Causality in Goal Conditioned RL: Return to No Future?

Ivana Malenica
Department of Statistics
Harvard University
Cambridge, MA
imalenica@fas.harvard.edu

Susan A. Murphy
Departments of Statistics and Computer Science
Harvard University
Cambridge, MA
samurphy@fas.harvard.edu

Abstract

The main goal of goal-conditioned RL (GCRL) is to learn actions that maximize the conditional probability of achieving the desired goal from the current state. To improve sample efficiency, GCRL utilizes either 1) imitation learning with expert demonstrations or 2) supervised learning with self-imitation, denoted goal-conditioned RL with supervised learning (GCRL-SL). The GCRL-SL algorithms directly estimate the probability of actions ($A = a$) given the current state ($S = s$), and a future, observed goal ($G = g$) from batch data generated under a behavior policy. Subsequently, the optimal action maximizes an estimate of $P(A \mid S = s, G = g)$. One crucial insight missing from empirical and theoretical work on GCRL relates to the causal interpretation of the policy learned by GCRL algorithms. In this study, we begin exploring a crucial question for ensuring safe and robust decision-making: *What causal biases arise in the GCRL training process and when can these causal biases lead to a poor policy?* Our theoretical and empirical analysis demonstrates that GCRL algorithms can result in learning poor policies when the training data follows particular causal graphs. This issue is particularly problematic when implementing GCRL in environments with potential unmeasured confounding, as often encountered in healthcare and mobile health applications.

1 Introduction

In the field of goal-conditioned policy training, several methods based on imitation learning have been proposed to enhance sample efficiency [20, 21, 10, 4]. Incorporating imitation into goal-conditioned reinforcement learning (GCRL) has demonstrated significant effectiveness [11]. Without loss of generality, the first step involves collecting a dataset of demonstrations performed by an expert. These demonstrations consist of state-action pairs that result in a desired goal (e.g., high final reward) in a given environment. Using the batch data consisting of expert demonstrations, the imitation learning based GCRL algorithm then attempts to learn a policy that can replicate the observed behavior.

When obtaining an adequate number of expert demonstrations is challenging, self-imitation learning has proven to be a useful alternative in GCRL [25, 14, 36, 22]. Including the future information in the training process, such as the observed final state or the reward-to-go, makes any trajectory suited for learning even without expert demonstrations. Recent work has highlighted the potential effectiveness of formulating RL objectives as supervised learning problems with self-imitation [15, 7, 13]. We denote the family of GCRL algorithms which rely on self-imitation as *goal conditioned RL with Supervised Learning* (GCRL-SL). GCRL-SL algorithms learn, from a batch of data collected under a behavior policy, the conditional distribution of actions ($A = a$) given states ($S = s$) and future goals ($G = g$), that is, $P(A \mid S = s, G = g)$. The GCRL-SL policy then selects the action that maximizes the learned conditional distribution of actions given the current state and a desired goal. Conditioning can be on a goal state [9, 11, 24, 15, 23, 13] or on reward values [19, 31, 7].

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

One key insight missing from both empirical and theoretical work in this domain is if, and when, the training process employed by GCRL causes *selection bias* — a central problem for valid causal inference. Selection bias can occur when trajectories are preferentially chosen from the data [33]. More generally, selection bias results from conditioning on a collider variable [5, 16, 17]. When conditioned on, colliders introduce spurious associations between otherwise unrelated variables with a common descendant [26]. The unblocked extraneous information between parents of a collider could result in an agent learning a poor policy (policy based on spurious associations), which affects the achieved return. This type of bias cannot be removed with large amount of data.

In this work, we start to investigate a central question for safe and robust decision making: **What causal biases arise in the GCRL training process and when can these causal biases lead to a poor policy?** We focus on the causal implications of (1) training on only expert demonstrations with good performance, and (2) conditioning on future, observed goals as part of the training process. Our study includes (1) complete graph settings, including the standard MDP and (2) incomplete graphs, where unknown and unmeasured variables might influence the decision process. This paper is motivated by the potential use of GCRL methods in healthcare and mobile health. Here, unmeasured confounding is abundant, and while GCRL might be an appealing alternative to value-based methods, one must be especially careful not to introduce causal biases. Our theoretical and empirical analysis demonstrates that GCRL algorithms can result in learning poorly performing policies when the training data follows particular causal graphs. We believe these findings will be useful to the RL community as they begin to untangle settings in which the preferential inclusion of trajectories and conditioning on the observed future during the training process can be more or less beneficial.

2 Problem Scope

Consider a finite process with horizon T described by a tuple $(\mathcal{S}, \mathcal{A}, \mathcal{R}, P, P_1, T)$. We denote the random state, action, and reward at time t as S_t, A_t, and R_t, where t belongs to the set $[T] = \{1, \ldots, T\}$. The reward at t is a bounded random variable $R_t \in [0, 1]$. Let $P(S_{t+1} \mid S_t = s, A_t = a)$ denote the probability of transitioning to the next state given $S_t = s$ and $A_t = a$. We denote the marginal distribution of the initial state as $P_1(S_1 = s)$. The policy π consists of a sequence of conditional distributions $\{\pi_t\}_{t=1}^{T}$ that map a state to an action. The subscript b is used to emphasize that data is generated under the behavior policy. We express a trajectory τ as $(s_1, a_1, r_1, \ldots, s_T, a_T, r_T)$, where τ_i indexes the i-th trajectory. The two types of goals encountered in GCRL are the *goal state*, representing the state an agent visits in the future ($g_t \in \mathcal{S}$), and the *goal reward*, which is a function of the future collected reward(s) ($g_t \in \mathcal{R}$). The agent is considered successful if it achieves the desired goal by the end of the trajectory. Let $\mathcal{D} = \{\tau_i\}_{i=1}^{N}$ denote a dataset of N collected trajectories, generated by some behavior policy. Assuming final reward is the goal, GCRL-SL algorithms directly estimate $P_b(A_t \mid S_t = s, R_T = g)$ from $\mathcal{D}$ by minimizing the empirical negative log likelihood loss [15, 7, 6]. Therefore, GCRL-SL aims to estimate the full longitudinal structure of the trajectory, as $P(A_t \mid S_t = s, R_T = g) = P_b(A_t, S_t = s, R_T = g)/P_b(S_t = s, R_T = g)$. Subsequently, the action that maximizes an estimate of the conditional probability of actions given the current state and *any* desired goal is deemed optimal. It is important to note that during the training process, the goal is the *observed goal* g (e.g., state, reward) corresponding to collected trajectories in $\mathcal{D}$.

3 Causality of Goal Conditioned RL

To describe the environments we use structural causal models, where actions represent modifications to functional relationships [28]. Specifically, each action *do(a)* on a causal model $\mathcal{M}$ results in a new model $\mathcal{M}_a = (U, V, F_a)$, where V denotes the set of observable variables, while U are the unknown and unobservable variables. For every A, F_a is derived by replacing $f_A \in F$ with a new function that outputs a constant value a defined by *do(a)*. Central to the analysis in this study is the concept of *g-recoverability*, which emphasizes the necessity for effects to be computable from the existing data and assumptions embedded in an augmented causal graph C_g [1]. Another vital concept is of *identifiability*, which signifies the necessity for causal effects to be computable based on a combination of data and assumptions inherent in the causal graph C [28]. The *recoverability* of a causal effect is defined as the combination of g-recoverability with identifiability using the rules of *do*-calculus. Formal definitions are included in the Appendix Section A.

3.1 The g- recoverability in GCRL

To learn the optimal policy, we first need to learn quantities that rely on the effect of action $A_t = a$ on the outcome R_l for $l \geq t$ given the current state. In this section, we study conditions under which such quantities cannot be learned when trajectories are preferentially chosen based on a desired outcome. For example, learning *only* based on expert demonstrations that achieve high rewards, or a high final state, results in selection bias as selection is dependent on the outcome [33]. To illustrate the nature of selection bias, consider Figure 3(a) in the Appendix Section B.1. Let's assume we are interested in obtaining a high reward at the end of the trajectory, $R_2 = 1$. The goal G is the indicator of a high reward ($R_2 = 1$). First, we collect trajectories from an expert and retain trajectories with $R_2 = 1$. The problem is that the batch data reflects $P(\tau \mid G = 1)$. Can we recover the conditional distributions needed for learning the optimal policy? Proposition 1 states that we cannot recover essential components of the likelihood from $P(\tau \mid G = 1)$ under certain graphs [1, 3]. Similar analysis follows when expert trajectories are based on the final state, or sum of all rewards (e.g., Figure 3(b) and (c)). In Theorem 1, we provide general rules for when conditional probabilities are g-recoverable from the data. All proofs are allocated to the Appendix Section B.

