Invariant Policy Learning: A Causal Perspective

Sorawit Saengkyongam [1] **Nikolaj Thams** [1] **Jonas Peters** [1] **Niklas Pfister** [1]

Abstract

In the past decade, contextual bandit and reinforcement learning algorithms have been successfully used in various interactive learning systems such as online advertising, recommender systems, and dynamic pricing. However, they have yet to be widely adopted in high-stakes application domains, such as healthcare. One reason may be that existing approaches assume that the underlying mechanisms are static in the sense that they do not change over different environments. In many real world systems, however, the mechanisms are subject to shifts across environments which may invalidate the static environment assumption. In this paper, we tackle the problem of environmental shifts under the framework of offline contextual bandits. We view the environmental shift problem through the lens of causality and propose multi-environment contextual bandits that allow for changes in the underlying mechanisms. We adopt the concept of invariance from the causality literature and introduce the notion of policy invariance. We argue that policy invariance is only relevant if unobserved confounders are present and show that, in that case, an optimal invariant policy is guaranteed to generalize across environments under suitable assumptions. Our results do not only provide a solution to the environmental shift problem but also establish concrete connections among causality, invariance and contextual bandits.

1. Introduction

The problem of learning decision-making policies lies at the heart of learning systems. To adopt these learning systems in high-stakes application domains such as personalized medicine or autonomous driving, it is crucial that the learnt policies are reliable even in environments that have never been encountered before. In this paper, we consider the problem of learning policies that are robust with respect to shifts in the environments. We consider this question in the setup of offline contextual bandits.

While recent studies in offline contextual bandits (Dudik et al., 2011; Bottou et al., 2013; Swaminathan & Joachims, 2015a;b; Kallus, 2018; Zhou et al., 2018) offer theoretical results and novel methodologies for policy learning from offline data, they primarily focus on an identically and independent distributed (i.i.d.) setting, in which the underlying mechanisms do over different environments. In practice, however, shifts between environments often occur, possibly invalidating the i.i.d. assumption. As a result, a learning agent that ignores environmental shifts may fail to generalize beyond the environment it was trained on.

In the supervised learning context, the environmental shift problem has been studied under different terms, such as domain generalization, distributional robustness or out-of-distribution generalization (Muandet et al., 2013; Volpi et al., 2018; Arjovsky et al., 2019; Christiansen et al., 2020). In domain generalization, the goal is to develop learning algorithms that are robust to changes in the test distribution. Thus a fundamental problem is how to characterize such changes. A promising direction relies on a causal framework to describe the changes through the concept of interventions (Schölkopf et al., 2012; Rojas-Carulla et al., 2018; Arjovsky et al., 2019; Christiansen et al., 2020; Magliacane et al., 2018). A key insight is that while purely predictive methods perform best if test and training distributions coincide, causal models generalize to arbitrarily strong interventions on the covariates.

In real world applications, however, causal knowledge may not be available. In recent years, invariance-based methods have been exploited to learn causal structure from data (Peters et al., 2016; Pfister et al., 2018; Heinze-Deml et al., 2018). The underpinning assumption is the invariance assumption, which posits the existence of a set of predictors X in which the mechanism between X and Y remains constant. A model based on such invariant predictors is guaranteed to generalize to all unseen environments. Although some recent studies have explored the use of causality and invariance for tackling the environmental shifts problem in reinforcement learning problems (Zhang et al., 2020; Sonar et al., 2020), the actual roles and benefits of causality and

[1]University of Copenhagen. Correspondence to: Sorawit Saengkyongam <ss@math.ku.dk>.

invariance remain unclear and under-explored.

Our paper delineates an explicit connection among causality, invariance, and the environmental shift problem in the context of contextual bandits. We develop a causal framework for characterizing the environmental shift problem, and provide a theoretically sound solution based on the proposed framework. Our framework differs from the framework of causal bandits (Lee & Bareinboim, 2018; Lattimore et al., 2016; Yabe et al., 2018; de Kroon et al., 2020). While causal bandits exploit causal knowledge for improving the finite sample performance in a single environment, our framework focuses on modeling distributional shifts and the ability to generalize to new environments.

