Policy Improvement Over an Alternate Observation Space

Gokul Swamy [1] **Sanjiban Choudhury** [2] **J. Andrew Bagnell** [3 1] **Zhiwei Steven Wu** [1]

Abstract

We consider the problem of improving upon a black-box policy which operates on a different observation space than the learner. Such problems occur when augmenting an existing hand-engineered system with a new machine learning model or in a shared autonomy / human-AI complementarity context. We prove that following the naive policy gradient can lead to a decrease in performance because of incorrect grounding in a different observation space. We then consider three settings: a *white-box* setting in which we, at train time, have access to both sets of observations at, a *grey-box* setting in which we see learner observations and behavior policy action probabilities, and a *black-box* setting in which we only see learner observations and behavior policy actions. We derive methods for correctly estimating a policy gradient in each of these settings via an application of the backdoor adjustment, importance sampling, and the proxy correction, respectively.

1. Introduction

Classical results tell us that in a Partially Observed Markov Decision Process (POMDP), finding high-value policies that map observations to actions is computationally hard (Papadimitriou & Tsitsiklis, 1987). Furthermore, a naive application of MDP techniques (e.g. value-iteration, policy-iteration, or Q-learning) empirically leads to unstable and often decaying performance over iterations (Singh et al., 1994; Littman, 1994). However, folk wisdom suggests that if one computes (or approximates well) $Q^\pi(o, a)$ – the long-term expected return of an observation-action pair – and then makes "small adjustments" to the policy to select higher value actions for a given observation, the policy can be reliably improved. Such small changes can then be repeated to achieve a notion of local optima. This folk wisdom finds precision for stochastic mixtures of policies (Jaakkola et al., 1994; Kakade & Langford, 2002; Daumé et al., 2009), parameterized stochastic policies (Baxter & Bartlett, 2001; Williams, 1992), and when searching over a set of deterministic policies (Bagnell et al., 2003).

An oft-underemphasized requirement for the above policy improvement theorems to hold is that the proposed change to the policy uses *the same observation space* as the policy it is improving over. In practice, this often doesn't hold, as ML practitioners may be attempting to improve a policy that comes from an existing, effectively black-box system to them, or even to provide action recommendations to expert humans who operate with a fundamentally different observation space.

For example, consider a medical assistance system to help doctors at a walk-in clinic. In response to observing inflamed tonsils, a doctor might prescribe an antibiotic. The automated system, which operates purely based on numerical information, does not have access to this feature but does observe that patients always get better after being prescribed amoxicillin. It would then over-estimate $Q^{\text{doctor}}(o, \text{amoxicillin})$, leading to it suggesting unnecessary prescriptions of a strong medication that could have unfortunate side-effects for patients with issues other than strep throat. More generally, naively computing $Q^\pi(o, a)$ where π uses a different observation space than o can lead to estimates of cumulative reward that will not match what the learner would actually receive at test time.

In this work, we study the perhaps initially surprising failure of these policy improvement algorithms. We identify their fundamental issue: ***when observations spaces differ, $Q^\pi(o, a)$ becomes causally confounded, rendering the standard Monte-Carlo estimate inconsistent.*** We then study three regimes – first where a richer observation space is available when we train a policy, but not when we execute that policy ; second, a more difficult setting where at both train and test we have access only to a history of an observation space more limited than the initial policy we wish to improve over. We also consider a setting in between these two in which we are given an *action probability* by the behavior policy, rather than just a sample from the policy. Our technical tools leverage recent advances in proxy variable correction techniques (Miao et al., 2018; Tchetgen et al.,

[*]Equal contribution [1]Carnegie Mellon University [2]Cornell University [3]Aurora Innovation. Correspondence to: Gokul Swamy <gswamy@cmu.edu>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA. Copyright 2023 by the author(s).

2020) and approximate back-door corrections (Pearl, 2009; Xu & Gretton, 2022) to provide correct policy gradients in confounded settings.

More explicitly, our work makes the following four contributions.