Proposition 1. *The causal effect $P(R_2 \mid do(A_t = a), S_t = s)$ for $t = 1, 2$ is not g-recoverable from the causal graph C_g in Figure 3(a).*

Theorem 1. *The distribution $P(A_t \mid S_t)$ is g-recoverable from C_g if and only if $G \perp\!\!\!\perp A_t \mid S_t$. Similarly, $P(R_l \mid A_t, S_t)$ with $t \leq l$ and $P(S_j \mid A_t, S_t)$ with $t < j$ are g-recoverable if and only if $G \perp\!\!\!\perp R_l \mid A_t, S_t$ and $G \perp\!\!\!\perp S_j \mid A_t, S_t$.*

3.2 Recovering Causal Effects in GCRL-SL

In this section, we illustrate instances where key causal quantities for learning the optimal policy become unattainable due to conditioning on specific variables, as observed in GCRL-SL. [18, 17]. Instead of having a separate G variable, recoverability issues equivalently occur due to conditioning on variables known as *colliders*. The common structure of selection bias strives from conditioning on a variable which is caused by two other variables: one that is (or is associated) with an action, and one that is (or is associated) with the outcome. Conditioning on the common outcome of two independent variables induces a spurious association between them, as they are now associated within levels of it. Conditioning on a collider can also distort the association between dependent variables. In the following, we focus on the causal graph depicted in Figure 1(b), which contains an unknown, and unmeasured, variable ϵ. While atypical in RL literature, this causal structure is prevalent in healthcare and mobile health applications. In Figure 1(a), we study a subgraph of Figure 1(b) which is an MDP.

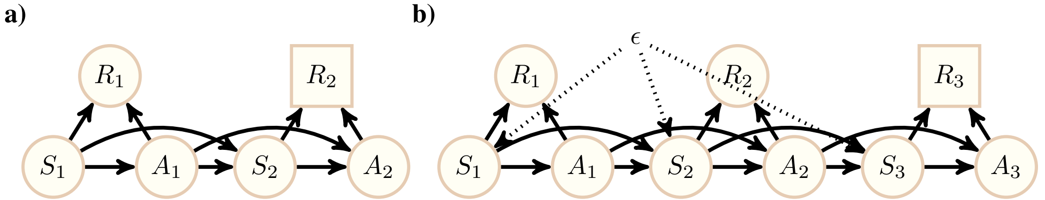

Figure 1: Directed Acyclic Graphs depicting (a) complete causal graph, and (b) incomplete causal graph with unknown, and unmeasured variable ϵ. Variables conditioned on are depicted as rectangles.

We aim to quantify the change in R_l induced by raising A_t one unit, while keeping $S_t = s$ constant. In the language of *do*-calculus, we express this target parameter as $\frac{\delta}{\delta a}\mathbb{E}[R_l \mid do(A_t = a), S_t = s]$. This causal effect can be defined using the partial regression coefficient $\beta_{R_l A_t \cdot S_t}$, representing the slope of the regression line of R_l on A_t given $S_t = s$. In a linear model, $\beta_{R_l A_t \cdot S_t}$ is crucial for learning the optimal policy. The GCRL-SL objective in a linear model then corresponds to $\beta_{A_t R_l \cdot S_t}$, the slope of the regression line of A_t on R_l for $S_t = s$. We employ linear structural equation models and path tracing to illustrate that, for causal graphs depicted in Figure 1 and $l \geq t$, the partial coefficient $\beta_{A_t R_l \cdot S_t}$ is not always equal to $\beta_{R_l A_t \cdot S_t}$ [12, 29]. The closed-form expressions for $\beta_{R_l A_t \cdot S_t}$ and $\beta_{A_t R_l \cdot S_t}$ are provided in the Appendix Section B.3 as part of Lemma 1. Challenges with the GCRL-SL policy arise when $\beta_{A_t R_l \cdot S_t}$ and $\beta_{R_l A_t \cdot S_t}$ display opposing signs, indicating a contradictory relationship between the goal and action. We illustrate this empirically in Section 4.

Lemma 1. *For graphs in Figure 1(a) and (b), the partial coefficient $\beta_{A_t R_l \cdot S_t}$ is not always equal to $\beta_{R_l A_t \cdot S_t}$. For certain path coefficients in linear SEMs, $\beta_{A_t R_l \cdot S_t}$ and $\beta_{R_l A_t \cdot S_t}$ have different signs.*

In Lemma 2, we establish that the causal effect of an action at time t, denoted as $do(A_t = a)$, on R_l for $l \geq t$ given the current state cannot be recovered when employing GCRL-SL methods for the causal graph in Figure 1. Moreover, conditioning on R_T in a causal graph depicted in Figure 1 opens a back-door path from S_t to A_t because $(S_t \perp\!\!\!\perp A_t \mid R_T)_{C_{\underline{S}_t}}$ is violated for $t \in [T]$. As a result, the GCRL-SL policy produces actions influenced by spurious associations, even when actions have no impact on the rewards and states. The proofs for Lemma 2 and Corollary 1 are available in the Appendix Section B.4.

Lemma 2. *Given the causal graph depicted in Figure 1 with unmeasured and unknown variable ϵ, $P(R_l \mid do(A_t = a), S_t = s)$ for $l \geq t$ is not recoverable using GCRL-SL algorithms.*

Corollary 1. *For the causal graphs in Figure 1, $P(A_t \mid S_t = s)$ is not g-recoverable from $P(A_t \mid S_t = s, R_l = g)$.*

4 Experiments

We consider two environments for which GCRL-SL policies perform poorly due to causal biases: (1) complete graph in Figure 1(a) denoted as CG1, and (2) incomplete graph in Figure 1(b) denoted IG1. For clarity, we consider finite trajectories of length $T = 7$ with linear dynamics. The action space is binary for both, while the CG1 data-generating process is non-stationary. We run multiple experiments to show that performance differences between GCRL-SL and value-based RL (e.g., fitted Q iteration) stem from GCRL-SL methods learning poor policies due to causal biases introduced during the training phase. Reported return is an average over 100 random iterations, each evaluated over 20 validation trajectories. More details on the experiments and results are available in the Appendix section C, D and E.

Why does GCRL under-perform? In CG1, GCRL-SL policy assigns $A = 1$ at early time-points due to spurious associations with actions at later time-points, where $A = 1$ does lead to a higher return. In IG1, the GCRL-SL policy consistently assigns $A = 1$, despite the negative impact of actions on the outcome. This behavior is influenced by the spurious association with the exogenous variable ϵ, which does have a positive impact on the return.

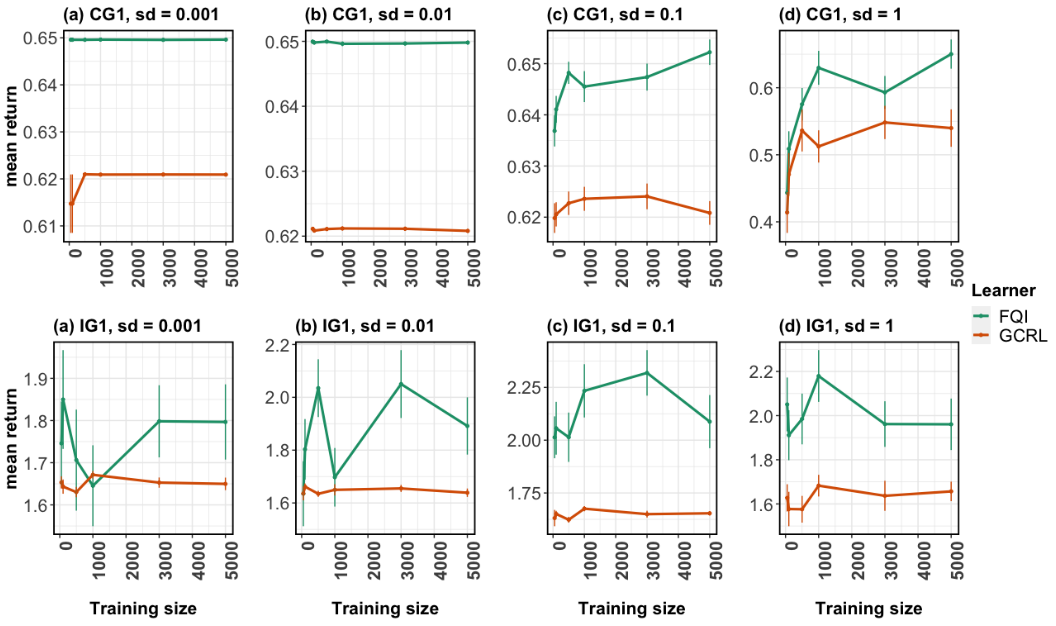

Figure 2: Mean return and standard error for CG1 (upper panels (a)-(d)) and IG1 (lower panels (a)-(d)) over 100 Monte Carlo iterations. Variations in the level of stochasticity of the data-generating process is indicated by distinct standard error levels, denoted as σ.

References

[1] Elias Bareinboim and Judea Pearl. Controlling selection bias in causal inference. In Neil D. Lawrence and Mark Girolami, editors, *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pages 100–108, La Palma, Canary Islands, 21–23 Apr 2012. PMLR.

[2] Elias Bareinboim and Jin Tian. Recovering causal effects from selection bias. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1), Mar. 2015.

[3] Elias Bareinboim, Jin Tian, and Judea Pearl. Recovering from selection bias in causal and statistical inference. *Probabilistic and Causal Inference*, 2014.

[4] Mark Beliaev, Andy Shih, Stefano Ermon, Dorsa Sadigh, and Ramtin Pedarsani. Imitation learning by estimating expertise of demonstrators. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvari, Gang Niu, and Sivan Sabato, editors, *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 1732–1748. PMLR, 17–23 Jul 2022.

[5] Joseph Berkson. Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2(3):47–53, 1946.

