant we will choose later. We need to show two things, first if G_i occurs, then $T_i(n) \leq u_i$. Second, The complement event G_i^c occurs with low probability (governed by our future choice of u_i). Because $T_i(n) \leq n$, we will have

$$\mathbb{E}[T_i(n)] = \mathbb{E}[\mathbb{I}\{G_i\}T_i(n)] + \mathbb{E}[\mathbb{I}\{G_i^c\}T_i(n)] \leq u_i + \mathbb{P}(G_i^c)n. \tag{A.72}$$

For the case where G_i is true, we can show by contradiction that $T_i(n) \leq u_i$, similar to Lattimore and Szepesvári [2018]. The next step is to upper bound $\mathbb{P}(G_i^c)$. By its definition,

$$G_i^c = \{c_1^\alpha \geq \min_{t\in[n]} \mathrm{UCB}_1^\alpha(t)\} \cup \{\mathrm{UCB}_i^\alpha(u_i) \geq c_1^\alpha\}$$

The first set can be decomposed into a union of inequalities $\bigcup_{t=1}^n \{c_1^\alpha \geq UCB_1^\alpha(t)\}$. We apply the concentration inequality from Brown [2007, Theorem 4.2]. Combining all probability bounds of individual events with a union bound, we bound the probability of $\{c_1^\alpha \geq \min_{t\in[n]} \mathrm{UCB}_1^\alpha(t)\}$ as $n\delta$.

$$\begin{aligned}
\mathbb{P}\left(c_1^\alpha \geq \min_{t\in[n]} UCB_1^\alpha(t)\right) &\leq \mathbb{P}\left(\bigcup_{t=1}^n \{c_1^\alpha \geq UCB_1^\alpha(t)\}\right) \\
&\leq \sum_{t=1}^n \left(\mathbb{P}(c_1^\alpha \geq \hat{c}_1^\alpha + U\sqrt{\frac{5\ln(3/\delta)}{\alpha t}}\right) \leq n\delta
\end{aligned}$$

Since the second event is contained in $\{UCB_i^\alpha(u_i) \geq c_1^\alpha\}$ we can simply apply the Brown [2007, Corollary 3.1] concentration inequality to the second set as well. Assume that u_i is chosen large enough that $\Delta_i^\alpha - U\sqrt{\frac{5\ln(3/\delta)}{\alpha u_i}} \geq c\Delta_i^\alpha$:

$$\begin{aligned}
\mathbb{P}\left(\hat{c}_i^\alpha + U\sqrt{\frac{5\ln(3/\delta)}{\alpha u_i}} \geq c_1^\alpha\right) &= \mathbb{P}\left(\hat{c}_i^\alpha - c_i^\alpha \geq \Delta_i^\alpha - U\sqrt{\frac{5\ln(3/\delta)}{\alpha u_i}}\right) \\
&\leq \mathbb{P}\left(\hat{c}_i^\alpha - c_i^\alpha \geq c\Delta_i^\alpha\right) \\
&\leq \exp\left(-2u_i\frac{\alpha^2(c\Delta_i^\alpha)^2}{U^2}\right)
\end{aligned}$$

Substituting into (Equation A.72), we obtain

$$\mathbb{E}[T_i(n)] \leq u_i + n\left(n\delta + \exp\left(-2u_i\frac{\alpha^2(c\Delta_i^\alpha)^2}{U^2}\right)\right)$$

Choosing $u_i = \left\lceil \frac{5U^2\ln(3/\delta)}{\alpha(1-c)^2(\Delta_i^\alpha)^2} \right\rceil$ and $c = \frac{1}{2}$ then yields

$$\mathbb{E}[T_i(n)] \leq \frac{40U^2\ln(\sqrt{3}n)}{\alpha(\Delta_i^\alpha)^2} + 3$$

when combined with the CVaR-regret decomposition results in

$$R_n^\alpha = \sum_{i=1}^{k} \Delta_i^\alpha \mathbb{E}[T_i(n)] \leq \sum_{i=1}^{K} \frac{40 \ln(\sqrt{3}n) U^2}{\alpha \Delta_i^\alpha} + 3 \sum_{i=1}^{K} \Delta_i^\alpha$$

To get a final result not dependent on each arm's optimality gap, we decompose the CVaR-regret further as

$$\begin{aligned}
R_n^\alpha &= \sum_{i=1}^{k} \Delta_i^\alpha \mathbb{E}[T_i(n)] \\
&= \sum_{i:\Delta_i^\alpha < \Delta} \Delta_i^\alpha \mathbb{E}[T_i(n)] + \sum_{i:\Delta_i^\alpha \geq \Delta} \Delta_i^\alpha \mathbb{E}[T_i(n)] \\
&\leq n\Delta + \sum_{i:\Delta_i^\alpha \geq \Delta} \left(3\Delta_i^\alpha + \frac{40U^2 \ln(\sqrt{3}n)}{\alpha \Delta_i^\alpha} \right) \\
&\leq 4\sqrt{\frac{10kn \ln(\sqrt{3}n)}{\alpha}} U + 3 \sum_{i=1}^{k} \Delta_i^\alpha \\
&\leq 4\sqrt{\frac{10kn \ln(\sqrt{3}n)}{\alpha}} U + 3KU
\end{aligned}$$

where the inequality follows because $\sum_{i:\Delta_i^\alpha < \Delta} T_i(n) \leq n$. Choosing $\Delta = \sqrt{\frac{40kU^2 \log(\sqrt{3}n)}{n\alpha}}$ and simplifying produces the desired problem-independent bound. □

A.3.3 Proxy Regret

Cassel et al. [2018] introduced the notion of proxy regret for risk aware multi-arm bandits as:

$$\bar{R}_\pi(n) = \text{CVaR}_\alpha(F_{p^\star}) - \mathbb{E}[\text{CVaR}_\alpha(F_n^\pi)]$$

where $p^\star = \text{argmax}_{p \in \Delta_{K-1}} \text{CVaR}_\alpha(F_p)$ where Δ_{K-1} is the $K-1$ dimensional simplex:

$$\Delta_{K-1} = \left\{ p = (p_1, \ldots, p_K) \in \mathbb{R}^K | \sum_{i=1}^{K} p_i = 1, p_i \geq 0 \right\}$$

and

$$F_p = \sum_{i=1}^{K} p_i F_i$$

$$F_n^\pi = \frac{1}{n}\sum_{t=1}^{n} F_{\pi_t}$$

Where $F^{(i)}$ is the distribution of arm i and π_t is the policy at step t. Here we establish a formal relation between this notion and CVaR-regret, defined in section 2.7.1.