1. We formalize the problem of improving upon a policy with a different observation space. We prove that correct estimation of $Q^{\pi}(o, a)$ values is sufficient for policy improvement. We then give a family of examples under which the standard, Monte-Carlo estimate for $Q^{\pi}(o, a)$ will not lead to policy improvement.

2. We consider the "white-box" □ setting when both sets of observations are available at training time. We show how we can use Pearl's backdoor adjustment (Pearl, 2009) to de-confound our Q-value estimates. We derive a procedure efficiently approximating the backdoor adjustment using the output of a simple classifier.

3. We consider the "black-box" ■ setting in which the learner never observes the features the baseline policy used to make decisions. Under some assumptions, we show how we can use the proxy correction (Miao et al., 2018; Tchetgen et al., 2020; Bennett & Kallus, 2021) to correct estimate Q-values in the learner's observation space. We derive two algorithms for doing so, one of a generative-modeling flavor and one of a game-theoretic flavor.

4. We consider the "grey-box" ▨ setting in which we don't observe the features the baseline policy used to make decisions but we are given the probability of the observed decision. We derive a simple importance-sampling based procedure to correct the naive Q-value estimates.

We now turn our attention to formalizing the problem we consider.

2. Formalism

Consider a POMDP (Kaelbling et al., 1998) with two different observation spaces, $\mathcal{O}_1$ and $\mathcal{O}_2$. We can represent our setup via the following structural causal model (SCM).

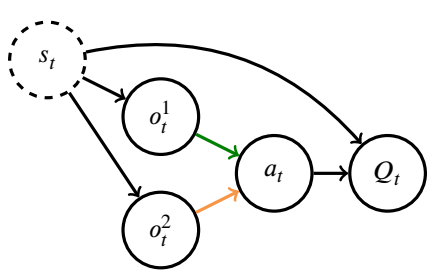

Figure 1. The causal model we consider in this paper. The green arrow denotes the dependence in the training data, the orange arrow denotes the dependence of the policy we're learning.

Let

$$J(\pi_\theta) = \mathbb{E}_{\xi\sim\pi_\theta}\left[\sum_t^T r(s_t, a_t)\right] \tag{1}$$

denote the performance of a policy. We assume we see data generated by some behavior policy $\pi_b : \mathcal{O}^1 \to \Delta(\mathcal{A})$ that consists of tuples of (o_t^2, a_t, Q_t) and would like to find a direction in policy-space that improves $J(\pi_b)$. More formally, we're looking for a policy $\pi : \mathcal{O}^2 \to \Delta(\mathcal{A})$ such that

$$J((1-\epsilon)\pi_b + (\epsilon)\pi) \geq J(\pi_b) \tag{2}$$

for small enough ϵ.

2.1. Policy Improvement in POMDPs

The first complexity we have to deal with is that our policies depend on observations rather than states. Following an argument similar to that of Jaakkola et al. (1994), we prove the following theorem.

Theorem 2.1. *Define $\pi_\epsilon = (1-\epsilon)\pi_b + (\epsilon)\pi$, where π_b operates on $\mathcal{O}^2$ and π operates on $\mathcal{O}^1$. Then, we have that*

$$J(\pi_\epsilon) - J(\pi_b) = \epsilon T \mathbb{E}_{o\sim\rho_{\pi_\epsilon}}[\mathbb{E}_{a\sim\pi(a|o)}[\mathbb{E}_{s\sim\rho_{\pi_b}(s|o)}[A^{\pi_b}(s,a)]]] + O(\epsilon^2 T^2). \tag{3}$$

Define

$$A^{\pi_b}(o, a) = \mathbb{E}_{s\sim\rho_{\pi_b}(s|o)}[A^{\pi_b}(s, a) - V^{\pi_b}(s)], \tag{4}$$

and $Q^{\pi_b}(o, a)$, $V^{\pi_b}(o)$ analogously. The above result tells us that as long as we are able to estimate $Q^{\pi_b}(o, a)$ values correctly, putting more probability mass on actions with positive advantages with a step of size $\epsilon < O(\frac{1}{T})$ is sufficient to guarantee improvement.