In this work, we propose a multi-environment contextual bandit framework that represents mechanisms underlying a contextual bandit problem by structural causal models (SCMs) (Pearl, 2009). The framework allows for changes in environments and thereby relaxes the i.i.d. assumption. We define environments as different perturbations on the underlying SCM, and we evaluate the policy according to its worst-case performance in all possible environments. By using the proposed framework, we generalize the invariance assumption used in methods such as invariant causal prediction (Peters et al., 2016) and define an invariant policy that, under certain assumptions, is guaranteed to generalize across different environments.

2. A Causal Framework for Multi-environment Contextual Bandits

We consider a contextual bandit setup in which part of the contexts are unobserved, where we assume that the context variables can be partitioned into observed and unobserved variables X and U. We introduce a collection $\boldsymbol{\mathcal{E}}$ of environments such that, the context (X, U) and reward R are drawn from an environment specific distribution $\mathbb{P}^e_{X,U,R}$. We now introduce a framework that puts assumptions on how environments change the distributions of X, U and R. The assumptions are constructed via an underlying class of SCMs indexed by the environment and policy.

Setting 1 (Multi-environment Contextual Bandits)**.** *Let $\boldsymbol{\mathcal{X}} = \mathcal{X}^1 \times \ldots \times \mathcal{X}^d$, $\boldsymbol{\mathcal{U}} = \mathcal{U}^1 \times \ldots \times \mathcal{U}^p$ and $\mathcal{A} = \{a^1, \ldots, a^k\}$ be measureable spaces, let $\Pi := \{\boldsymbol{\mathcal{X}} \to \Delta(\mathcal{A})\}$ be the set of all policies and let $\boldsymbol{\mathcal{E}}$ be a collection of environments. For all $\pi \in \Pi$ and all $e \in \boldsymbol{\mathcal{E}}$ we consider the following SCMs,*

$$\mathcal{S}(\pi, e): \quad \begin{cases} U := s(X, \epsilon_U) \\ X := h_e(X, U, \epsilon_X) \\ A := g_\pi(X, \epsilon_A) \\ R := f(X, U, A, \epsilon_R), \end{cases} \tag{1}$$

where $(X, U, A, R) \in \boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{U}} \times \mathcal{A} \times \mathbb{R}$, s, $(h_e)_{e \in \boldsymbol{\mathcal{E}}}$, and f are measurable functions, $\epsilon = (\epsilon_U, \epsilon_X, \epsilon_A, \epsilon_R)$ is a random vector with a distribution $Q_\epsilon = Q_{\epsilon_U} \otimes Q_{\epsilon_X} \otimes Q_{\epsilon_A} \otimes Q_{\epsilon_R}$ whose ϵ_A component is uniform on $(0, 1)$ and g_π is a function such that for all $x \in \boldsymbol{\mathcal{X}}$ it holds that $g_\pi(x, \epsilon_A)$ is a discrete distribution on $\mathcal{A}$ with probabilities $\pi(x)$.

We assume there exists a probability measure μ on $\boldsymbol{\mathcal{X}} \times \boldsymbol{\mathcal{U}} \times \mathcal{A} \times \mathbb{R}$ such that for all $\pi \in \Pi$ and all $e \in \boldsymbol{\mathcal{E}}$ the SCM $\mathcal{S}(\pi, e)$ induces a unique distribution $\mathbb{P}^{\pi,e}$ over (X, U, A, R) (see (Bongers et al., 2016) for details) which is dominated by μ and has full support on X. We denote the corresponding density by $p^{\pi,e}$ and the corresponding expectations by $\mathbb{E}^{\pi,e}$. Whenever a probability, density or expectation does not depend on π, we omit π and write $\mathbb{E}^e[X]$ rather than $\mathbb{E}^{\pi,e}[X]$, for example.

Some remarks regarding these assumptions are in order: (1) We only use the SCMs as a flexible way of modeling the changes in the joint distribution with respect to the environment e and the policy π. In particular, we do not use it to model any intervention distributions that do not correspond to a change of policy or environment. (2) In practice, the precise form of the SCM is unkown. In particular, we neither assume knowledge of the functions of the SCM nor of the structure of the SCM.