Proposition 4. CVaR *is a convex function of the CDF. Concretely, if* $\sum_{\alpha_i} = 1$ *and* $\alpha_i \geq 0$*:*

$$\mathrm{CVaR}_\alpha\left(\sum_i \alpha_i F_i(x)\right) \leq \sum_i \alpha_i \,\mathrm{CVaR}_\alpha(F_i(x))$$

Proof. Define the mixture distribution $\hat{F}(x) = \sum_i \alpha_i F_i(x)$:

$$\begin{aligned}
\mathrm{CVaR}_\alpha(\sum_i \alpha_i F_i(x)) &= \frac{1}{\alpha}\int_0^{\hat{F}^{-1}(\alpha)} (\alpha - \sum_i \alpha_i F_i(x))dx \\
&= \frac{1}{\alpha}\int_0^{\hat{F}^{-1}(\alpha)} \sum_i \alpha_i(\alpha - F_i(x))dx \\
&= \sum_i \frac{\alpha_i}{\alpha}\int_0^{\hat{F}^{-1}(\alpha)} (\alpha - F_i(x))dx \\
&\leq \sum_i \frac{\alpha_i}{\alpha}\int_0^{{F_i}^{-1}(\alpha)} (\alpha - F_i(x))dx \\
&= \sum_i \alpha_i \,\mathrm{CVaR}_\alpha(F_i(x))
\end{aligned}$$

Where the last inequality followed by the fact that $\int_0^y(\alpha - F(x))dx$ attains its maximum at $F^{-1}(y)$ □

Proposition 5. *Consider a notion of proxy regret* $\bar{R}_\pi(n)$ *defined in [Cassel et al., 2018] as:*

$$\bar{R}_\pi(n) = \mathbb{E}[\mathrm{CVaR}_\alpha(F_{p^\star}) - \mathrm{CVaR}_\alpha(F_n^\pi)]$$

The notion of regret R_n^α *defined in equation 2.11 satisfies the following inequality.*

$$\bar{R}_\pi(n) \geq \frac{1}{n} R_\alpha^n$$

Proof. First note that By the convexity of CVaR shown in proposition 4 we have $\mathrm{CVaR}_\alpha(F_{p^\star}) = \mathrm{CVaR}_\alpha\left(F_{i^\star}\right)$ where $i^\star = \mathrm{argmax}_i \,\mathrm{CVaR}_\alpha(F_i)$.

$$\mathrm{CVaR}_\alpha(F_{p^\star}) = \mathrm{CVaR}_\alpha\left(\sum_{i=1}^{K} p_i^\star F_i\right)$$

$$\leq \sum_{i=1}^{K} p_i^\star \operatorname{CVaR}_\alpha(F_i)$$
$$\leq \max_i \operatorname{CVaR}_\alpha(F_i)$$

The bound is tight by setting $p^*_{i^\star} = 1$ and $p^*_i = 0 : i \neq i^\star$. Then by using the linearity of expectation and the convexity of CVaR we have:

$$\begin{aligned}
\bar{R}_\pi(n) &= \operatorname{CVaR}_\alpha(F_{p^\star}) - \mathbb{E}[\operatorname{CVaR}_\alpha(F_n^\pi)] \\
&= \operatorname{CVaR}_\alpha(F_{i^\star}) - \mathbb{E}[\operatorname{CVaR}_\alpha(F_n^\pi)] \\
&= \operatorname{CVaR}_\alpha(F_{i^\star}) - \mathbb{E}[\operatorname{CVaR}_\alpha\left(\frac{1}{n}\sum_{t=1}^{n} F_{\pi(t)}\right)] \\
&\geq \operatorname{CVaR}_\alpha(F_{i^\star}) - \frac{1}{n}\mathbb{E}[\sum_{t=1}^{n} \operatorname{CVaR}_\alpha(F_{\pi_i})] \\
&= \frac{1}{n}\left(n \operatorname{CVaR}_\alpha(F_{i^\star}) - \mathbb{E}[\sum_{t=1}^{n} \operatorname{CVaR}_\alpha F_{\pi_i})]\right) = \frac{1}{n} R_\alpha^n
\end{aligned}$$

This completes the proof. □

Here we reiterate and prove the proposition 2.

Proposition (2). *Consider a stochastic K-armed bandit problem with rewards bounded in* $[0, U]$. *For any given horizon* n *and risk level* α, *both* CVaR-*UCB and U-UCB incur proxy regret with* $O(\frac{\log n}{n})$ *and* $O(1/\alpha^2)$ *dependency on the horizon and risk level, respectively.*

Proof. First note that proxy regret decomposes as (Equation 15 in Cassel et al. [2018]):

$$\bar{R}_\pi \leq \frac{L}{n} \sum_{i \neq i^\star} \mathbb{E}[T_i(n)] ||F_{i^\star} - F_i||$$

Where $i^\star = \arg\max_i \operatorname{CVaR}_\alpha(F_i)$ is the optimal arm, and:

$$L = b\left(1 + \max_{i,j \in \{1,\dots,K\}} ||F_i - F_j||^{q-1}\right)$$
$$||F|| = \max\{||F||_\infty, \int_{-\infty}^{0} x dF\}$$

For $\operatorname{CVaR}_\alpha$, $q = 2$ and $b = \frac{1}{\alpha}(1 + \frac{3}{\min\{\alpha, 1-\alpha\}})$ (proposition 4 in Appendix E.2. Cassel et al. [2018]). Following results from Theorem A.3.1, follows that for CVaR-UCB:

$$\bar{R}_\pi \leq \frac{L}{n} \sum_{i \neq i^\star} \left(3 + \frac{4\log(\sqrt{2}n)}{\alpha^2 \Delta_i^2} U_i\right) ||F_{i^\star} - F_i||$$

The right hand side has $O(1/\alpha^2)$ and $O(\log(n)/n)$ dependency on the risk level and horizon, respectively.

Similarly for U-UCB we have (Theorem 2 in Cassel et al. [2018])

$$\bar{R}_{U-UCB} \leq \frac{L}{n} \sum_{i \neq i^\star} \left(\frac{\alpha' \log n}{\phi(\Delta_i/2)} + \frac{\alpha' + 6}{\alpha' - 2} \right) ||F_{i^\star} - F_i||$$

For $\alpha' \geq 2$. Where

$$\phi(y) = \min\{a(\frac{y}{2b})^2, a(\frac{y}{2b})^{2/q}\}$$

For CVaR_α we have $b \leq \frac{1}{\alpha}$ and $q = 2$, which yields $\phi(\Delta_i/2) \leq a\alpha^2\Delta_i^2$. Hence the RHS has $O(1/\alpha^2)$ and $O(\log(n)/n)$ dependency on the risk level and horizon. □

Appendix B

Supplementary Materials for Chapter 3

B.1 Proof of basic lemmas

Before we give the proof of our main results, we give a set of essentially standard lemmas that we build on in the rest of the chapter. In the following, we use a notational shorthand for (nested) expectations under observable distributions: for all $1 \leq t \leq T$ and $1 \leq t_1 \leq t_2 \leq T$,

$$\mathbb{E}^{t}_{a_t}[X] := \mathbb{E}[X \mid H_t, A_t = a_t] \quad \text{and} \tag{B.1a}$$

$$\mathbb{E}^{t_1:t_2}_{a_{t_1:t_2}}[X] := \mathbb{E}^{t_1}_{a_{t_1}}[\mathbb{E}^{t_1+1}_{a_{t_1+1}}[\cdots \mathbb{E}^{t_2}_{a_{t_2}}[X] \cdots]]. \tag{B.1b}$$