For comparison, consider the (vanilla) policy gradient for MDPs.

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\xi\sim\pi_\theta}\left[\sum_t^T \nabla_\theta \log(\pi_\theta(a|s)) Q^{\pi_\theta}(s, a)\right]. \tag{5}$$

One usually subtracts a baseline to reduce variance. [1] The optimal control variate is the value function (Sutton & Barto, 2018), which gives us the following expression

$$\nabla_\theta J(\pi_\theta) = \mathbb{E}_{\xi\sim\pi_\theta}\left[\sum_t^T \nabla_\theta \log(\pi_\theta(a|s)) A^{\pi_\theta}(s, a)\right]. \tag{6}$$

[1] One usually pre-conditions with the Fisher information matrix to follow the natural policy gradient (Kakade, 2001), but we ignore those details for simplicity.

Similar to our above result, the theory of policy gradients (Kakade & Langford, 2002; Agarwal et al., 2020) tells us that mixing in a small enough portion of the greedily optimal policy is sufficient to guarantee improvement. Thus, the main difference comes in the definition of the advantage function: for MDPs, advantages depend on states while for our setup, advantages depend on the observation space of the policy we're mixing in. We therefore turn our attention to estimating Q-values / advantages.

2.2. An Example of why the Naive Q-value Estimate Fails

We illustrate the issue with the naive Q-value estimator with a simple, one-step example.

Problem 2.2 (Cancer Treatment)**.** Consider the following one-step (contextual bandit) problem. A doctor observes whether a patient has cancer ($\mathcal{O}^1 = \{\text{cancer}, !\text{cancer}\}$) and decides whether to instruct them to begin chemotherapy. The rewards are as follows:

$$Q(\text{cancer}, \text{chemo}) = 1, Q(!\text{cancer}, \text{chemo}) = 0, \quad (7)$$
$$Q(\text{cancer}, !\text{chemo}) = 0, Q(!\text{cancer}, !\text{chemo}) = 0.5$$

We assume α fraction of the population has cancer. The doctor makes a mistake (i.e. picking the action with the lower Q value) with probability ϵ. We want to learn a policy that, without any observations, improves upon the doctor ($\mathcal{O}^2 = \emptyset$).

To solve this problem, we compute both $Q(\emptyset, \text{chemo})$ and $Q(\emptyset, !\text{chemo})$ by averaging the rewards received when a patient was / wasn't given chemotherapy, regardless of their underlying cancer status. More formally,

$$\hat{Q}(\emptyset, a) = \mathbb{E}_{\pi_b}[Q(S, a)] \quad (8)$$

We pick the action with the higher Q value always (i.e. $\pi(\emptyset) = \max_a Q(\emptyset, a)$). We visualize the region of (α, ϵ) in which we pick !chemo in yellow in the below figure. Given access to the base rate of cancer (α), we can also compute the ground-truth Q values, which tell us when giving no patients chemotherapy would be preferable to giving all of them chemotherapy. Intuitively, this is true when few patients have cancer. We visualize this in teal in the figure.

Mixing in *any* amount of the greedy optimal policy is harmful when $\hat{Q}$ and Q rank the two actions in the opposite order, which correspond to the red dots in the above figure. Intuitively, the cluster of red dots in the bottom left of the figure corresponds to the learner under-estimating the value of chemotherapy when few people need it and doctors rarely spuriously prescribe it.

If such a problem can happen even in a single-step problem, it can clearly happen with multiple steps of interaction. We therefore turn our attention to using techniques from causal inference to de-bias our Q-value estimates.

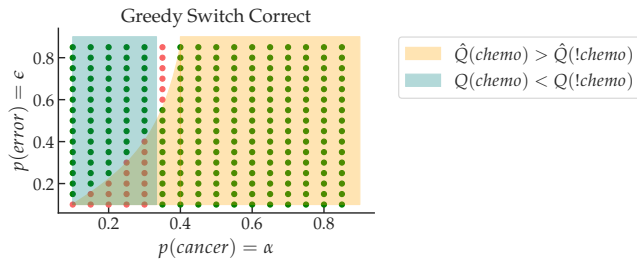

Figure 2. We see that when the base rate of cancer is low and the doctor makes few mistakes, we under-estimate the value of chemo and would end up providing poor care to people. This is indicated in the figure by the red dots.