We are now ready to define contextual bandits with multiple environments.

Definition 1 (Multi-environment Contextual Bandits)**.** *Assume Setting 1, then in a multi-environment contextual bandit setup, before the beginning of each round, the system is in an environment $e \in \boldsymbol{\mathcal{E}}$, while the agent deploys a policy $\pi : \boldsymbol{\mathcal{X}} \to \Delta(\mathcal{A})$. Then, the system generates a context (X, U) and reveals only the observable X and the environment label e to the agent. Based on the observed context X, the agent selects an action A according to the policy π. The agent then receives a reward R, depending on the chosen action A and on both the observed and unobserved contexts (X, U).*

More precisely, we assume for all $i \in \{1, \ldots, n\}$ that (X_i, U_i, A_i, R_i) are sampled independently according to $\mathbb{P}^{\pi_i, e_i}_{X,U,A,R}$ (see Setting 1). When $|\boldsymbol{\mathcal{E}}| = 1$, the setup reduces to a standard contextual bandit setup.

In the multi-environment contextual bandit setup, the context variables on different rounds are not identically distributed due to the environment perturbations, this permits us to consider situations in which the training environments differ from the test environments. The assumptions in Setting 1 constrain how the environments affect the context and reward variables. Specifically, an environment e can only perturb the distribution of the reward R through altering the structural assignments of X. This constraint makes it possible to generalize information learned from one set of environments to another.

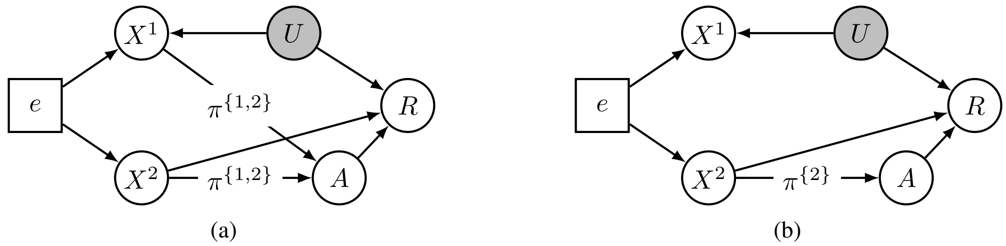

Figure 1. The graphs induced by the SCM $S(\pi, e)$ in Example 1 with a policy $\pi^{\{1,2\}}$ that takes both X^1, X^2 as inputs (left), and with a policy $\pi^{\{2\}}$ that takes only X^2 as an input (right). The policy $\pi^{\{2\}}$ is invariant w.r.t. $S = \{2\}$, which is therefore an invariant set. If data are only available from the system in the left figure, we require testing under distributional shift to test this invariance of $\pi^{\{2\}}$.

In this formulation, even though the conditional distribution of reward $\mathbb{P}^{\pi,e}_{R|X,U,A}$ is assumed to be invariant across environments, the distribution after marginalizing out U, $\mathbb{P}^{\pi,e}_{R|X,A}$, is not necessarily invariant, and since the variables U are unobserved we only have partial information about $\mathbb{P}^{\pi,e}_{R|X,A}$ in practice.

Example 1 (Linear Confounded Multi-environment Contextual Bandits)**.** *Consider a linear confounded multi-environment contextual bandit with the following SCMs*

$$\mathcal{S}(\pi, e): \quad \begin{cases} U := \epsilon_U \\ X^1 := \gamma_e U + \epsilon_{X^1} \\ X^2 := \alpha_e + \epsilon_{X^2} \\ A := g_\pi(X^1, X^2, \epsilon_A) \\ R := \begin{cases} X^2 + U + \epsilon_R, & \textit{if } A = 0 \\ X^2 - U + \epsilon_R, & \textit{if } A = 1, \end{cases} \end{cases}$$