Similarly, we write for all $1 \leq t_0 \leq t_1 \leq t_2 \leq T$

$$\mathbb{E}^{t_2}_{a_{t_1:t_2}}[X] := \mathbb{E}[X \mid H_t(A_{1:t_1-1}, a_{t_1:t_2-1}), A_{t_2} = a_{t_2}] \quad \text{and} \tag{B.2a}$$

$$\mathbb{E}^{t_1:t_2}_{a_{t_0:t_2}}[X] := \mathbb{E}^{t_1}_{a_{t_0:t_1}}[\mathbb{E}^{t_1+1}_{a_{t_0:t_1+1}}[\cdots \mathbb{E}^{t_2}_{a_{t_0:t_2}}[X] \cdots]]. \tag{B.2b}$$

The cumulative rewards $\mathbb{E}[Y(\bar{A}_{1:T})]$ under the candidate policy has an alternate representation, which we draw on heavily in the rest of the proofs. See Section B.1.1 for a derivation.

Lemma 11. *If sequential ignorability (Assumption 2) holds for the evaluation policy $\bar{\pi}$, we have the identity*

$$\mathbb{E}\left[Y(\bar{A}_{1:T})\right] = \sum_{a_{1:T}} \mathbb{E}\left[Y(a_{1:T}) \prod_{t=1}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(a_{1:t-1}))\right].$$

To ease notation, denote each integrand in the above sum by

$$Y(a_{1:T};\bar{\pi}) := Y(a_{1:T})\prod_{t=1}^{T}\bar{\pi}_t(a_t \mid \bar{H}_t(a_{1:t-1})). \tag{B.3}$$

We will also use the following two identities heavily. Recall that we denote by $W := \{W(a_{1:T})\}_{a_{1:T}}$, the tuple of all potential outcomes, which takes values in $\mathcal{W}$. See Section B.1.2 for a proof of the following result.

Lemma 12. *Let sequential ignorability (Assumption 2) hold for the behavioral policy π in the time steps $t_1 : t_2$, where $1 \le t_1 < t_2 \le T$. Then, for any measurable $f : \mathcal{W} \to \mathbb{R}$*

$$\mathbb{E}[f(W) \mid H_{t_1}(a_{1:t_1-1})] = \mathbb{E}\left[\mathbb{E}^{t_1:t_2}_{a_{1:t_2}}[f(W)] \mid H_{t_1}(a_{1:t_1-1})\right]$$

for any $a_{1:t_2} \in \mathcal{A}_1 \times \cdots \times \mathcal{A}_{t_2}$.

The following identity—whose proof we give in Section B.1.3—is a simple consequence of the definition of conditional expectations, and the tower law.

Lemma 13. *For any measurable function $f : \mathcal{W} \to \mathbb{R}$, and $1 \le t_1 \le t_2 \le T$,*

$$\mathbb{E}^{t_1:t_2}_{a_{1:t_2}} f(W) = \mathbb{E}\left[f(W)\prod_{t=t_1}^{t_2}\frac{\mathbf{1}\{A_T = a_t\}}{\pi_t(a_t \mid H_t(a_{1:t-1}))} \mid H_{t_1}(a_{1:t_1-1})\right]$$

B.1.1 Proof of Lemma 11

Similar to the notational shorthand (Equation B.1), define

$$\overline{\mathbb{E}}^{t}_{a_{1:t}}[X] := \mathbb{E}[X \mid \bar{H}_t(a_{1:t-1}), \bar{A}_t = a_t] \quad \text{and} \quad \overline{\mathbb{E}}^{t:T}_{a_{1:T}}[X] = \overline{\mathbb{E}}^{t}_{a_{1:t}}[\overline{\mathbb{E}}^{t+1}_{a_{1:t+1}}[\cdots \overline{\mathbb{E}}^{T}_{a_{1:T}}[X]\cdots]].$$

Begin by noting that by definition of conditional expectation

$$\begin{aligned}\mathbb{E}[Y(\bar{A}_{1:T}) \mid \bar{H}_1] &= \sum_{a_1 \in \mathcal{A}_1}\bar{\pi}(a_1 \mid \bar{H}_1)\mathbb{E}[Y(a_1, \bar{A}_{2:T}) \mid \bar{H}_1, \bar{A}_1 = a_1] \\ &= \sum_{a_1 \in \mathcal{A}_1}\bar{\pi}(a_1 \mid \bar{H}_1)\overline{\mathbb{E}}^{1}_{a_1}[Y(a_1, \bar{A}_{2:T})],\end{aligned}$$

and similarly, conditioning on $\bar{H}_2(a_1) = (S_1, \bar{A}_1 = a_1, S_2(a_1))$ yields

$$\begin{aligned}\mathbb{E}[Y(a_1, \bar{A}_{2:T}) \mid \bar{H}_2(a_1)] &= \sum_{a_2 \in \mathcal{A}_2}\bar{\pi}_2(a_2 \mid \bar{H}_2(a_1))\mathbb{E}[Y(a_{1:2}, \bar{A}_{3:T}) \mid \bar{H}_2(a_1), \bar{A}_2 = a_2] \\ &= \sum_{a_2 \in \mathcal{A}_2}\bar{\pi}_2(a_2 \mid \bar{H}_2(a_1))\overline{\mathbb{E}}^{2}_{a_{1:2}}[Y(a_{1:2}, \bar{A}_{3:T})].\end{aligned}$$

From the tower law, the above two equalities yield

$$\begin{aligned}\mathbb{E}[Y(\bar{A}_{1:T})] &= \mathbb{E}\Bigg[\sum_{a_1\in\mathcal{A}_1}\bar{\pi}_1(a_1\mid\bar{H}_1)\times \\ &\qquad \mathbb{E}\Bigg[\sum_{a_2\in\mathcal{A}}\bar{\pi}_2(a_2\mid\bar{H}_2(a_1))\cdot\mathbb{E}[Y(a_{1:2},\bar{A}_{3:T})\mid\bar{H}_2(a_1),\bar{A}_2=a_2]\Bigg|\bar{H}_1,\bar{A}_1=a_1\Bigg]\Bigg] \\ &= \mathbb{E}\Bigg[\sum_{a_1\in\mathcal{A}_1}\bar{\pi}_1(a_1\mid\bar{H}_1)\overline{\mathbb{E}}^1_{a_1}\Bigg[\sum_{a_2\in\mathcal{A}}\bar{\pi}_2(a_2\mid\bar{H}_2(a_1))\cdot\overline{\mathbb{E}}^2_{a_{1:2}}[Y(a_{1:2},\bar{A}_{3:T})]\Bigg]\Bigg].\end{aligned}$$