[6] David Brandfonbrener, Alberto Bietti, Jacob Buckman, Romain Laroche, and Joan Bruna. When does return-conditioned supervised learning work for offline reinforcement learning? In *NeurIPS*, 2022.

[7] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Michael Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 15084–15097, 2021.

[8] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, New York, NY, USA, 2016. ACM.

[9] Felipe Codevilla, Matthias Müller, Antonio López, Vladlen Koltun, and Alexey Dosovitskiy. End-to-end driving via conditional imitation learning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 4693–4700, 2018.

[10] Christopher R. Dance, Julien Perez, and Théo Cachet. Demonstration-conditioned reinforcement learning for few-shot imitation. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 2376–2387. PMLR, 18–24 Jul 2021.

[11] Yiming Ding, Carlos Florensa, Mariano Phielipp, and P. Abbeel. Goal-conditioned imitation learning. *ArXiv*, abs/1906.05838, 2019.

[12] Otis Dudley Duncan. Chapter 4: Structural coefficients in recursive models. In Otis Dudley Duncan, editor, *Introduction to Structural Equation Models*, Studies in Population, pages 51–66. Academic Press, San Diego, 1975.

[13] Scott Emmons, Benjamin Eysenbach, Ilya Kostrikov, and Sergey Levine. RvS: What is essential for offline RL via supervised learning? *CoRR*, abs/2112.10751, 2021.

[14] Johan Ferret, Olivier Pietquin, and Matthieu Geist. Self-imitation advantage learning. In *Proceedings of the 20th International Conference on Autonomous Agents and MultiAgent Systems*, AAMAS '21, page 501–509, Richland, SC, 2021. International Foundation for Autonomous Agents and Multiagent Systems.

[15] Dibya Ghosh, Abhishek Gupta, Ashwin Reddy, Justin Fu, Coline Manon Devin, Benjamin Eysenbach, and Sergey Levine. Learning to reach goals via iterated supervised learning. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*, 2021.

[16] S. Greenland. Quantifying biases in causal models: classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, May 2003.

[17] M.A. Hernan, S. Hernandez-Díaz, and J. M. Robins. A structural approach to selection bias. *Epidemiology*, 15(5):615–625, Sep 2004.

[18] Miguel A. Hernán, Sonia Hernández-Díaz, Martha M. Werler, and Allen A. Mitchell. Causal Knowledge as a Prerequisite for Confounding Evaluation: An Application to Birth Defects Epidemiology. *American Journal of Epidemiology*, 155(2):176–184, 01 2002.

[19] Aviral Kumar, Xue Bin Peng, and Sergey Levine. Reward-conditioned policies, 2019.

[20] Sang-Hyun Lee and Seung-Woo Seo. Learning compound tasks without task-specific knowledge via imitation and self-supervised learning. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 5747–5756. PMLR, 13–18 Jul 2020.

[21] Youngwoon Lee, Andrew Szot, Shao-Hua Sun, and Joseph J Lim. Generalizable imitation learning from observation via inferring goal proximity. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 16118–16130. Curran Associates, Inc., 2021.

[22] Yao Li, YuHui Wang, and XiaoYang Tan. Self-imitation guided goal-conditioned reinforcement learning. *Pattern Recognition*, 144:109845, 2023.

[23] Minghuan Liu, Menghui Zhu, and Weinan Zhang. Goal-conditioned reinforcement learning: Problems and solutions, 2022.

[24] Corey Lynch, Mohi Khansari, Ted Xiao, Vikash Kumar, Jonathan Tompson, Sergey Levine, and Pierre Sermanet. Learning latent plans from play, 2019.

[25] Junhyuk Oh, Yijie Guo, Satinder Singh, and Honglak Lee. Self-imitation learning. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 3878–3887. PMLR, 10–15 Jul 2018.

[26] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

[27] Judea Pearl. Causal diagrams for empirical research. *Biometrika*, 82(4):669–688, 1995.

[28] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, 2009.

[29] Judea Pearl. Linear models: A useful “microscope” for causal analysis. 2013.

[30] Judea Pearl and Azaria Paz. Confounding equivalence in causal inference. *Journal of Causal Inference*, 2(1):75–93, 2014.

[31] Rupesh Kumar Srivastava, Pranav Shyam, Filipe Mutz, Wojciech Jaśkowski, and Jürgen Schmidhuber. Training agents using upside-down reinforcement learning, 2021.

[32] M.J. van der Laan, E.C. Polley, and A.E. Hubbard. Super learner. Technical Report Working Paper 222., U.C. Berkeley Division of Biostatistics Working Paper Series, 07 2007.

[33] Christopher Winship and Robert D. Mare. Models for sample selection bias. *Annual Review of Sociology*, 18:327–350, 1992.

[34] S.N Wood. *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC, 2 edition, 2017.

[35] Marvin N. Wright and Andreas Ziegler. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(1):1–17, 2017.

[36] Rui Yang, Yiming Lu, Wenzhe Li, Hao Sun, Meng Fang, Yali Du, Xiu Li, Lei Han, and Chongjie Zhang. Rethinking goal-conditioned supervised learning and its connection to offline RL, 2022.

A Background and Notation

Let X, Y and Z be arbitrary disjoint sets of nodes in a causal graph C. We denote V and U as the set of all observable and unobservable variables. A convenient way of characterizing the set of distributions compatible with the causal graph C is to list the set of all (conditionally) independent variables that such distribution must satisfy. We make use of the *d-separation* criterion for reading constraints imposed over distributions in the induced graph, as stated in Definition 1 [26, 28].

Definition 1 (d-separation)**.** *A path p is said to be d-separated by a set of vertices in Z if and only if*

1. *p contains a chain $i \rightarrow m \rightarrow j$ or a fork $i \leftarrow m \rightarrow j$ such that m is in Z (at least one arrow-emitting node in Z), or*
2. *p contains an inverted fork $i \rightarrow m \leftarrow j$ such that m is not in Z, and no descendant of m is in Z (at least one collision node is outside Z and its descendants).*

A set Z d-separates X from Y if and only if Z blocks every path from node in X to a node in Y.

We also make use of *do*-calculus, which is a collection of syntactic rules that permit the manipulation of causal expressions involving the *do*-operator [27]. Let $C_{\bar{X}}$ be the graph obtained by removing all arrows pointing to nodes in X from the original graph C. The $C_{\underline{X}}$ is a graph obtained by removing all arrows emerging from nodes in X. An expression of the type $Q = P(y \mid do(x), z)$ is compatible with C if the interventional distribution described by Q can be generated by parameterizing the graph with a set of functions and exogenous variables. The following inference rules of the proposed calculus (rules of *do*-calculus) are valid for every interventional distribution compatible with C [28]:

Rule 1 (Insertion/deletion of observations)

$$P(y \mid do(x), z, w) = P(y \mid do(x), w)$$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{C_{\bar{X}}}$.

Rule 2 (Action/observation exchange)

$$P(y \mid do(x), do(z), w) = P(y \mid do(x), z, w)$$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{C_{\bar{X},\underline{Z}}}$.

Rule 3 (Insertion/deletion of actions)

$$P(y \mid do(x), do(z), w) = P(y \mid do(x), w)$$

if $(Y \perp\!\!\!\perp Z \mid X, W)_{C_{\bar{X},Z(\bar{W})}}$ and $Z(\bar{W})$ is the set of Z-nodes that are not ancestors of any W-node in $C_{\bar{X}}$.

We provide the formal definitions of *identifiability* and *g-recoverability* in Definition 2 and 3. Additionally, Theorem 2 from [2] outlines *recoverability* as a combination of g-recoverability of conditional distributions with identifiability employing the rules of *do*-calculus.

Definition 2 (Identifiability of Causal Effects)**.** *The causal effect of an action do(X=x) on a set of variables Y is said to be identifiable from P in a causal graph C if $P(Y \mid$ do(x)$)$ is uniquely computable from $P(v)$ in any model that induces C.*

Definition 3 (g-Recoverability)**.** *Given a causal graph C_g augmented with a node G encoding the selection mechanism, the distribution $Q = P(y \mid x)$ is said to be g-recoverable from selection biased data in C_g if the assumptions embedded in the causal model renders Q expressible in terms of the distribution under selection bias $P(v \mid G = 1)$. Formally, for every two probability distributions P_1 and P_2 compatible with C_g, $P_1(v \mid G = 1) = P_2(v \mid G = 1) > 0$ implies $P_1(y \mid x) = P_2(y \mid x)$.*

Theorem 2 (Recoverability: g-recoverability and identifiability via *do*-calculus)**.** *The causal effect $Q = P(y \mid do(x))$ is recoverable from selection biased data if using the rules of the do-calculus, Q is reducible to an expression in which there is no do-operator, and g-recoverability can be determined.*

Finally, Definition 4 extends the backdoor condition to selection bias problems by identifying a set of variables W which are 1) backdoor admissible, and 2) ensure recoverability from selection bias [30, 3]. The Corollary 2, proved in [3], gives a graphical condition for recovering causal effects which generalizes the back-door adjustment.