where $\epsilon_R, \epsilon_A, \epsilon_{X^1}, \epsilon_{X^2}, \epsilon_{X^3}$ are jointly independent noise variables with zero mean, $\gamma_e, \alpha_e \in \mathbb{R}$ for all $e \in \boldsymbol{\mathcal{E}}$ and $\mathcal{A} = \{0, 1\}$. Figure 1 depicts the induced graph $\mathcal{G}^\pi$. In this example, the environments influence the observable contexts in two ways: (a) they change the mean of X^2 via α_e and (b) they change the conditional mean of X^1 given U via γ_e, while the rest of the components remain fixed across different environments. Here, the environment-specific coefficient γ_e modifies the correlation between the observable X^1 and the unobserved variable U, and consequently between X^1 and the reward R. Thus, an agent that uses information from X^1 to predict the reward R in the training environments may fail to generalize to other environments. To see this, consider a training environment $e = 1$ and a test environment $e = 2$ and let $\gamma_1 = 1$, $\gamma_2 = -1$ be the environment-specific coefficients in the training and test environment, respectively. In the training environment, we have a large positive correlation between X^1 and U, and consequently the agent will learn that the action $A = 0$ yields higher reward when observing positive value of X^1 (and $A = 1$ otherwise). However, the correlation between X^1 and U becomes negative (and large in absolute value) in the test environment, which means that the policy that the agent learnt from the training environment will now be harmful. We will see in Section 3 that a policy which depends on invariant variables (X^2 in this example) does not suffer from this generalization problem and is guaranteed to generalize across different environments.

2.1. Distributionally Robust Policies

To evaluate the performance of an agent across different environments, we define a policy value that takes into account environments. In particular, we focus on the worst-case performance of the agent across environments.

Definition 2 (Robust Policy Value)**.** *For a fixed policy* $\pi \in \Pi$*, and a set of environments* $\boldsymbol{\mathcal{E}}$*, we define the* robust policy value $V^{\boldsymbol{\mathcal{E}}}(\pi) \in \mathbb{R}$ *as the worst-case expected reward*

$$V^{\boldsymbol{\mathcal{E}}}(\pi) := \inf_{e \in \boldsymbol{\mathcal{E}}} \mathbb{E}^{\pi,e}\big[R\big].$$

Intuitively, an agent that maximizes the robust policy value is expected to perform well regardless of the environment. We now assume that for several environments $e \in \boldsymbol{\mathcal{E}}^{\mathrm{obs}}$, we are given an i.i.d. sample from a multi-environment contextual bandit, see Definition 1. More precisely, we assume to observe $D := \{(X_i, A_i, R_i, \pi_i(X_i), e_i)\}_{i=1}^n$, where $e_i \in \boldsymbol{\mathcal{E}}^{\mathrm{obs}}$, $A_i \sim \pi_i(X_i)$, $(X_i, A_i, R_i) \overset{\text{ind.}}{\sim} \mathbb{P}^{\pi_i, e_i}_{X,A,R}$ for all $i \in \{1, \ldots, n\}$. Using only D, we aim to solve the following maximin problem:

$$\arg\max_{\pi \in \Pi} V^{\boldsymbol{\mathcal{E}}}(\pi). \qquad (2)$$

Directly solving the maximin problem (2) is not feasible if we do not observe all the environments. A naive approach to this problem is to pool the data from all training environments and learn a policy that maximizes the policy value ignoring the environment structure. We show in in Appendix A that this is indeed optimal if all relevant context variables have been observed. If, however, hidden confounding is present, the naive approach no longer guarantees to find an optimal policy and the learnt policy may fail to generalize to unseen test environments. In Section 3, we introduce the notion of policy invariance. We will show that under certain assumptions solving the maximin problem (2) amounts to finding an optimal invariant policy which is then guaranteed to generalize across environments.

3. Invariant Policy for Distributional Robustness

We now introduce invariant policies and show that the maximin problem (2) can, under certain assumptions, be reduced to finding an optimal invariant policy. For all subsets $S \subseteq \{1, \dots, d\}$, let us denote the set of all policies that depend only on X^S by $\Pi^S := \{\pi \in \Pi \mid \exists \pi^S : \boldsymbol{\mathcal{X}}^S \to \Delta(\mathcal{A}) \text{ s.t. } \forall x \in \boldsymbol{\mathcal{X}}, \pi(\cdot|x) = \pi^S(\cdot|x^S)\}$.