Proceeding iteratively as before and expanding each $\mathbb{E}[Y(a_{1:t-1},\bar{A}_{t:T})\mid\bar{H}_t(a_{1:t-1})]$, we arrive at

$$\begin{aligned}\mathbb{E}[Y(\bar{A}_{1:T})] = &\\ &\mathbb{E}\Bigg[\sum_{a_1\in\mathcal{A}_1}\bar{\pi}_1(a_1\mid\bar{H}_1)\overline{\mathbb{E}}^1_{a_1}\Bigg[\overline{\mathbb{E}}^2_{a_{1:2}}\Bigg[\sum_{a_2\in\mathcal{A}_2}\bar{\pi}_2(a_2\mid\bar{H}_2(a_1))\overline{\mathbb{E}}^3_{a_{1:3}}\times \\ &\Bigg[\cdots\sum_{a_T\in\mathcal{A}_T}\bar{\pi}_T(a_T\mid\bar{H}_T(a_{1:T-1}))\overline{\mathbb{E}}^T_{a_{1:T}}[Y(a_{1:T})]\Bigg]\Bigg]\Bigg]\Bigg].\end{aligned}$$

Now, we proceed backwards from the inner most expectation to take the outer sum inside the expectation. By Assumption 2, we have

$$\begin{aligned}&\sum_{a_T\in\mathcal{A}_T}\bar{\pi}_T(a_T\mid\bar{H}_T(a_{1:T-1}))\overline{\mathbb{E}}^T_{a_{1:T}}[Y(a_{1:T})] \\ &\qquad=\sum_{a_T\in\mathcal{A}_T}\bar{\pi}_T(a_T\mid\bar{H}_T(a_{1:T-1}))\cdot\mathbb{E}\left[Y(a_{1:T})\;\middle|\;\bar{H}_T(a_{1:T-1})\right] \\ &\qquad=\mathbb{E}\left[\sum_{a_T\in\mathcal{A}_T}\bar{\pi}_T(a_T\mid\bar{H}_T(a_{1:T-1}))\cdot Y(a_{1:T})\;\middle|\;\bar{H}_T(a_{1:T-1})\right].\end{aligned}$$

Noting that $\mathbb{E}[\cdot\mid\bar{H}_T(a_{1:T-1})]=\mathbb{E}[\cdot\mid\bar{H}_{T-1}(a_{1:T-2}),S_T(a_{1:T-1}),\bar{A}_{T-1}{=}a_{T-1}]$, the tower law and preceding display yield

$$\begin{aligned}&\overline{\mathbb{E}}^{T-1}_{a_{1:T-1}}\left[\sum_{a_T\in\mathcal{A}_T}\bar{\pi}_T(a_T\mid\bar{H}_T(a_{1:T-1}))\cdot\overline{\mathbb{E}}^T_{a_{1:T}}[Y(a_{1:T})]\right] \\ &\qquad=\overline{\mathbb{E}}^{T-1}_{a_{1:T-1}}\left[\sum_{a_T\in\mathcal{A}_T}\bar{\pi}_T(a_T\mid\bar{H}_T(a_{1:T-1}))\cdot Y(a_{1:T})\right].\end{aligned}$$

We repeat an identical process for the sum over a_{T-1}. Similarly as above, applying Assumption 2

gives

$$\sum_{a_{T-1}\in\mathcal{A}_{T-1}} \bar{\pi}_{T-1}(a_{T-1} \mid \bar{H}_{T-1}(a_{1:T-2})) \cdot \overline{\mathbb{E}}_{a_{1:T-1}}^{T-1}\left[\sum_{a_T\in\mathcal{A}_T} \bar{\pi}_T(a_T \mid \bar{H}_T(a_{1:T-1})) \cdot Y(a_{1:T})\right]$$
$$= \mathbb{E}\Big[\sum_{a_{T-1}\in\mathcal{A}_{T-1}} \bar{\pi}_{T-1}(a_{T-1} \mid \bar{H}_{T-1}(a_{1:T-2}))$$
$$\sum_{a_T\in\mathcal{A}_T} \bar{\pi}_T(a_T \mid \bar{H}_T(a_{1:T-1})) \cdot Y(a_{1:T}) \;\Big|\; \bar{H}_{T-1}(a_{1:T-2})\Big].$$

By the tower law, we again get

$$\overline{\mathbb{E}}_{a_{1:T-2}}^{T-2}\left[\sum_{a_{T-1}\in\mathcal{A}_{T-1}} \bar{\pi}_{T-1}(a_{T-1} \mid \bar{H}_{T-1}(a_{1:T-2}))\overline{\mathbb{E}}_{a_{1:T-1}}^{T-1}\left[\sum_{a_T\in\mathcal{A}_T} \bar{\pi}_T(a_T \mid \bar{H}_T(a_{1:T-1})) \cdot Y(a_{1:T})\right]\right]$$
$$= \overline{\mathbb{E}}_{a_{1:T-2}}^{T-2}\left[\sum_{a_{T-1}\in\mathcal{A}_{T-1}} \bar{\pi}_{T-1}(a_{T-1} \mid \bar{H}_{T-1}(a_{1:T-2})) \cdot \sum_{a_T\in\mathcal{A}_T} \bar{\pi}_T(a_T \mid \bar{H}_T(a_{1:T-1})) \cdot Y(a_{1:T})\right].$$

Iterating the above process over the indices $t = T-2, \ldots, 1$, we arrive at the desired formula.

B.1.2 Proof of Lemma 12

From the tower law and sequential ignorability of π,

$$\begin{aligned}\mathbb{E}[f(W) \mid H_{t_1}(a_{1:t_1-1})] &= \mathbb{E}[f(W) \mid H_{t_1}(a_{1:t_1-1}), A_{t_1} = a_{t_1}] \\ &= \mathbb{E}[\mathbb{E}[f(W) \mid H_{t_1+1}(a_{1:t_1})] \mid H_{t_1}(a_{1:t_1-1}), A_{t_1} = a_{t_1}]\end{aligned}$$

Applying the tower law to the inner expectation, and applying sequential ignorability again, we get

$$\mathbb{E}[f(W) \mid H_{t_1+1}(a_{1:t_1})] = \mathbb{E}\left[\mathbb{E}[f(W) \mid H_{t_1+2}(a_{1:t_1+1})] \mid H_{t_1+1}(a_{1:t_1}), A_{t_1+1} = a_{t_1+1}\right]$$

Plugging this back into the original display, we have

$$\mathbb{E}[f(W) \mid H_{t_1}(a_{1:t_1-1})] = \mathbb{E}_{a_1:t_1+1}^{t_1:t_1+1}\left[\mathbb{E}[f(W) \mid H_{t_1+2}(a_{1:t_1+1})]\right]$$

Repeating this argument over $t = t_1+2, \ldots, t_2$, we conclude the result.

B.1.3 Proof of Lemma 13

From the definition of conditional expectations, we have

$$\mathbb{E}[f(W) \mid H_t(a_{1:t-1}), A_t = a_t] = \mathbb{E}\left[f(W)\frac{\mathbf{1}\{A_t = a_t\}}{\pi_t(a_t \mid H_t(a_{1:t-1}))} \mid H_t(a_{1:t-1})\right].$$

The result follows by applying this equality at $t = t_2$, applying the tower law, and iterating the same argument over $t = t_2 - 1, \ldots, t_1$.