2.3. White Box □: Estimating Q-values when both Observations are Available

We first consider the relatively simple setting when both sets of observations are provided to the learner in the behavioral data. Formally, we want to compute

$$Q(o_t^2, a) = \mathbb{E}[\sum_{\tau=t}^{T} r(s_\tau, a_\tau) | do(a_t), o_t^2]. \quad (9)$$

Using Q_t as a shorthand for the sum of rewards over the horizon, we can apply the rules of do-calculus (equivalently, the backdoor adjustment formula (Pearl, 2009)) to compute the above causal effect.

$$\begin{aligned}
\mathbb{E}[Q_t | do(a_t), o_t^2] &= \sum_{o_t^1} P(o_t^1 | do(a_t), o_t^2) \mathbb{E}[Q_t | do(a_t), o_t^1, o_t^2] \\
&= \sum_{o_t^1} P(o_t^1 | do(a_t), o_t^2) \mathbb{E}[Q_t | a_t, o_t^1, o_t^2] \\
&= \sum_{o_t^1} P(o_t^1 | o_t^2) \mathbb{E}[Q_t | a_t, o_t^1, o_t^2],
\end{aligned}$$

where the second equality follows from $(Q_t \perp\!\!\!\perp a | o_t^1, o_t^2)_{\mathcal{G}_{\underline{a_t}}}$ and Rule 2 and the third equality follows from $(o_t^1 \perp\!\!\!\perp a | o_t^2)_{\mathcal{G}_{\overline{a_t}}}$ and Rule 3. Lastly, simple algebra tells us that

$$\begin{aligned}
\mathbb{E}[Q_t | do(a_t), o_t^2] &= \mathbb{E}_{o_t^1}\left[\frac{P(o_t^1, o_t^2)}{P(o_t^1)P(o_t^2)} \mathbb{E}[Q_t | a_t, o_t^1, o_t^2]\right] \\
&= \mathbb{E}_{o_t^1}\left[r(o_t^1, o_t^2) \mathbb{E}[Q_t | a_t, o_t^1, o_t^2]\right], \quad (10)
\end{aligned}$$

where we use $r(o_t^1, o_t^2)$ to denote the density ratio between the joint and the product of marginals.

The above expression gives us the following procedure for policy improvement.

1. Estimate $\mathbb{E}[Q_t|a_t, o_t^1, o_t^2]$ via regression.
2. Fit a density ratio estimate for $r(o_t^1, o_t^2)$.
3. Fit $Q(o_t^2, a_t)$ via a secondary regression with targets determined via the previous two components.
4. Compute the greedily optimal policy and plug into your favorite policy improvement method to control the step size.

We can perform the seconds step via classification between paired and unpaired samples of o_t^1, o_t^2 (Hyvarinen & Morioka, 2016).

2.4. Grey Box ▨: Estimating Q-values with Learner Observations and Behavior Policy Action Probabilities

Even if we don't have access to both sets of observations, if we are given access to the full distribution of possible action probabilities from the behavior policy (e.g. $p(a_t|o_t^1)$), we can use an importance sampling technique to correct the naive Q-value estimate. Let us build up intuition for why this is the case.