Because of the hidden confounding, $\mathbb{E}^{\pi,e}[R \mid X = x]$ may not be independent of the environments. We will see below that our model described in Setting 1 nevertheless ensures that parts of the conditional mean may be invariant. To make this precise, we first define an invariant policy.

Definition 3 (Invariant Policies). *A policy π is said to be* invariant *with respect to a subset $S \subseteq \{1, \dots, d\}$ if $\pi \in \Pi^S$ and for all $e, f \in \boldsymbol{\mathcal{E}}$ and all $x \in \boldsymbol{\mathcal{X}}^S$ it holds that*

$$\mathbb{E}^{\pi,e}\left[R \mid X^S = x\right] = \mathbb{E}^{\pi,f}\left[R \mid X^S = x\right]. \tag{3}$$

The conditioning set S in the above definition is important and motivates the following definition.

Definition 4 (Invariant Sets). *A subset $S \subseteq \{1, \dots, d\}$ is said to be an* invariant set *if there exists $\pi \in \Pi^S$ that satisfies* (3).

By definition, the invariance property of a subset S is related to the invariance of a single policy. However, under Assumption 1 (see Appendix B), the following proposition shows that if a subset S is invariant then all policies in Π^S are also invariant.

Proposition 1. *Assume Setting 1 and Assumption 1. Then, for every subset $S \subseteq \{1, \dots, d\}$ and for all policies $\pi, \tilde{\pi} \in \Pi^S$, it holds that*

$$\pi \text{ satisfies (3)} \iff \tilde{\pi} \text{ satisfies (3)}.$$

Proof. See Appendix D.3. □

In other words, a policy $\pi \in \Pi^S$ is an invariant policy if and only if S is an invariant set. This allows us to define the set of all invariant policies in terms of the invariant subsets. Formally, we denote the set of all invariant policies by

$$\Pi_{\text{inv}} := \{\pi \in \Pi \mid \exists S \text{ s.t. } \pi \text{ is invariant w.r.t. } S\}. \tag{4}$$

Equivalently, this set can also be written as

$$\Pi_{\text{inv}} = \bigcup_{S \in \mathbf{S}^{\text{inv}}} \Pi^S,$$

where $\mathbf{S}^{\text{inv}} := \{S \subseteq \{1, \dots, d\} \mid S \text{ is invariant}\}$ is the set of all invariant subsets.

Invariant sets and policies play a central role in solving the maximin problem (2). The invariance condition (3) ensures that the conditional expectation of the reward under an invariant policy has to remain identical across different environments. Intuitively, this means that an invariant policy π that yields a high reward in the training environments is expected to also yield a high reward in the unseen test environments.

Proposition 2. *Assume Setting 1 and Assumption 1. Let Π_{inv} be the set of all invariant policies defined in* (4)*. Furthermore, let $\boldsymbol{\mathcal{E}}^{\text{obs}} \subset \boldsymbol{\mathcal{E}}$ be a set of training environments. We then have that*

$$\underset{\pi \in \Pi_{\text{inv}}}{\arg\max}\, V^{\boldsymbol{\mathcal{E}}}(\pi) = \underset{\pi \in \Pi_{\text{inv}}}{\arg\max}\, V^{\boldsymbol{\mathcal{E}}^{\text{obs}}}(\pi) \tag{5}$$

Proof. See Appendix D.4. □

In general, this is not true for non-invariant policies. In fact, we will see in Theorem 1 that under certain assumptions on the set $\boldsymbol{\mathcal{E}}$ (see Appendix B) of environments, the worst-case reward of any non-invariant policy that is optimal on the training environments will be upper bounded by the worst-case reward of the optimal invariant policy.