B.2 Proof of key identities

B.2.1 Proof of Lemma 1

Recalling the notation (Equation B.3), sequential ignorability of $\bar{\pi}$ and Lemma 11 gives the following representation

$$\mathbb{E}\left[Y(\bar{A}_{1:T})\right] = \sum_{a_{1:T}} \mathbb{E}\left[Y(a_{1:T}; \bar{\pi})\right].$$

We deal with each term $\mathbb{E}[Y(a_{1:T}; \bar{\pi})]$ in the summation separately, for each fixed sequence of actions $a_{1:T}$. From sequential ignorability of π and Lemma 12,

$$\mathbb{E}[Y(a_{1:T}; \bar{\pi})] = \mathbb{E}[\mathbb{E}^1_{a_1}[\cdots \mathbb{E}^T_{a_{1:T}}[Y(a_{1:T}; \bar{\pi})] \cdots]] = \mathbb{E}[\mathbb{E}^{1:T}_{a_{1:T}}[Y(a_{1:T}; \bar{\pi})]].$$

Applying Lemma 13, we get

$$\mathbb{E}[Y(a_{1:T}; \bar{\pi})] = \mathbb{E}\left[Y(a_{1:T}; \bar{\pi}) \prod_{t=1}^{T} \frac{\mathbf{1}\{A_t = a_t\}}{\pi_t(a_t \mid H_t(a_{1:t-1}))}\right].$$

Summing the preceeding display over $a_{1:T}$, we obtain the desired result.

B.2.2 Proof of Proposition 3

From Lemma 11, we have

$$\mathbb{E}[Y(\bar{A}_{1:T})] = \mathbb{E}\left[\sum_{a_{1:T}} Y(a_{1:T}) \prod_{t=1}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(a_{1:t-1}))\right].$$

Since sequential ignorability for π holds at any $t < t^\star$, Lemma 12 implies that the preceeding display is equal to

$$\mathbb{E}\left[\sum_{a_{1:t^\star-1}} \mathbb{E}^{1:t^\star-1}_{a_{1:t^\star-1}}\left[\sum_{a_{t^\star:T}} Y(a_{1:T}) \prod_{t=1}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(a_{1:t-1}))\right]\right].$$

Applying Lemma 13 to the inner expectations, we get

$$\begin{aligned}
\mathbb{E}[Y(\bar{A}_{1:T})] &= \mathbb{E}\left[\sum_{a_{1:t^\star-1}} \prod_{t=1}^{t^\star-1} \frac{\mathbf{1}\{A_t = a_t\}}{\pi_t(a_t \mid H_t(a_{1:t-1}))} \sum_{a_{t^\star:T}} Y(a_{1:T}) \prod_{t=1}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(a_{1:t-1}))\right] \\
&= \mathbb{E}\left[\prod_{t=1}^{t^\star-1} \frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t-1}))}{\pi_t(A_t \mid H_t)} \sum_{a_{t^\star:T}} Y(A_{1:t^\star-1}, a_{t^\star:T}) \prod_{t=t^\star}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1}))\right].
\end{aligned}$$

From the tower law, we arrive at

$$\begin{aligned}
\mathbb{E}[Y(\bar{A}_{1:T})] &= \mathbb{E}\Bigg[\mathbb{E}\Bigg[\prod_{t=1}^{t^\star-1} \frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t-1}))}{\pi_t(A_t \mid H_t)} \sum_{a_{t^\star:T}} Y(A_{1:t^\star-1}, a_{t^\star:T}) \\
&\qquad \prod_{t=t^\star}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1})) \,\Bigg|\, H_{t^\star}\Bigg]\Bigg] \\
&= \mathbb{E}\Bigg[\prod_{t=1}^{t^\star-1} \frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t-1}))}{\pi_t(A_t \mid H_t)} \mathbb{E}\Bigg[\sum_{a_{t^\star:T}} Y(A_{1:t^\star-1}, a_{t^\star:T}) \\
&\qquad \prod_{t=t^\star}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1})) \,\Bigg|\, H_{t^\star}\Bigg]\Bigg]. \qquad \text{(B.5)}
\end{aligned}$$

Applying the tower law to the inner expectation in the final display, we can write

$$\begin{aligned}
&\mathbb{E}\left[\sum_{a_{t^\star:T}} Y(A_{1:t^\star-1}, a_{t^\star:T}) \prod_{t=t^\star}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1})) \,\middle|\, H_{t^\star}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left[\sum_{a_{t^\star:T}} Y(A_{1:t^\star-1}, a_{t^\star:T}) \prod_{t=t^\star}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1})) \,\middle|\, H_{t^\star}, A_{t^\star}\right] \,\middle|\, H_{t^\star}\right] \\
&= \sum_{a_{t^\star}, a'_{t^\star}} \bar{\pi}_{t^\star}(a_{t^\star} \mid \bar{H}_{t^\star}(A_{1:t^\star-1})) \pi_{t^\star}(a'_{t^\star} \mid H_{t^\star}) \\
&\qquad \times \mathbb{E}\left[\sum_{a_{t^\star+1:T}} Y(A_{1:t^\star-1}, a_{t^\star:T}) \prod_{t=t^\star+1}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1})) \,\middle|\, H_{t^\star}, A_{t^\star} = a'_{t^\star}\right] \\
&= \sum_{a_{t^\star}, a'_{t^\star}} \bar{\pi}_{t^\star}(a_{t^\star} \mid \bar{H}_{t^\star}(A_{1:t^\star-1})) \pi_{t^\star}(a'_{t^\star} \mid H_{t^\star})
\end{aligned}$$

$$\times \mathbb{E}\Bigg[\mathcal{L}(W; H_{t^\star}, a_{t^\star}, a'_{t^\star}) \sum_{a_{t^\star+1:T}} Y(A_{1:t^\star-1}, a_{t^\star:T}) \times \prod_{t=t^\star+1}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1})) \;\Bigg|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\Bigg]$$

where in the last equality, we used the definition

$$\mathcal{L}(\cdot; H_{t^\star}, a_{t^\star}, a'_{t^\star}) := \frac{dP_W(\cdot \mid H_{t^\star}, A_{t^\star} = a'_{t^\star})}{dP_W(\cdot \mid H_{t^\star}, A_{t^\star} = a_{t^\star})}.$$

Again, by the tower law,

$$\mathbb{E}\Bigg[\mathcal{L}(W; H_{t^\star}, a_{t^\star}, a'_{t^\star}) \sum_{a_{t^\star+1:T}} Y(A_{1:t^\star-1}, a_{t^\star:T}) \prod_{t=t^\star+1}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1})) \;\Bigg|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\Bigg]$$

$$= \mathbb{E}\Bigg[\mathbb{E}\Big[\mathcal{L}(W; H_{t^\star}, a_{t^\star}, a'_{t^\star}) \sum_{a_{t^\star+1:T}} Y(A_{1:t^\star-1}, a_{t^\star:T}) \times \prod_{t=t^\star+1}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1})) \Big| H_{t^\star+1}(A_{1:t^\star-1}, a_{t^\star})\Big] \;\Bigg|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\Bigg]$$