Pretend for a second that we were able to intervene and set a_t based on the outcome of a random coin flip (e.g. a randomized control trial). Then, we'd clearly be computing the desired conditional average treatment effect (CATE). We now argue that access to behavior policy action probabilities is sufficient to simulate such an unbiased coin flip. Recall that

$$\mathbb{E}[Q_t|do(a_t), o_t^2] = \sum_{o_t^1} P(o_t^1|o_t^2)\mathbb{E}[Q_t|a_t, o_t^1, o_t^2] \tag{11}$$

and

$$\mathbb{E}[Q_t|a_t, o_t^2] = \sum_{o_t^1} P(o_t^1, a_t|o_t^2)\mathbb{E}[Q_t|a_t, o_t^1, o_t^2]. \tag{12}$$

Importantly, this second quantity can be calculated simply by averaging the observed cumulative returns after an action-observation pair. The critical step is to realize that importance weights

$$\frac{P(o_t^1|o_t^2)}{P(o_t^1, a_t|o_t^2)} = \frac{P(o_t^1, o_t^2)}{P(o_t^1, a_t, o_t^2)} = \frac{1}{P(a_t|o_t^1, o_t^2)} = \frac{1}{P(a_t|o_t^1)},$$

where the last equality comes from the fact that $a_t \perp\!\!\!\perp o_t^2|o_t^1$. Thus, simply by re-weighting the observed returns by the inverse of the observed action probabilities, we are able to estimate the causal effect, *all without ever seeing the behavior policy's observation*. Putting it all together, our proposed estimator for the grey-box setting is

$$\mathbb{E}[Q_t|do(a_t), o_t^2] = \mathbb{E}\left[\frac{Q_t}{P(a_t|o_t^1)}\Big|o_2^t, a_t\right]. \tag{13}$$

We would then plug in these Q values estimates into a policy improvement method like CPI (Kakade & Langford, 2002) to ensure we don't take too large of a step towards the greedily optimal policy.

2.5. Black Box ■: Estimating Q-values with only Learner Observations

We first describe the general technique we use before specializing to our case of POMDPs. The argument we present below is a slight generalization of that in Xu et al. (2021).

2.5.1. The Proxy Correction

We are attempting to determine causal effect $\mathbb{E}[Y|do(a), x]$. Unfortunately, Y, A, X all have a common parent: unobserved confounder U. However, we have access to two "proxies", W, Z, for the unobserved U that we can use to estimate the effect. Formally, we require the following independence conditions (Miao et al., 2018):

$$\text{C1: } W \perp\!\!\!\perp (A, Z)|U, X \tag{14}$$

$$\text{C2: } Y \perp\!\!\!\perp Z|A, U, X \tag{15}$$

An example of a graph that satisfies these conditions is drawn in Figure 3, left. Consider the following set of conditional moment restrictions (CMR) on g:

$$\mathbb{E}[Y|a, z, x] = \mathbb{E}[g(a, x, W)|a, z, x], \forall(a, z, x) \in (\mathcal{A}\times\mathcal{Z}\times\mathcal{X}). \tag{16}$$

We prove the following theorem.

Theorem 2.3. *Define*

$$h(a, x) = \mathbb{E}_{W|x}[g(a, x, W)]. \tag{17}$$

If g satisfies Equation 16, then $\mathbb{E}[Y|do(a), x] = h(a, x)$.

Swamy et al. (2022) provide two methods for learning a g that satisfies the CMR in one of two ways. The first is via game-solving:

$$\min_g \max_f \mathbb{E}[f(A, X, Z)(Y - g(A, X, W)) - 0.5 f^2(A, X, Z)].$$

The second is via generative modeling. First, we learn a model of $P(W|a, z, x)$. Then, we minimize the following MSE

$$\min_g \mathbb{E}_{A,Z,X}[(\mathbb{E}[Y|a, z, x] - \mathbb{E}[g(a, x, W)|a, z, x])^2]. \tag{18}$$