Definition 4 (Selection-backdoor criterion)**.** *Let a set W of variables be partitioned into $W^+ \cup W^-$ such that W^+ contains all non-descendants of X and W^- the descendants of X. The set W is said to satisfy the selection backdoor criterion (g-backdoor) relative to an ordered pairs of variables (X, Y) in a graph C_g if W^+ and W^- satisfy the following:*

1. *W^+ blocks all back door paths from X to Y.*
2. *X and W^+ block all paths between W^- and Y.*
3. *X and W block all paths between G and Y.*
4. *Measurements of W for the entire population are available.*

Corollary 2 (Selection-backdoor adjustment)**.** *If a set W satisfies the g-backdoor criterion relative to the ordered pair (X, Y), then the effect of X on Y is identifiable and g-recoverable.*

B Proofs

B.1 Proposition 1

Proof. Let P_1 be compatible with the causal graph C_g depicted in Figure 3(a). We construct another causal model, such that P_2 corresponds to the subgraph $C_g \backslash \{R_2 \to G\}$. We set the parameters of P_1 through its factors, and notice that $(V \perp\!\!\!\perp G)_{P_2}$, where V is the set of all variables in C_g without the selection variable. As $P_2(V \mid G) = P_2(V)$, we compute the parameters of P_2 by enforcing $P_1(V \mid G) = P_2(V)$. We assume all variables are binary for ease of derivations, as recoverability should hold for any parametrization.

We focus on the recoverability of the probability of R_2 conditional on action and state at $t = 2$. The same arguments follows for the probability of R_2 conditional on action and state at $t = 1$. First we observe that for the causal graph C_g in Figure 3(a), $Y \in Pa(G)$, and there exists no sets of variables for which one can d-separate G from R_2. We can write the conditional distribution corresponding to the second causal model at $S_2 = s$ under MDP dynamics as

$$\begin{aligned} P_2(R_2 \mid A_2, S_2 = s) &= P_1(R_2 \mid A_2, S_2 = s, G = 1) \\ &= \frac{P_1(R_2, A_2, S_2 = s, G = 1)}{P_1(A_2, S_2 = s, G = 1)} \\ &= \frac{P_1(R_2 \mid A_2, S_2 = s) P_1(G = 1 \mid R_2)}{P_1(G = 1 \mid R_2) P_1(R_2 \mid A_2, S_2 = s) + P_1(G = 1 \mid \tilde{R}_2) P_1(\tilde{R}_2 \mid A_2, S_2 = s)}. \end{aligned}$$

Let $P_1(R_2 \mid A_2, S_2 = s) = 1/2$ and $P_1(\tilde{R}_2 \mid A_2, S_2 = s) = 1/2$. Further, let $P_1(G = 1 \mid R_2) = \alpha$ and $P_1(G = 1 \mid \tilde{R}_2) = \beta$, where $0 < \alpha$, $\beta < 1$ and $\alpha \neq \beta$. The result follows as under this parametrization $P_2(R_2 \mid A_2, S_2 = s) = \alpha/(\alpha + \beta)$, while $P_1(R_2 \mid A_2, S_2 = s) = 1/2$. Therefore, we have that

$$\begin{aligned} P_2(R_2 \mid do(A_2 = a), S_2 = s) &= P_2(R_2 \mid A_2 = a, S_2 = s) \\ &\neq P_2(R_2 \mid A_2 = a, S_2 = s, G = 1), \end{aligned}$$

where the first equality follows from the second rule of *do*-calculus as $(R_2 \perp\!\!\!\perp A_2 \mid S_2)_{C_{\underline{A_2}}}$. By Theorem 2, $P_2(R_2 \mid do(A_2 = a), S_2 = s)$ is not recoverable. The same argument follows for $P_2(R_2 \mid do(A_1 = a), S_1 = s)$. Finally, note that G and R_1 are d-separated by (A_1, S_1). Therefore, $G \perp\!\!\!\perp R_1 \mid A_1, S_1$, and $P(R_1 \mid A_1 = a, S_1 = s)$ is g-recoverable from C_g. By the second rule of *do*-calculus, the causal effect $P(R_1 \mid do(A_1 = a), S_1 = s)$ is recoverable. □

B.2 Theorem 1

Proof. We show that the distribution of A_t given S_t is g-recoverable from C_g if and only if $G \perp\!\!\!\perp A_t \mid S_t$. The same argument can be used to show $P(R_l \mid A_t, S_t)$ for $t \leq l$ is g-recoverable from C_g if and only if $G \perp\!\!\!\perp R_l \mid A_t, S_t$, and $P(S_j \mid A_t, S_t)$ for $t < j$ is g-recoverable from C_g if and only if $G \perp\!\!\!\perp S_j \mid A_t, S_t$. The argument follows from [3], adapted to longitudinal settings.

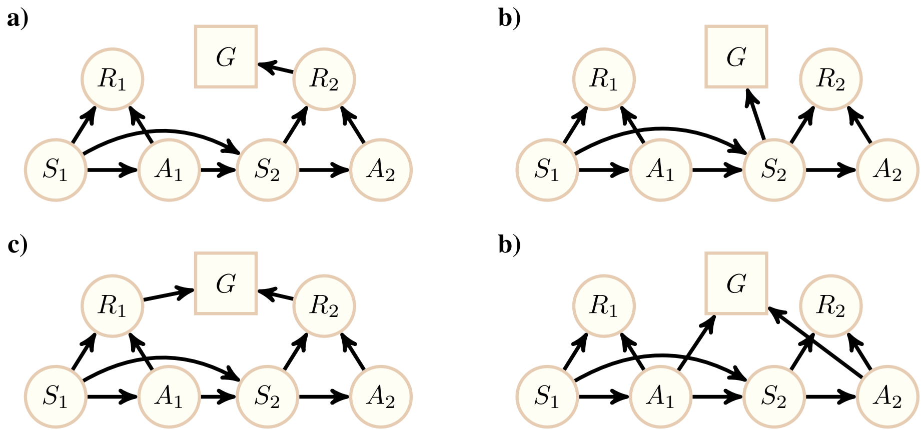

Figure 3: Directed Acyclic Graph (DAG) where the selection variable G is an indicator of (a) a high final reward, (b) a high final state, (c) a high reward-to-go and (d) action selection. Variables conditioned on are depicted as rectangles.

(if) We note that if S_t d-separates G and A_t in C_g, $P(A_t \mid S_t)$ is g-recoverable.

(only if) To establish necessity, we must show that for all graphs where d-separation fails, none of these structures enable g-recoverability.

Let P_1 be compatible with the causal graph C_g. We construct another causal model, such that P_2 corresponds to a subgraph C_2 where all edges pointing to G are removed. Let V denote the set of all variables in C_g without G. Given a Markovian data-generating model, P_1 can be parameterized through its factors [28]. As $(V \perp\!\!\!\perp G)_{P_2}$, it follows that $P_2(V \mid G = g) = P_2(V)$. We can compute the parameters of P_2 by enforcing $P_1(V \mid G = g) = P_2(V)$. We assume all variables are binary for ease of derivations, as recoverability should hold for any parametrization. We assume S_t is directly connected to A_t, meaning there are no nodes in between S_t and A_t $\forall t \in [T]$. We proceed to show that whenever there is an open path between G and A_t that is not blocked by S_t, P_1 and P_2 can be constructed such that $P_1(V \mid G = 1) = P_2(V \mid G = 1)$, but $P_1(A_t \mid S_t) \neq P_2(A_t \mid S_t)$.

Case 1: A_t is a parent of G. Case 1 corresponds a graph where $A_t \in Pa(G)$, and G is not separable from A_t in C_g. We follow the same construction given in Lemma 1. We can write the conditional distribution corresponding to the second causal model under MDP dynamics as

$$\begin{aligned} P_2(A_t \mid S_t) = P_1(A_t \mid S_t, G = 1) &= \frac{P_1(A_t, S_t, G = 1)}{P_1(S_t, G = 1)} \\ &= \frac{P_1(A_t \mid S_t) P_1(G = 1 \mid A_t)}{P_1(G = 1 \mid A_t) P_1(A_t \mid S_t) + P_1(G = 1 \mid \tilde{A}_t) P_1(\tilde{A}_t \mid S_t)}, \end{aligned} \tag{1}$$

where the first equality is by construction. Consider a subgraph C^*, such that all nodes $V \backslash \{S_t, A_t, G\}$ in C_g are disconnected from $\{S_t, A_t, G\}$, and we can parametrize the complete model. Note that all parameters of P_2 are free, and can be chosen to match P_1. Still, for almost all parametrizations, $P_1(A_t \mid S_t)$ and $P_2(A_t \mid S_t)$ in Equation 1 will not be the same. Let $P_1(G = 1 \mid A_t) = \alpha$, $P_1(G = 1 \mid \tilde{A}_t) = \beta$, where $\alpha \neq \beta$, $0 < \alpha$ and $\beta < 1$. We set every distribution but the selection node equal to $1/2$, so $P_1(A_t \mid S_t) = 1/2$. The result follows as $P_2(A_t \mid S_t) = \alpha/(\alpha + \beta)$, while $P_1(A_t \mid S_t) = 1/2$.