From sequential ignorability of π for $t > t^\star$ and Lemma 12, the preceeding display is equal to

$$\mathbb{E}\Bigg[\sum_{a_{t^\star+1:T}} \mathbb{E}^{t^\star+1:T}_{a_{t^\star:T}} \mathcal{L}(W; H_{t^\star}, a_{t^\star}, a'_{t^\star}) Y(A_{1:t^\star-1}, a_{t^\star:T}) \prod_{t=t^\star+1}^{T} \bar{\pi}_t(a_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star:t-1})) \;\Bigg|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\Bigg].$$

From Lemma 13, we can rewrite the above expression as

$$\mathbb{E}\left[\mathcal{L}(W; H_{t^\star}, a_{t^\star}, a'_{t^\star}) Y_{t^\star}(a_{t^\star}) \prod_{t=t^\star+1}^{T} \frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t^\star-1}, a_{t^\star}, A_{t^\star+1:t}))}{\pi_t(A_t \mid H_t(A_{1:t^\star-1}, a_{t^\star}, A_{t^\star+1:t}))} \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\right].$$

Plugging these expressions back into the equality (Equation B.5), we obtain the result.

B.3 Proof of bounds under unobserved confounding

B.3.1 Naive bound

We show the below more general result.

Lemma 14. *Let Assumptions 2, 3, 4 hold. Then, we have*

$$\mathbb{E}[Y(\bar{A}_{1:T})] \geq \mathbb{E}\left[Y(A_{1:T})\prod_{t=1}^{T}\frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t-1}))}{\pi_t(A_t \mid H_t)(\Gamma_t^{-1}\mathbf{1}\{Y(A_{1:T})<0\}+\Gamma_t\mathbf{1}\{Y(A_{1:T})>0\})}\right].$$

Proof of Lemma From an identical argument as the proof of Lemma 1, Assumption 3 yields

$$\mathbb{E}[Y(\bar{A}_{1:T})] = \mathbb{E}\left[Y(A_{1:T})\prod_{t=1}^{T}\frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t-1}), U_t)}{\pi_t(A_t \mid H_t, U_t)}\right].$$

From Assumption 2, the preceeding display is equal to

$$\mathbb{E}[Y(\bar{A}_{1:T})] = \mathbb{E}\left[Y(A_{1:T})\prod_{t=1}^{T}\frac{\bar{\pi}_t(A_t \mid \bar{H}_t(A_{1:t-1}))}{\pi_t(A_t \mid H_t, U_t)}\right]. \tag{B.6}$$

Now, we bound $\pi_t(A_t \mid H_t, U_t)$ by $\pi_t(A_t \mid H_t)$. Assumption 4 implies

$$\pi_t(a_t \mid H_t, U_t = u_t)\pi_t(a_t' \mid H_t, U_t = u_t') \leq \Gamma_t\pi_t(a_t' \mid H_t, U_t = u_t)\pi_t(a_t \mid H_t, U_t = u_t').$$

Multiplying by $p_{U_t}(u_t' \mid H_t)$ on both sides and integrating over u_t', we get

$$\pi_t(a_t \mid H_t, U_t = u_t)\pi_t(a_t' \mid H_t) \leq \Gamma_t\pi_t(a_t' \mid H_t, U_t = u_t)\pi_t(a_t \mid H_t).$$

Summing over a_t' on both sides, we conclude that

$$\pi_t(a_t \mid H_t, U_t = u_t) \leq \Gamma_t\pi_t(a_t \mid H_t).$$

almost surely, for any t, a_t, H_t, u_t. Using this relation to lower bound expression (Equation B.6), we obtain the result.

B.3.2 Proof of Theorem 3

By rewriting the original infimization problem over $L(W; H_{t^\star})$ to $L(W, A_{t^\star+1:T}; H_{t^\star})$, we have

$$\eta^\star(H_{t^\star}; a_{t^\star}) = \inf_{L\geq 0}\left\{\mathbb{E}\left[L(W, A_{t^\star+1:T}; H_{t^\star})Y(A_{1:T})\prod_{t=t^\star+1}^{T}\rho_t \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star}\right]:\right.$$
$$\left.\mathbb{E}[L(W, A_{t^\star+1:T}; H_{t^\star}) \mid H_{t^\star}, A_{t^\star} = a_{t^\star}] = 1,\right.$$

$$L(w, a_{t^\star+1:T}; H_{t^\star}) = L(w, a'_{t^\star+1:T}; H_{t^\star}),$$
$$L(w, a_{t^\star+1:T}; H_{t^\star}) \le \Gamma L(w', a'_{t^\star+1:T}; H_{t^\star}) \text{ a.s. all } w, a_{t^\star+1:T}, w', a'_{t^\star+1:T} \Bigg\}.$$

Relaxing the equality constraint $L(w, a_{t^\star+1:T}; H_{t^\star}) = L(w, a'_{t^\star+1:T}; H_{t^\star})$, we arrive at

$$\eta^\star(H_{t^\star}; a_{t^\star}) \ge \inf_{L \ge 0} \Bigg\{ \mathbb{E}\left[L(W, A_{t^\star+1:T}; H_{t^\star}) Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star} \right] :$$
$$\mathbb{E}[L(W, A_{t^\star+1:T}; H_{t^\star}) \mid H_{t^\star}, A_{t^\star} = a_{t^\star}] = 1,$$
$$L(w, a_{t^\star+1:T}; H_{t^\star}) \le \Gamma L(w', a'_{t^\star+1:T}; H_{t^\star}) \text{ a.s. all } w, a_{t^\star+1:T}, w', a'_{t^\star+1:T} \Bigg\}.$$

The preceeding optimization problem is convex, and Slater's condition holds for $L \equiv 1$. By strong duality [Luenberger, 1969, Thm. 8.6.1 and Problem 8.7], we obtain the dual formulation

$$\sup_{\mu} \inf_{L \ge 0} \Bigg\{ \mathbb{E}\left[L(W, A_{t^\star+1:T}; H_{t^\star}) \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - \mu \right) \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star} \right] + \mu :$$
$$L(w, a_{t^\star+1:T}; H_{t^\star}) \le \Gamma L(w', a'_{t^\star+1:T}; H_{t^\star}) \text{ a.s. all } w, a_{t^\star+1:T}, w', a'_{t^\star+1:T} \Bigg\}.$$

By inspection, the solution to the inner infimum takes the form

$$L(w, a_{t^\star+1:T}; H_{t^\star}) = c \left(\Gamma \mathbf{1}\left\{ Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - \mu < 0 \right\} + \mathbf{1}\left\{ Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - \mu \ge 0 \right\} \right)$$

for some constant $c > 0$. Let $\ell'_\Gamma(z) := (z)_+ - \Gamma(z)_-$, the derivative of the weighted squared loss $\ell_\Gamma(z) = \frac{1}{2}(\Gamma(z)_-^2 + (z)_+^2)$. Plugging the preceeding display into the dual formulation, we get

$$\sup_{\mu} \inf_{c \ge 0} \left\{ c\mathbb{E}\left[\ell'_\Gamma \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - \mu \right) \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star} \right] + \mu \right\}$$
$$= \sup_{\mu} \left\{ \mu : \mathbb{E}\left[\ell'_\Gamma \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - \mu \right) \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star} \right] \ge 0 \right\}.$$

Since the function $\mu \mapsto \mathbb{E}\left[\ell'_\Gamma \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - \mu \right) \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star} \right]$ is strictly decreasing, the optimal solution (and its value) in the preceeding display is given by the unique zero of this function.