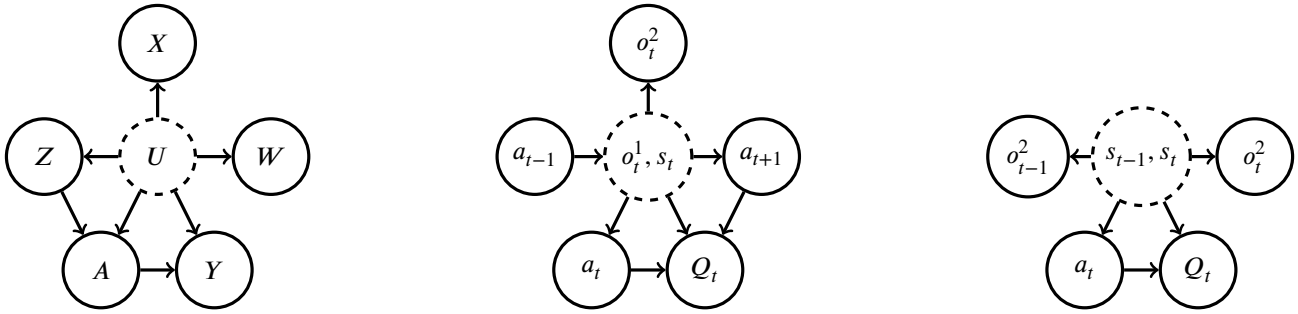

Figure 3. Causal models that satisfy the independence assumptions required to apply the proxy correction.

Importantly, by drawing separate samples of W for both evaluations of g inside the square, we avoid the so-called "double-sample" problem that can lead to inconsistent estimates of a causal effect (Baird, 1995).

A third family of techniques, described by Xu et al. (2021), alternates between solving two regression problems, each of which corresponds to one of the stages of the above generative modeling approach. More formally, we assume that we can write g and the conditional feature mean as

$$g(a, x, w) = u^T(\Psi_{A(2)}(a) \otimes \Psi_{X(2)}(x) \otimes \Psi_W(w))$$

$$\mathbb{E}_{W|a,z,x}[\Psi_W(w)] = V(\Psi_{A(1)}(a) \otimes \Psi_Z(z) \otimes \Psi_{X(1)}(x)),$$

where all Ψ are learned neural networks and u and V are a vector and matrix with closed-form expressions, respectively. Observe that because we take a tensor product $\otimes$ between the output of neural network features which can be arbitrarily high dimensional, the above separation into functions of each random variable is without meaningful restriction on functions representable. We then alternate between minimizing two loss functions over subsets of the parameters. We refer interested readers to Xu et al. (2021)'s excellent paper for more details.

2.5.2. Applying the Proxy Correction to POMDPs

Due to space concerns, we only write out the full form of the game-solving variant.

Approach 1. Consider the SCM in Figure 3, center.

Observe that we satisfy the requisite independence assumptions:

$$\text{C1: } a_{t-1} \perp\!\!\!\perp (a_t, a_{t+1}) | s_t, o_t^2, o_t^1 \tag{19}$$

$$\text{C2: } Q_t \perp\!\!\!\perp a_{t-1} | a_t, s_t, o_t^2, o_t^1 \tag{20}$$

Putting it all together, we suggest the following method for policy improvement

1. Solve game

$$\min_g \max_f \mathbb{E}[f(a_t, o_t^2, a_{t-1})(Q_t - g(a_t, o_t^2, a_{t+1})) - 0.5 f^2(a_t, o_t^2, a_{t-1})]$$

2. Fit generative model of $P(a_{t+1}|o_t)$.
3. Fit a model of $Q^{\pi_b}(o_t^2, a_t) = \mathbb{E}_{a_{t+1}|o_t}[g(a_t, o_t, a_{t+1})]$ via regression.
4. Compute the greedily optimal policy and plug into a policy improvement method to control the step size.

Note that the generative model we have to fit in the second step is just a policy.

Approach 2. An alternative approach which would allow us to estimate the same quantity would be to use the past and current observations as proxies (Bennett & Kallus, 2021). We visualize this approach graphically in Figure 3, right.

Note that because we directly use the observation as one of the proxies, we don't need to perform a secondary averaging. We can solve the following game to recover Q_t:

$$\min_g \max_f \mathbb{E}[f(a_t, o_{t-1})(Q_t - g(a_t, o_t)) - 0.5 f^2(a_t, o_{t-1})]$$

After solving for g, we can directly plug it into an advantage estimation procedure (Schulman et al., 2018), allowing us to compute causally correct policy gradients.