Case 2: path from A_t to G passes through an offspring of A_t. Case 2 corresponds to a graph where there is a directed path p from A_t to G, which does not pass through S_t. As such, G is not separable from A_t in C_g, and S_t cannot d-separate G and A_t. Consider R_t as the immediate child of A_t in path p. For simplicity, we assume the segment from R_t to G is of length one, although the distance between R_t and G can be otherwise arbitrary. This assumption is valid by construction, which ensures only chains exist along this segment. We can write the conditional distribution corresponding to the second causal model under MDP dynamics as

$$\begin{aligned} P_2(A_t \mid S_t) = P_1(A_t \mid S_t, G=1) &= \frac{P_1(A_t, S_t, G=1)}{P_1(S_t, G=1)} \\ &= \frac{P_1(A_t \mid S_t) \sum_{R_t} \phi_{R_t}(A_t) f(R_t)}{P_1(A_t \mid S_t) \sum_{R_t} \phi_{R_t}(A_t) f(R_t) + P_1(\tilde{A}_t \mid S_t) \sum_{R_t} \phi_{R_t}(\tilde{A}_t) f(R_t)}, \end{aligned} \tag{2}$$

where $\phi_{R_t}(A_t) = P_1(R_t \mid A_t)$ and $f(R_t) = P_1(G = 1 \mid R_t)$. Consider a subgraph C^*, such that all nodes $V \backslash \{S_t, A_t, R_t, G\}$ in C_g are disconnected from $\{S_t, A_t, R_t, G\}$, and we can parametrize the complete model. We first set $P_1(A_t \mid S_t) = 1/2$ for all values of A_t and S_t. Using the same construction presented in [3] with $k = 0$, we set $f(R_t) = 7/12 + \epsilon$ and $f(\tilde{R_t}) = 7/12 - \epsilon$ for $0 < \epsilon \leq 1$, and let $\phi_{R_t}(A_t) = 1/3$ and $\phi_{R_t}(\tilde{A}_t) = 3/4$. Substituting back into Equation 2, we get that $P_2(A_t \mid S_t) = 1/2 - (2/7)\epsilon$, which is never equal to $P_1(A_t \mid S_t) = 1/2$.

Case 3: path from A_t to G passes through an ancestor of A_t. Let W represent the set of nodes that are ancestors of A_t. In Case 3, we consider a graph where $W \backslash S_t$ is not d-separated from A_t in C_g, and there exists a path p from $Z_t \in W \backslash S_t$ that cannot be blocked by S_t. Two cases must be considered: (a) p is a directed path from Z_t to A_t that does not pass through S_t, and (b) p contains converging arrows into S_t.

Case 3a considers graphs where p is a directed path from Z_t to A_t. For simplicity, we assume the segment from Z_t to G and Z_t to A_t is of length one. This assumption is possible as, by construction, there can be only chains along these segments. We note that the distance between Z_t to G and Z_t to A_t can be otherwise arbitrary. Under this configuration, we have the flexibility to modify C_g while remaining in the same equivalence class by reversing the direction in p in a way that Z_t ceases to be an ancestor of A_t. Consequently, Case 3a simplifies to Case 2.

Case 3b considers a graph where p contains converging arrows into S_t. Let W_t denote an ancestor which, together with Z_t, has converging arrows into S_t ($W_t \to \cdots \to S_t \leftarrow \cdots \leftarrow Z_t$). We again assume that segments $W_t \to \cdots \to A_t$, $W_t \to \cdots \to S_t$, $Z_t \to \cdots \to S_t$ and $Z_t \to \cdots \to G$ are of length one. We can write the conditional distribution corresponding to the second causal model as

$$\begin{aligned} P_2(A_t \mid S_t) = P_1(A_t \mid S_t, G=1) &= \frac{P_1(A_t, S_t, G=1)}{P_1(S_t, G=1)} \\ &= \frac{\sum_{W_t} \phi_{W_t}(A_t) f(Z_t)}{\sum_{W_t} \phi_{W_t}(A_t) f(Z_t) + \sum_{W_t} \phi_{W_t}(\tilde{A}_t) f(Z_t)}, \end{aligned} \tag{3}$$

where $\phi_{W_t}(A_t) = P_1(A_t \mid W_t) P_1(W_t)$ and $f(Z_t) = \sum_{Z_t} P_1(S_t \mid Z_t, W_t) P(Z_t) P_1(G = 1 \mid Z_t)$. Consider a subgraph C^*, such that all nodes $V \backslash \{W_t, S_t, A_t, Z_t, G\}$ in C_g are disconnected from $\{W_t, S_t, A_t, Z_t, G\}$, and we can parametrize the complete model. Consider the following parametrization where $\phi_{W_t}(A_t) = 1/3$, $\phi_{\tilde{W}_t}(A_t) = 1/3$, $\phi_{\tilde{W}_t}(\tilde{A}_t) = 2/9$ and $\phi_{W_t}(\tilde{A}_t) = 1/9$. We also let $P_1(Z_t) = 1/2$, $P_1(\tilde{Z}_t) = 1/2$, $P_1(S_t \mid Z_t, W_t) = 1/2 + \epsilon$, $P_1(S_t \mid \tilde{Z}_t, W_t) = 1/2 - \epsilon$, $P_1(S_t \mid Z_t, \tilde{W}_t) = 1/2$ and $P_1(S_t \mid \tilde{Z}_t, \tilde{W}_t) = 1/2$ for $0 < \epsilon < 1/2$. Finally, let $P_1(G = 1 \mid Z_t) = \alpha$ and $P_1(G = 1 \mid \tilde{Z}_t) = \beta$ where $\alpha > \beta$. With some algebra, we derive that $P_1(A_t \mid S_t) = 2/3$ while $P_2(A_t \mid S_t) = \frac{2}{3}(\frac{\alpha+\beta+\epsilon(\alpha-\beta)}{\alpha+\beta+8/9\epsilon(\alpha-\beta)})$. □

B.3 Lemma 1

Proof. Let $\beta_{R_l A_t \cdot S_t}$ denote the partial regression coefficient, or the slope of the regression line of R_l on A_t given $S_t = s$. We denote $\sigma_{R_l A_t}$ as the covariance of R_l and A_t. We assume all variables are standardized and S_t is a singleton.

Case 1: Causal Graph in Figure 1(a). Consider the following Linear Structural Equation model (SEM) corresponding to the causal graph in Figure 1(a) with $T = 2$:

$$\begin{aligned}
S_1 &= U_{S_1} \\
A_1 &= \beta_1 S_1 + U_{A_1} \\
R_1 &= \delta_1 S_1 + \nu_1 A_1 + U_{R_1}, \\
S_2 &= \gamma_1 S_1 + \alpha_1 A_1 + U_{S_2}, \\
A_2 &= \beta_2 S_2 + \eta_1 A_1 + U_{A_2} \\
R_2 &= \delta_2 S_2 + \nu_2 A_2 + U_{R_2}.
\end{aligned} \tag{4}$$

All SEM coefficients in Equation 4 carry causal information, and are refered to as path coefficients. For example, ν_1 stands for the change in R_1 induced by raising A_1 one unit, while keeping all other variables constant. In terms of do-calculus, ν_1 can be interpreted as the slope $\nu_1 = \delta/\delta a \mathbb{E}[R_1 \mid do(a), do(s)]$. Using Wright's path-tracing rules and d-separation in a standardized model where all variables are unity we have that

$$\begin{aligned}
\beta_{R2A1\cdot S1} &= \frac{\alpha_1\delta_2(1-\beta_1^2) + \alpha_1\beta_2\nu_2(1-\beta_1^2) + \eta_1\nu_2(1-\beta_1^2)}{1-\beta_1^2} \\
&= \alpha_1\delta_2 + \alpha_1\beta_2\nu_2 + \eta_1\nu_2.
\end{aligned} \tag{5}$$

The closed-form expression for the partial coefficient $\beta_{A1R2\cdot S1}$ is

$$\beta_{A1R2\cdot S1} = \frac{(\alpha_1\delta_2 + \alpha_1\beta_2\nu_2 + \eta_1\nu_2)(1-\beta_1^2)}{1 - (\beta_1\alpha_1\delta_2 + \beta_1\alpha_1\beta_2\nu_2 + \beta_1\eta_1\nu_2 + \gamma_1\delta_2 + \gamma_1\beta_2\nu_2)^2}. \tag{6}$$

Let $\beta_1 = 0.01$ and $\gamma_1 = 10$, while the rest of the path coefficients are 0.5. Then $\beta_{A1R2\cdot S1} < 0$ while $\beta_{R2A1\cdot S1} > 0$.

Case 2: Causal Graph in Figure 1(b). Consider the following Linear Structural Equation model (SEM) corresponding to the causal graph in Figure 1(b) with $T = 2$:

$$\begin{aligned}
S_1 &= \xi_1\epsilon + U_{S_1} \\
A_1 &= \beta_1 S_1 + U_{A_1} \\
R_1 &= \delta_1 S_1 + \nu_1 A_1 + U_{R_1}, \\
S_2 &= \gamma_1 S_1 + \alpha_1 A_1 + \xi_2\epsilon + U_{S_2}, \\
A_2 &= \beta_2 S_2 + \eta_1 A_1 + U_{A_2} \\
R_2 &= \delta_2 S_2 + \nu_2 A_2 + U_{R_2}.
\end{aligned} \tag{7}$$

Using Wright's path-tracing rules and d-separation in a standardized model where all variables are unity we have that

$$\begin{aligned}
\beta_{R2A1\cdot S1} &= \frac{\alpha_1\delta_2(1-\beta_1^2) + \alpha_1\beta_2\nu_2(1-\beta_1^2) + \eta_1\nu_2(1-\beta_1^2)}{1-\beta_1^2} \\
&= \alpha_1\delta_2 + \alpha_1\beta_2\nu_2 + \eta_1\nu_2.
\end{aligned} \tag{8}$$

The closed-form expression for the partial coefficient $\beta_{A1R2\cdot S1}$ is

$$\beta_{A1R2\cdot S1} = \frac{(\alpha_1\delta_2 + \alpha_1\beta_2\nu_2 + \eta_1\nu_2)(1-\beta_1^2)}{1 - (\beta_1\alpha_1\delta_2 + \beta_1\alpha_1\beta_2\nu_2 + \beta_1\eta_1\nu_2 + \gamma_1\delta_2 + \gamma_1\beta_2\nu_2 + \xi_1\xi_2\delta_2 + \xi_1\xi_2\beta_2\nu_2)^2}. \tag{9}$$

Let $\beta_1 = 0.01$ and $\gamma_1 = 10$, while the rest of the path coefficients are 0.5. Then $\beta_{A1R2\cdot S1} < 0$ while $\beta_{R2A1\cdot S1} > 0$.