We now show that the solution to our loss minimization problem

$$\kappa(H_{t^\star}; a_{t^\star}) = \operatorname*{argmin}_{f(H_{t^\star})} \left\{ \mathbb{E}\left[\frac{\mathbf{1}\left\{A_{t^\star} = a_{t^\star}\right\}}{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star})} \times \ell_\Gamma\left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - f(H_{t^\star}) \right) \right] \right.$$
$$\left. = \mathbb{E}\left[\mathbb{E}\left[\ell_\Gamma\left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - f(H_{t^\star}) \right) \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star} \right] \right] \right\}$$

is in fact the unique zero of the function $\mu \mapsto \mathbb{E}\left[\ell_\Gamma'\left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - \mu \right) \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star} \right]$. The (almost sure) uniqueness of the solution follows from strong convexity of ℓ_Γ. Since the optimization is over all $H_{t^\star}$-measurable functions, the argmin is given by

$$\operatorname*{argmin}_{f(H_{t^\star})} \; \mathbb{E}\left[\ell_\Gamma\left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - f(H_{t^\star}) \right) \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star} \right].$$

So long as

$$\mathbb{E}[Y(A_{1:T})^2 \prod_{t=t^\star+1}^{T} \rho_t^2 \mid A_{t^\star} = a_{t^\star}, H_{t^\star}] < \infty \qquad \text{almost surely}$$

first order optimality conditions of this loss minimization problem is equivalent to

$$\mathbb{E}\left[\ell_\Gamma'\left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - f(H_{t^\star}) \right) \;\middle|\; H_{t^\star}, A_{t^\star} = a_{t^\star} \right] = 0$$

which gives our result.

B.3.3 Proof of Theorem 4

Our result is based on epi-convergence theory [King and Wets, 1991, Rockafellar and Wets, 1998], which shows (uniform) convergence of convex functions, and solutions to convex optimization problems.

Definition 2. *Let $\{A_n\}$ be a sequence of subsets of $\mathbb{R}^d$. The* limit supremum *(or* limit exterior *or* outer limit*) and* limit infimum *(*limit interior *or* inner limit*) of the sequence $\{A_n\}$ are*

$$\limsup_n A_n := \left\{ v \in \mathbb{R}^d \mid \liminf_{n\to\infty} \operatorname{dist}(v, A_n) = 0 \right\} \quad \textit{and}$$
$$\liminf_n A_n := \left\{ v \in \mathbb{R}^d \mid \limsup_{n\to\infty} \operatorname{dist}(v, A_n) = 0 \right\}.$$

The epigraph of a function $h : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ is $\operatorname{epi} h := \{(x,t) \in \mathbb{R}^d \times \mathbb{R} \mid h(x) \le t\}$. We say

$\lim_n A = A_\infty$ if $\limsup_n A_n = \liminf_n A_n = A_\infty \subset \mathbb{R}^d$. We define a notion of convergence for functions in terms of their epigraphs.

Definition 3. *A sequence of functions* h_n epi-converges to a function h, *denoted* $h_n \overset{\text{epi}}{\to} h$, *if*

$$\operatorname{epi} h = \liminf_{n\to\infty} \operatorname{epi} h_n = \limsup_{n\to\infty} \operatorname{epi} h_n. \tag{B.7}$$

If h is proper ($\operatorname{dom} h \neq \varnothing$), epigraphical convergence (Equation B.7) is characterizaed by pointwise convergence on a dense set.

Lemma 15 (Theorem 7.17, Rockafellar and Wets [1998])**.** *Let* $h_n : \mathbb{R}^d \to \overline{\mathbb{R}}, h : \mathbb{R}^d \to \overline{\mathbb{R}}$ *be closed, convex, and proper. Then* $h_n \overset{\text{epi}}{\to} h$ *is equivalent to either of the following two conditions.*

(i) There exists a dense set $A \subset \mathbb{R}^d$ *such that* $h_n(v) \to h(v)$ *for all* $v \in A$.

(ii) For all compact $C \subset \operatorname{dom} h$ *not containing a boundary point of* $\operatorname{dom} h$,

$$\lim_{n\to\infty} \sup_{v\in C} |h_n(v) - h(v)| = 0.$$

The last characterization of epigraphical convergence is powerful as it gives convergence of solution sets.

Lemma 16 (Theorem 7.31, Rockafellar and Wets [1998])**.** *Let* $h_n : \mathbb{R}^d \to \overline{\mathbb{R}}, h : \mathbb{R}^d \to \overline{\mathbb{R}}$ *satisfy* $h_n \overset{\text{epi}}{\to} h$ *and* $-\infty < \inf h < \infty$. *Let* $S_n(\varepsilon) = \{\theta \mid h_n(\theta) \leq \inf h_n + \varepsilon\}$ *and* $S(\varepsilon) = \{\theta \mid h(\theta) \leq \inf h + \varepsilon\}$. *Then* $\limsup_n S_n(\varepsilon) \subset S(\varepsilon)$ *for all* $\varepsilon \geq 0$, *and* $\limsup_n S_n(\varepsilon_n) \subset S(0)$ *whenever* $\varepsilon_n \downarrow 0$.

From Lemmas 15, 16, it suffices to show that the expected loss function and its empirical counterpart satisfies appropriate regularity conditions (proper and closed), and show that our empirical loss pointwise converges to the population loss almost surely. Recall that $\mathcal{D}_n$ is the split of data used to estimate $\widehat{\pi}$, and let $\mathcal{D}_\infty$ be the σ-algebra defined by $\mathcal{D}_n$ as $n \to \infty$. Our subsequent argument will be conditional on $\mathcal{D}_\infty$, and the event

$$\mathcal{E} := \left\{\widehat{\pi}_t \to \pi_t \ \text{ pointwise, } \widehat{\rho}_t \leq 2C, \text{ and } \ \widehat{\pi}_{t^\star}(a_{t^\star} \mid H_{t^\star}) \in [(2C)^{-1}, 1]\right\}.$$

We assume w.l.o.g. (increasing C if necessary) that $c \leq (2C)^{-1}$. Note that $\mathbb{P}(\mathcal{E}) = 1$ by assumption.