3. Acknowledgments

We thank Arthur Gretton for insightful comments on the proxy correction. ZSW is supported in part by the NSF FAI Award #1939606, a Google Faculty Research Award, a J.P. Morgan Faculty Award, a Facebook Research Award, an Okawa Foundation Research Grant, and a Mozilla Research Grant. GS is supported by his family and friends.

References

Agarwal, A., Jiang, N., Kakade, S. M., and Sun, W. Reinforcement learning: Theory and algorithms. *CS Dept., UW Seattle, Seattle, WA, USA, Tech. Rep*, pp. 10–4, 2019.

Agarwal, A., Kakade, S. M., Lee, J. D., and Mahajan, G. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*, pp. 64–66. PMLR, 2020.

Bagnell, J., Kakade, S. M., Schneider, J., and Ng, A. Policy search by dynamic programming. *Advances in neural information processing systems*, 16, 2003.

Baird, L. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pp. 30–37. Elsevier, 1995.

Baxter, J. and Bartlett, P. L. Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350, 2001.

Bennett, A. and Kallus, N. Proximal reinforcement learning: Efficient off-policy evaluation in partially observed markov decision processes. *arXiv preprint arXiv:2110.15332*, 2021.

Daumé, H., Langford, J., and Marcu, D. Search-based structured prediction. *Machine learning*, 75:297–325, 2009.

Hyvarinen, A. and Morioka, H. Unsupervised feature extraction by time-contrastive learning and nonlinear ica. *Advances in neural information processing systems*, 29, 2016.

Jaakkola, T., Singh, S., and Jordan, M. Reinforcement learning algorithm for partially observable markov decision problems. *Advances in neural information processing systems*, 7, 1994.

Kaelbling, L. P., Littman, M. L., and Cassandra, A. R. Planning and acting in partially observable stochastic domains. *Artificial intelligence*, 101(1-2):99–134, 1998.

Kakade, S. and Langford, J. Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pp. 267–274, 2002.

Kakade, S. M. A natural policy gradient. *Advances in neural information processing systems*, 14, 2001.

Littman, M. L. Memoryless policies: Theoretical limitations and practical results. In *From Animals to Animats 3: Proceedings of the third international conference on simulation of adaptive behavior*, volume 3, pp. 238. MIT Press Cambridge, MA, USA, 1994.

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

Papadimitriou, C. H. and Tsitsiklis, J. N. The complexity of markov decision processes. *Mathematics of operations research*, 12(3):441–450, 1987.

Pearl, J. Causal inference in statistics: An overview. 2009.

Schulman, J., Moritz, P., Levine, S., Jordan, M., and Abbeel, P. High-dimensional continuous control using generalized advantage estimation, 2018.

Singh, S. P., Jaakkola, T., and Jordan, M. I. Learning without state-estimation in partially observable markovian decision processes. In *Machine Learning Proceedings 1994*, pp. 284–292. Elsevier, 1994.

Sutton, R. S. and Barto, A. G. *Reinforcement learning: An introduction*. MIT press, 2018.

Swamy, G., Choudhury, S., Bagnell, J. A., and Wu, Z. S. Game-theoretic algorithms for conditional moment matching, 2022.

Tchetgen, E. J. T., Ying, A., Cui, Y., Shi, X., and Miao, W. An introduction to proximal causal learning. *arXiv preprint arXiv:2009.10982*, 2020.

Williams, R. J. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Reinforcement learning*, pp. 5–32, 1992.

Xu, L. and Gretton, A. A neural mean embedding approach for back-door and front-door adjustment, 2022.

Xu, L., Kanagawa, H., and Gretton, A. Deep proxy causal learning and its application to confounded bandit policy evaluation. *Advances in Neural Information Processing Systems*, 34:26264–26275, 2021.