□

B.4 Lemma 2 and Corollary 1

Proof. We consider a simplified case with $T = 3$, as recoverability should hold for any T. Let the goal reward at T be g, such that if $R_3 = g$, we have achieved the desired goal. For the graph illustrated in Figure 4(a), we can write the following structural equations model:

$$\begin{aligned}
\epsilon &= f_\epsilon(U_\epsilon) \\
S_1 &= f_{S_1}(\epsilon, U_{S_1}) \\
A_1 &= f_{A_1}(S_1, U_{A_1}) \\
R_1 &= f_{R_1}(S_1, A_1, U_{R_1}) \\
S_t &= f_{S_t}(S_t, A_t, \epsilon, U_{S_t}) && \text{for } 1 < t \leq T \\
A_t &= f_{A_t}(S_t, A_t, U_{A_t}) && \text{for } 1 < t \leq T \\
R_t &= f_{R_t}(S_t, A_t, U_{R_t}) && \text{for } 1 < t \leq T
\end{aligned}$$

where $U = (U_{S_1}, U_{A_1}, U_{R_1}, U_{S_2}, U_{A_2}, U_{R_2}, U_{S_3}, U_{A_3}, U_{R_3})$ is the set of all independent and unobservable variables. We denote the set of all observable variables as $V = (S_1, A_1, R_1, S_2, A_2, R_2, S_3, A_3, R_3)$. Given an input (U, V), structural equations f_{S_t}, f_{A_t} and f_{R_t} for each t deterministically assign a value to each of the nodes. The structural equations do not restrict the functional form of causal relationships.

We start with what the policy estimand should be at $t = 2$, then attempt to derive from it the GCRL-SL objective, $P(A_2 \mid S_2 = a, R_3 = g)$. The target optimal policy at $t = 2$ can then be written as an argmax over a of the following quantity:

$$\frac{P_b(f_{R_3}(f_{S_3}(s, a, \epsilon, U_{R_3}), f_{A_3}(f_{S_3}(s, a, \epsilon, U_{R_3}), a, U_{A_3}), U_{R_3}) = g \mid f_{S_2}(\epsilon, S_1, A_1, U_{S_2}) = s)}{\sum_{\bar{a}} {}_b(f_{R_3}(f_{S_3}(s, \bar{a}, \epsilon, U_{R_3}), f_{A_3}(f_{S_3}(s, \bar{a}, \epsilon, U_{R_3}), \bar{a}, U_{A_3}), U_{R_3}) = g \mid f_{S_2}(\epsilon, S_1, A_1, U_{S_2}) = s)}. \tag{10}$$

Let $A_2 = a^*$. It follows that

$$\frac{P_b(\{A_2 = a^*\} \cap f_{R_3}(f_{S_3}(s, a^*, \epsilon, U_{R_3}), f_{A_3}(f_{S_3}(s, a, \epsilon, U_{R_3}), a^*, U_{A_3}), U_{R_3}) = g \mid f_{S_2}(\epsilon, S_1, A_1, U_{S_2}) = s)}{\sum_{\bar{a}} P_b(\{A_2 = a^*\} \cap f_{R_3}(f_{S_3}(s, \bar{a}, \epsilon, U_{R_3}), f_{A_3}(f_{S_3}(s, \bar{a}, \epsilon, U_{R_3}), \bar{a}, U_{A_3}), U_{R_3}) = g \mid f_{S_2}(\epsilon, S_1, A_1, U_{S_2}) = s)}$$
$$\neq P_b(A_2 = a^* \mid S_2 = s) P_b(R_3 = g \mid A_2 = a^*, S_2 = s) / P_b(R_3 = g, S_2 = s).$$

The second inequality follows from the fact that $R_3 \not\perp\!\!\!\perp A_2 \mid S_2$ and the backdoor-path $R_3 \rightarrow S_3 \rightarrow \epsilon \leftarrow S_1 \rightarrow A_1 \rightarrow A_2$ is open. Therefore, we cannot write the joint over $f_{A_2}(\cdot)$ and $f_{R_3}(\cdot)$ conditional on $f_{S_2}(\cdot)$ as a product of conditional probabilities, and $P_b(R_3 = g \mid do(A_2 = a^*), S_2 = s)$ cannot be written in terms of the observed data, $P_b(R_3 = g \mid A_2 = a^*, S_2 = s)$. Therefore, the GCRL-SL objective does not correspond to the desired policy estimand.

Under the dynamics illustrated in Figure 4(a), $P_b(A_t \mid S_t = s)$ is not g-recoverable from $P_b(A_t \mid S_t = s, R_l = g)$ for $l \geq t$ by Theorem 1. Consider a subgraph of Figure 4(a), as illustrated in Figure 4(b). Even when actions have no effect on the future rewards and states, $P_b(A_t \mid S_t = s)$ is still not g-recoverable from $P_b(A_t \mid S_t = s, R_l = g)$, as $R_l \perp\!\!\!\perp A_t \mid S_t$ does not apply in the causal graph illustrated in Figure 4(b). □

B.5 Lemma 3

In Lemma 3, we show that if the behavior policy is directly re-weighed based on the distribution of future returns, no causal biases occur.

Lemma 3. *In a MDP illustrated in Figure 5(a) $P(R_l \mid do(A_t = a), S_t = s)$ for $l \geq t$ is recoverable under sequential randomization and corresponds to the GCRL-SL objective.*

Proof. We consider a simplified case with $T = 2$, as recoverability should hold for any T. Let the goal reward at T be g, such that if $R_2 = g$, we have achieved the desired goal. For the graph illustrated in Figure 5(a), we can write the following structural equations model:

$$\begin{aligned}
S_1 &= f_{S_1}(U_{S_1}) \\
A_1 &= f_{A_1}(f_{S_1}(U_{S_1}), U_{A_1}) \\
R_1 &= f_{R_1}(f_{S_1}(U_{S_1}), f_{A_1}(f_{S_1}(U_{S_1}), U_{A_1}), U_{R_1}) \\
S_2 &= f_{S_2}(f_{S_1}(U_{S_1}), f_{A_1}(f_{S_1}(U_{S_1}), U_{A_1}), U_{S_2}) \\
A_2 &= f_{A_2}(f_{S_2}(f_{S_1}(U_{S_1}), f_{A_1}(f_{S_1}(U_{S_1}), U_{A_1}), U_{S_2}), U_{A_2}) \\
R_2 &= f_{R_2}(f_{S_2}(f_{S_1}(U_{S_1}), f_{A_1}(f_{S_1}(U_{S_1}), U_{A_1}), U_{S_2}), \\
&\qquad f_{A_2}(f_{S_2}(f_{S_1}(U_{S_1}), f_{A_1}(f_{S_1}(U_{S_1}), U_{A_1}), U_{S_2}), U_{A_2}), U_{R_2}),
\end{aligned}$$

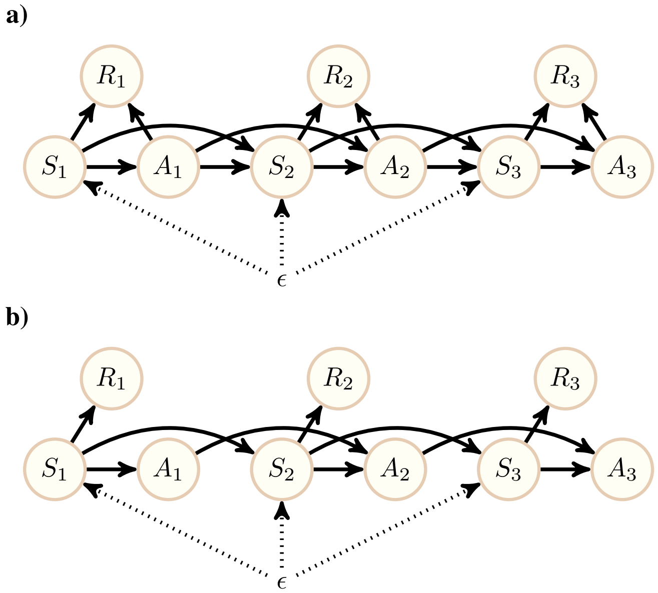

Figure 4: Example DAG for $T = 3$ and non-MDP settings where (a) actions have an effect, and (b) actions have no effect on the states and rewards.