First, note that since $\theta \mapsto f_\theta$ is linear, $\theta \mapsto \ell_\Gamma(Y(A_{1:T}) \prod_{t=t^\star+1}^T \widehat{\rho}_t - f_\theta(H_{t^\star}))$ is convex. Both the empirical and population loss

$$\theta \mapsto \widehat{\mathbb{E}}_n \left[\frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\widehat{\pi}_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma\left(Y(A_{1:T}) \prod_{t=t^\star+1}^T \widehat{\rho}_t - f_\theta(H_{t^\star})\right)\right] =: \widehat{h}_n(\theta),$$

$$\theta \mapsto \mathbb{E} \left[\frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma\left(Y(A_{1:T}) \prod_{t=t^\star+1}^T \rho_t - f_\theta(H_{t^\star})\right)\right] =: h(\theta),$$

are proper since they are nonnegative, and finite a.s. at $\theta = 0$. Since the functions

$$\theta \mapsto \frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\widehat{\pi}_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \widehat{\rho}_t - f_\theta(H_{t^\star}) \right),$$
$$\theta \mapsto \frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \rho_t - f_\theta(H_{t^\star}) \right),$$

are continuous by linearity of $\theta \mapsto f_\theta$, dominated convergence shows continuity of both the empirical loss $\widehat{h}_n(\theta)$ (a.s.) and population loss $h(\theta)$.

Next, we show that the empirical plug-in loss converges pointwise to its population counterpart almost surely. Since $S(0) = \{\theta^\star\}$ by hypothesis, Lemmas 15, 16 will give the final result. Defining the function

$$h_n(\theta) := \mathbb{E}\left[\frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\widehat{\pi}_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \widehat{\rho}_t - f_\theta(H_{t^\star}) \right) \right],$$

we write

$$|\widehat{h}_n(\theta) - h(\theta)| \le |h_n(\theta) - h(\theta)| + |\widehat{h}_n(\theta) - h_n(\theta)|,$$

and show that each term in the right hand side converges to 0 almost surely.

To show that the first term goes to zero, note that since $\widehat{\pi}_{t^\star} \to \pi_{t^\star}$ a.s., we have $\pi_{t^\star}(a_{t^\star} \mid H_t) \ge (2C)^{-1}$ a.s. for all $a_{t^\star}$. This gives

$$\begin{aligned}
|h_n(\theta) - h(\theta)| &\le \left| h_n(\theta) - \mathbb{E}\left[\frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \widehat{\rho}_t - f_\theta(H_{t^\star}) \right) \right] \right| \\
&\quad + \left| \mathbb{E}\left[\frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\pi_{t^\star}(a_{t^\star} \mid H_{t^\star})} \ell_\Gamma \left(Y(A_{1:T}) \prod_{t=t^\star+1}^{T} \widehat{\rho}_t - f_\theta(H_{t^\star}) \right) \right] - h(\theta) \right| \\
&\le \Gamma C^2 \mathbb{E}\Bigg[|\pi_{t^\star}(a_{t^\star} \mid H_{t^\star}) - \widehat{\pi}_{t^\star}(a_{t^\star} \mid H_{t^\star})| \cdot \left(Y(A_{1:T})^2 (2C)^{2T} + 2|f_\theta(H_{t^\star})|^2 \right) \\
&\quad + \Gamma C \mathbb{E}\Bigg[\left(Y(A_{1:T}) 2(2C)^T + 2|f_\theta(H_{t^\star})| \right) \cdot Y(A_{1:T}) \cdot \left| \prod_{t=t^\star+1}^{T} \widehat{\rho}_t - \prod_{t=t^\star+1}^{T} \rho_t \right| \Bigg],
\end{aligned}$$

which has an integrable envelope function under our assumptions and conditional on $\mathcal{E}$. By dominated convergence, we have the result since $\widehat{\pi}_t \to \pi_t$ almost surely (and hence $\widehat{\rho}_t \overset{a.s.}{\to} \rho_t$).

To show that the second term converges to zero, we use the following strong law of large numbers for triangular arrays.

Lemma 17 (Hu et al. [1989, Theorem 2])**.** *Let $\{\xi_{ni}\}_{i=1}^n$ be a triangular array where $X_{n1}, X_{n2}, \ldots$ are independent random variables for any fixed n. If there exists ξ such that $|\xi_{ni}| \le \xi$ and $\mathbb{E}[\xi^2] < \infty$,*

then $\frac{1}{n}\sum_{i=1}^{n}(\xi_{ni} - \mathbb{E}[\xi_{ni}]) \overset{a.s.}{\to} 0$.

The random variable

$$\frac{\mathbf{1}\{A_{t^\star} = a_{t^\star}\}}{\widehat{\pi}_{t^\star}(a_{t^\star} \mid H_{t^\star})}\ell_\Gamma\left(Y(A_{1:T})\prod_{t=t^\star+1}^{T}\widehat{\rho}_t - f_\theta(H_{t^\star})\right)$$

are i.i.d. for each trajectory, conditional on $\mathcal{D}_\infty$. By convexity, the below random variable upper bounds the preceeding display

$$\xi = 16\Gamma(2C)^{2T}\left(f_\theta(H_{t^\star})^2 + Y(A_{1:T})^2\right)$$

on the event $\mathcal{E}$. From hypothesis, we have $\mathbb{E}[\xi^2 \mid \mathcal{D}_\infty, \mathcal{E}] < \infty$. Applying Lemma 17 conditional on $\mathcal{D}_\infty$ and $\mathcal{E}$, we conclude

$$|\widehat{h}_n(\theta) - h_n(\theta)| \overset{a.s.}{\to} 0.$$

Applying Lemmas 15, 16, we conclude that for any $\varepsilon_n \downarrow 0$, $\liminf_{n\to\infty} \operatorname{dist}(\theta^\star, S_{\varepsilon_n}) \overset{p}{\to} 0$ conditional on $\mathcal{D}_\infty$ and $\mathcal{E}$. Now, note that for any $\Delta > 0$,

$$\begin{aligned}
\mathbb{P}\left(|\liminf_{n\to\infty} \operatorname{dist}(\theta^\star, S_{\varepsilon_n})| \geq \Delta\right) &= \mathbb{P}\left(|\liminf_{n\to\infty} \operatorname{dist}(\theta^\star, S_{\varepsilon_n})| \geq \Delta \mid \mathcal{E}\right) \\
&= \mathbb{E}\left[\mathbb{P}\left(|\liminf_{n\to\infty} \operatorname{dist}(\theta^\star, S_{\varepsilon_n})| \geq \Delta \mid \mathcal{D}_\infty, \mathcal{E}\right) \mid \mathcal{E}\right] \\
&= \mathbb{E}\left[\mathbb{P}\left(|\liminf_{n\to\infty} \operatorname{dist}(\theta^\star, S_{\varepsilon_n})| \geq \Delta \mid \mathcal{D}_\infty, \mathcal{E}\right)\right]
\end{aligned}$$

where the first and the last equality used since $\mathbb{P}(\mathcal{E}) = 1$. By dominated convergence, the preceeding display goes to 0 as $n \to \infty$.