A. Proofs

A.1. Proof of Theorem 2.1

Proof. Via the performance difference lemma (Kakade & Langford, 2002),

$$J(\pi_\epsilon) - J(\pi_b) = T\mathbb{E}_{a,o\sim\rho_{\pi_\epsilon}}[\mathbb{E}_{s\sim\rho_\epsilon(s|o)}[A^{\pi_b}(s,a)]]. \tag{21}$$

Fix $o \in \mathcal{O}^2$. Consider

$$\begin{aligned}
\mathbb{E}_{a\sim\rho_{\pi_\epsilon}(o)}[\mathbb{E}_{s\sim\rho_\epsilon(s|o)}[A^{\pi_b}(s,a)]] &= \sum_{a\in\mathcal{A}}(\pi_b(a|o)+\Delta(a|o))\sum_{s\in\mathcal{S}}(\rho_{\pi_b}(s|o)+\Delta(s|o))(Q^{\pi_b}(s,a)-V^{\pi_b}(s)) \\
&= \sum_{a\in\mathcal{A}}\Delta(a|o)\sum_{s\in\mathcal{S}}(\rho_{\pi_b}(s|o)+\Delta(s|o))(Q^{\pi_b}(s,a)-V^{\pi_b}(s)) \\
&= \epsilon\sum_{a\in\mathcal{A}}\pi(a|o)\sum_{s\in\mathcal{S}}(\rho_{\pi_b}(s|o)+\Delta(s|o))(Q^{\pi_b}(s,a)-V^{\pi_b}(s)) \\
&= \epsilon\sum_{a\in\mathcal{A}}\pi(a|o)\sum_{s\in\mathcal{S}}(\rho_{\pi_b}(s|o))(Q^{\pi_b}(s,a)-V^{\pi_b}(s)) + O(\epsilon^2T^2).
\end{aligned}$$

The last line comes from the fact that

$$d_{TV}(\rho_{\pi_b}(s|o), \rho_{\pi_\epsilon}(s|o)) \le \epsilon T, \tag{22}$$

as proved in Agarwal et al. (2019).

Taking the average over observations tells us that

$$J(\pi_\epsilon) - J(\pi_b) = \epsilon T\mathbb{E}_{o\sim\rho_{\pi_\epsilon}}[\mathbb{E}_{a\sim\pi(a|o)}[\mathbb{E}_{s\sim\rho_{\pi_b}(s|o)}[A^{\pi_b}(s,a)]]] + O(\epsilon^2T^2). \tag{23}$$

□

A.2. Proof of Theorem 2.3

Proof. Via backdoor adjustment (Pearl, 2009),

$$\mathbb{E}[g(a,x,W)|a,z,x] = \mathbb{E}_{U|a,z,x}[\mathbb{E}[g(a,x,W)|a,z,x,u]] = \mathbb{E}_{U|a,z,x}[\mathbb{E}[g(a,x,W)|x,u]], \tag{24}$$

where the second equality comes from assumption C1. Next, we note that

$$\mathbb{E}[Y|a,z,x] = \mathbb{E}_{U|a,z,x}[\mathbb{E}[Y|a,z,x,u]] = \mathbb{E}_{U|a,z,x}[\mathbb{E}[Y|a,x,u]], \tag{25}$$

where the second equality comes from assumption C2. Now, if we satisfy the CMR uniformly, we know that

$$\mathbb{E}_{U|a,z,x}[\mathbb{E}[g(a,x,W)|x,u] - \mathbb{E}[Y|a,x,u]] = 0. \tag{26}$$

If we assume "completeness" of our confounder (Miao et al., 2018; Tchetgen et al., 2020), this implies that the above condition implies the following pointwise guarantee:

$$\mathbb{E}[g(a,x,W)|x,u] = \mathbb{E}[Y|a,x,u]. \tag{27}$$

This guarantee allows us to plug the LHS into the backdoor adjustment formula for the causal effect. More formally,

$$\mathbb{E}[Y|do(a),x] = \mathbb{E}_{U|x}[\mathbb{E}[Y|a,u,x]] = \mathbb{E}_{U|x}[\mathbb{E}[g(a,x,W)|x,u]] = \mathbb{E}_{W|x}[\mathbb{E}[g(a,x,W)], \tag{28}$$

where the first equality comes from the backdoor adjustment formula and the last from the law of iterated expectation. □