where $U = (U_{S_1}, U_{A_1}, U_{R_1}, U_{S_2}, U_{A_2}, U_{R_2})$ is the set of all independent and unobservable variables. We denote the set of all observable variables as $V = (S_1, A_1, R_1, S_2, A_2, R_2)$. The GCRL-SL optimal policy at $t = 1$ can then be written as

$$\operatorname*{argmax}_a \frac{P_b(f_{R_2}(f_{S_2}(s,a,U_{S_2}), f_{A_2}(f_{S_2}(s,a,U_{S_2}), U_{A_2}), U_{R_2}) = g \mid f_{S_1}(U_{S_1}) = s)}{\sum_{\tilde{a}} P_b(f_{R_2}(f_{S_2}(s,\tilde{a},U_{S_2}), f_{A_2}(f_{S_2}(s,\tilde{a},U_{S_2}), U_{A_2}), U_{R_2}) = g \mid f_{S_1}(U_{S_1}) = s)}. \tag{11}$$

We can identify the numerator in Equation 11 as

$$\begin{aligned}
&P_b(f_{R_2}(f_{S_2}(s,a,U_{S_2}), f_{A_2}(f_{S_2}(s,a,U_{S_2}), U_{A_2}), U_{R_2}) = g \mid f_{S_1}(U_{S_1}) = s) \qquad (12)\\
&= P_b(\{f_{A_1}(f_{S_1}(U_{S_1}), U_{A_1}) = a\} \cap \{f_{R_2}(f_{S_2}(s,a,U_{S_2}), f_{A_2}(f_{S_2}(s,a,U_{S_2}), U_{A_2}), U_{R_2}) = g\} \mid f_{S_1}(U_{S_1}) = s)\\
&= P_b(A_1 = a \mid f_{S_1}(U_{S_1}) = s) P_b(f_{R_2}(f_{S_2}(s,a,U_{S_2}), f_{A_2}(f_{S_2}(s,a,U_{S_2}), U_{A_2}), U_{R_2}) = g \mid f_{S_1}(U_{S_1}) = s)\\
&= P_b(A_1 = a \mid S_1 = s) P_b(R_2 = g \mid A_1 = a, S_1 = s).
\end{aligned}$$

Substituting back into Equation 11, we get that the target optimal policy at $t = 1$ is identified as

$$\operatorname*{argmax}_a \frac{P_b(A_1 = a \mid S_1 = s) P_b(R_2 = g \mid A_1 = a, S_1 = s)}{P_b(R_2 = g \mid S_1 = s)}. \tag{13}$$

The same algebra results in the target optimal policy at $t = 2$ to be identified as

$$\operatorname*{argmax}_a \frac{P_b(A_2 = a \mid S_2 = s) P_b(R_2 = g \mid A_2 = a, S_2 = s)}{P_b(R_2 = g \mid S_2 = s)}. \tag{14}$$

Note that, under the MDP dynamics illustrated in Figure 5(a), we have that

$$P_b(R_2 = g \mid do(A_2 = a), S_2 = s) = P_b(R_2 = g \mid A_2 = a, S_2 = s)$$

and

$$P_b(R_2 = g \mid do(A_1 = a), S_1 = s) = P_b(R_2 = g \mid A_1 = a, S_1 = s)$$

by Rule 2 of *do*-calculus since $(R_2 \perp\!\!\!\perp A_2 \mid S_2)_{\underline{A_2}}$ and $(R_2 \perp\!\!\!\perp A_1 \mid S_1)_{\underline{A_1}}$. Therefore the identified expression in Equation 13 and 14 is the causal effect. We can also apply Corollary 2 in order to check if the effect is identifiable. As there is no selection, C3 and C4 will hold. Let $W^+ = \{S_1\}$ and $W^- = \{S_2\}$ for $t = 1$, and $W^+ = \{S_2\}$ and $W^- = \{\}$ for $t = 2$. Then, C1, C2 are satisfied in Corollary 2.

Consider a subgraph of the MDP in Figure 5(a) where actions have no effect on the future rewards and states, as illustrated in Figure 5(b). By Theorem 1, $P_b(A_t \mid S_t = s)$ is g-recoverable from $P_b(A_t, S_t \mid R_T = g)$, as all actions are independent of the goal given the current state, $R_T \perp\!\!\!\perp A_t \mid S_t$. □

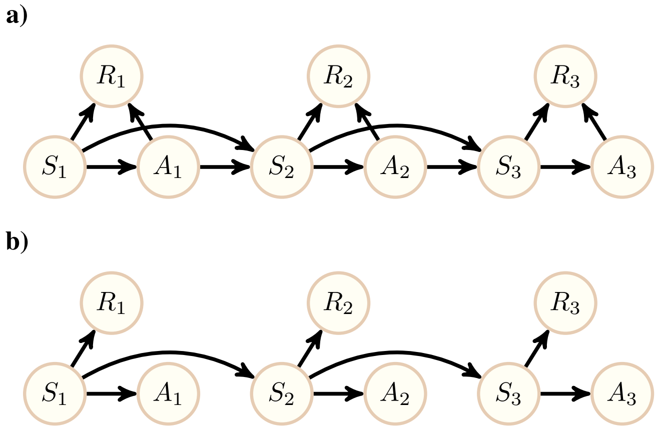

Figure 5: Example DAG for $T = 3$ and MDP settings where (a) actions have an effect, and (b) actions have no effect on the states and rewards.

C Details on Graphs in Section 4

C.1 Complete Graph Environment (CG1)

We consider a finite horizon MDP with $T = 7$. In this setting, all variables are endogenous, as represented by the corresponding DAG in Figure 6. The behavior policy takes as input the previous time-point action and state variables. The action space is binary, with $A = 1$ corresponding to "assign action" and $A = 0$ indicating "don't assign action". The data-generating process (DGP) corresponding to the state variables is as follows,

$$S_1 \sim Normal(0, \sigma)$$
$$S_t \sim Normal(\mu_{a,t}, \sigma), \text{ for } 1 < t \leq 3$$
$$S_t \sim Normal(\mu_{b,t}, \sigma), \text{ for } t \leq T$$

where $\mu_{a,t} = -0.7A_{t-1} + 0.4S_{t-1}$ and $\mu_{b,t} = 0.4A_{t-1} + 0.4S_{t-1}$. The process is not stationary. The reward at time t is equal to 1 if S_t exceeds the third quantile of the asymptotic distribution of S_t. Otherwise, it is 0. The goal state for CG1 is 0.7, which indicates a high value for the final state at $T = 7$. In the DAG shown in Figure 6, both states and actions are colliders. The rewards can be colliders or descendants of colliders. It's crucial to emphasize that actions exert a negative influence on the states until $t = 3$, after which they exhibit a positive effect from $t > 3$. By conditioning on the final state or reward, we may introduce spurious associations between actions at earlier time points (negatively affecting the return) and actions at later time points (positively affecting the return).

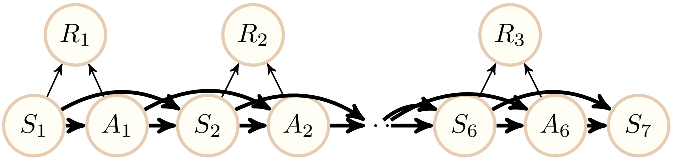

Figure 6: DAG corresponding to the Complete Graph environment (CG1) with a horizon of $T = 7$.

C.2 Incomplete Graph Environment (IG1)

We consider a finite horizon DGP with $T = 7$. In this setting, all variables are endogenous except for an *unknown*, exogenous variable $\epsilon \sim Normal(1, 0.2)$. The corresponding DAG is depicted in Figure 7. The behavior policy takes as input the previous time-point state and action variable. Similar to the previous scenario (CG1), the action space is binary. The data-generating process corresponding to the state variables is as follows,

$$S_1 \sim 0.8\epsilon$$
$$S_t \sim Normal(\mu_t, \sigma), \text{ for } 1 < t \leq T$$

where $\mu_t = -0.9A_{t-1} - 0.9S_{t-1} + 5\epsilon$. The process is stationary. The reward at time t is equal to 1 if S_t exceeds the third quantile of the asymptotic distribution of S_t. Otherwise, it is 0. The goal state for IG1 is 2.4, which indicates a high value for the final state at $T = 7$. In the DAG depicted in Figure 7, both states and actions are colliders. The rewards can be colliders or descendants of colliders. It's important to note that ϵ positively influences the outcome, while actions have a negative impact on the trajectory's return. By conditioning on the final state or reward, we can introduce spurious association between actions and ϵ.

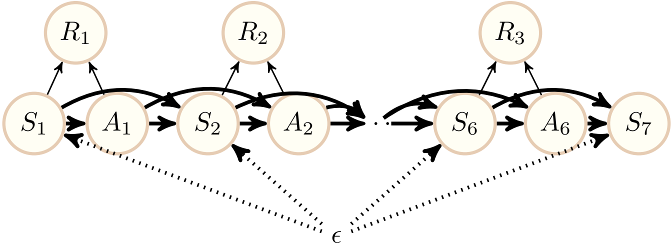

Figure 7: DAG corresponding to the Incomplete Graph environment (IG1) with $T = 7$. Dotted arrows represent paths from unknown, exogenous (unobserved) variable to endogenous (observed) state variables.

D Experiment Details

We run multiple experiments to show that performance differences between GCRL-SL and value-based RL stem from GCRL-SL methods learning poor policies due to causal biases introduced during the training phase. In all experiments, we focused on an offline RL setup, employing the most general form of GCRL-SL with final states as goals, as described in [13]. For comparison, we used a canonical version of Fitted Q Iteration (FQI). Each experiment was independently run 100 times, corresponding to 100 Monte Carlo simulations or iterations. The reported return represents an average over these 100 random iterations, with each evaluated over 20 validation trajectories.

In Table 1, we present details regarding the neural network architecture utilized in all experiments, unless explicitly stated otherwise. By default, we employ a feedforward fully-connected neural network for estimation. Notably, in the last row of Table 1, we highlight our exploration of regularization through dropout as a distinct estimator for various conditional policy estimators. The table also outlines all the algorithms incorporated in the ensemble learner's library, referred to as the Super Learner (SL) [32]. This library includes the following algorithms: (1) generalized linear model (glm), (2) single-layer neural network, (3) generalized additive model, (4) random forest, and (5) regularized gradient boosting [8, 35, 34]. We explored different configurations of random forests and gradient boosting based on their hyperparameters, such as the number of trees, maximum depth, and eta. The Super Learner utilized 10-fold cross-validation.

In Table 2, we present a detailed list of simulation parameters. In each iteration, we trained both GCRL-SL and FQI algorithm using training sets of varying sizes, specifically $n = \{50, 100, 500, 1000, 3000, 5000\}$. For every experiment, we utilized a validation set comprising 20 trajectories, and the reported return represents the average over 100 Monte Carlo simulations. To attain the target goal values for GCRL-SL, we set the target goal at 0.7 for CG1 and 2.4 for IG1. These goals were determined based on the asymptotic distribution consistent with the dynamics of the CG1 and IG1 DGP at the end of each trajectory. Specifically, they correspond to the upper tails (3rd quantile) of the asymptotic distribution observed in CG1 and IG1 DGPs and are supported by the training data used in each experiment.

E Experiment Results

E.1 Is stochasticity driving performance?

Recent studies suggest that GCRL-SL algorithms face challenges in stochastic environments [13, 6]. In our analysis, we assess the performance of GCRL-SL and FQI across different levels of variability,

Table 1: Algorithm specifications and design parameters used for considered experiments.

Hyperparameter	Value	Environment
Hidden Layers	2	All
Layer Width	1024	All
Nonlinearity	ReLU	All
Learning rate	1e-3	All
Epochs	20	All
Dropout	0	All
	0.1	All
Ensemble Learner	glm	All
	gam	All
	neural network	All
	random forest	All
	xgboost	All
CV	10	All

Table 2: Simulation parameters used for considered experiments.

Hyperparameter	Value	Environment
Number of MC iterations	100	All
Training size	50	All
	100	All
	500	All
	1000	All
	3000	All
	5000	All
Validation size	20	All
Goal max	0.7	CG1
Goal max	2.4	IG1

denoted by the parameter σ, in the data generating processes of CG1 and IG1. The results, depicted in Figure 8, show that FQI consistently outperforms GCRL-SL across all levels of stochasticity and training dataset sizes.

E.2 Do we need different policy estimators?

Practical recommendations suggest that simple implementations can achieve competitive, if not superior, performance compared to more complex architectures and value-based RL methods [13]. Others emphasize the importance of complex neural network architectures, as even if the behavior policy is simple, the conditional policy learned by GCRL-SL might not be [19, 7]. In our analysis, we explore various choices for policy estimation, including: (1) a simple main terms generalized linear model (glm), (2) Super Learner (SL), an ensemble learner based on cross-validation, (3) a high-capacity feed-forward fully-connected neural network, and (4) a high-capacity neural network with regularization. The SL is a convex combination of predictions made by glm, generalized additive model, shallow neural network, regularized gradient boosting, and random forest [32, 8, 35, 34]. Figure 9 presents the results for different policy estimators at $\sigma = 0.1$. It demonstrates that FQI consistently outperforms GCRL-SL across all considered policy estimators and sample sizes.

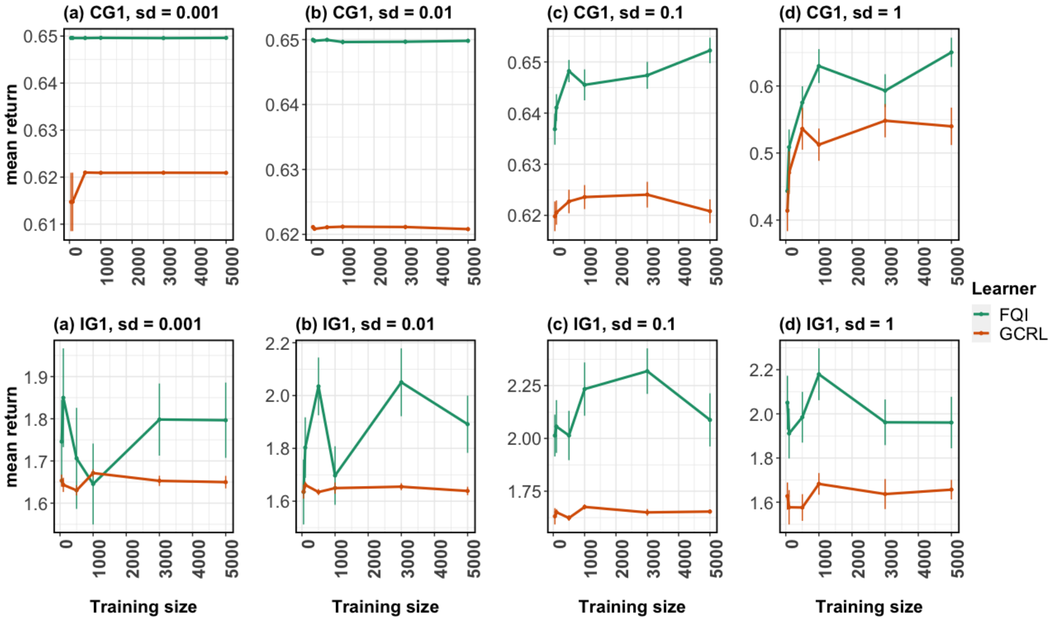

Figure 8: Mean return for CG1 and IG1 Data Generating Process (DGP) at $t = 7$ and its corresponding standard error, calculated over 100 Monte Carlo (MC) iterations. In the upper panels (a)-(d), we illustrate the CG1 DGP, while in the lower panels (a)-(d), we depict the IG1 DGP under different levels of $\sigma = \{0.001, 0.01, 0.1, 1\}$, indicating increasing stochasticity in the process.

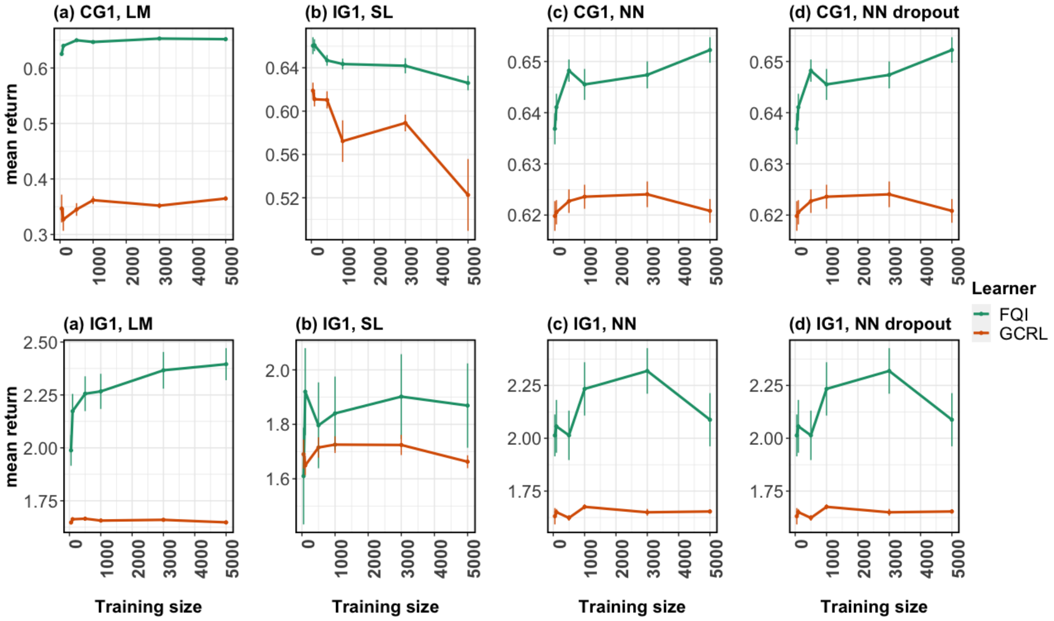

Figure 9: Mean return for CG1 and IG1 Data Generating Process (DGP) at $\sigma = 0.1$ and $t = 7$ with its corresponding standard error, calculated over 100 Monte Carlo (MC) iterations. In the upper panels (a)-(d), we illustrate the CG1 DGP, while in the lower panels (a)-(d), we depict the IG1 DGP with policy estimated using different estimators: linear models (LM), ensemble learner (SL), Neural Network (NN) and Neural Network with dropout.