Theorem 1. *Assume Setting 1, Assumptions 1 and 2, and that both Π_{inv} and $\Pi \setminus \Pi_{\text{inv}}$ are non-empty. Given a set of training environments $\boldsymbol{\mathcal{E}}^{\text{obs}} \subset \boldsymbol{\mathcal{E}}$, let us define $\bar{\pi}^* \in \arg\max_{\pi \in \Pi_{\text{inv}}} V^{\boldsymbol{\mathcal{E}}^{\text{obs}}}(\pi)$, $\underline{\pi}^* \in \arg\max_{\pi \in \Pi \setminus \Pi_{\text{inv}}} V^{\boldsymbol{\mathcal{E}}^{\text{obs}}}(\pi)$ as optimal policies under the training environments. Then*

$$V^{\boldsymbol{\mathcal{E}}}(\bar{\pi}^*) \geq V^{\boldsymbol{\mathcal{E}}}(\underline{\pi}^*).$$

Proof. See Appendix D.5. □

The above results motivate a procedure to solve the maximin problem (2). Proposition 2 implies that if we consider a policy class containing only the invariant policies, the maximin problem will reduce to a standard policy optimization problem, while Theorem 1 shows that an optimal invariant policy obtained in (5) always yields better or equal performance in terms of the robust policy value compared to a non-invariant one. In other words, given a training dataset D, we seek to operationalize the following two steps: (a) using D find the set Π_{inv} of all possible invariant policies, (b) use standard methods for offline contextual bandits to solve $\arg\max_{\pi \in \Pi_{\text{inv}}} V(\pi)$ on the data set D.

4. Learning an Invariant Policy

In the previous section, we showed that if a subset $S \subseteq \{1, \dots, d\}$ is invariant, policies $\pi^S \in \Pi^S$ generalize to unseen environments under suitable assumptions. We now consider the task of testing from data whether a set S is invariant. Let therefore $H_0(S, \pi)$ be the hypothesis that π and S satisfy the invariance property

$$H_0(S,\pi): \quad \forall e, f \in \boldsymbol{\mathcal{E}}, \forall x \in \boldsymbol{\mathcal{X}}^S : \qquad (6)$$
$$\mathbb{E}^{\pi,e}[R \mid X^S = x] = \mathbb{E}^{\pi,f}[R \mid X^S = x].$$

Consider now a fixed set S. By Definition 4, S is invariant if there exists a $\pi^S \in \Pi^S$, such that $H_0(S, \pi^S)$ is true. In this paper, we assume that the offline data are sampled from an initial policy π^0 that does not necessarily satisfy $\pi^0 \in \Pi^S$. Therefore, testing $H_0(S, \pi^0)$ is not the right step forward because it may happen that S is invariant but $H_0(S, \pi^0)$ is false. To see this, consider, e.g., the setting in Figure 1(b). By Lemma 1 (see Appendix D.3), the set $S := \{X^2\}$ is invariant, i.e., $H_0(S, \pi^S)$ is true. Yet, if the initial policy π^0 depends on both $\{X^1, X^2\}$, then $H_0(S, \pi^{\{1,2\}})$ is not true: In Figure 1(a), the path $e \to X^1 \to A \to R$ is open, and so (given Assumption 1) we do not have invariance of $\mathbb{E}^{\pi^{\{1,2\}},e}[R \mid X^{\{2\}}]$.

Thus, we cannot directly test the invariance of a set S by using the initial policy and the observed data. Instead, we need to test $H_0(S, \pi^S)$ for a policy $\pi^S \in \Pi^S$ that is different from π^0. As we detail in Appendix C, we can do so by applying the method for testing invariance under distributional shifts (Thams et al., 2021) by resampling the data to mimick the policy under π^S.

5. Simulations

To verify our theoretical findings we perform two simulation experiments, where we consider a linear multi-environment contextual bandit setting. Appendix F contains the simulation details.

5.1. Generalization of Invariant Sets

Here, we first consider an oracle setting, where we know a priori which subsets are invariant. From our data generating process, it follows that $\{X_2\}$ is the the only invariant set. We then compare an invariant policy which depends only on X_2 with a policy that uses both X_1 and X_2. We train both policies on a data set of size $10'000$ obtained from multiple training environments under a fixed initial policy π_0. Then we evaluate both policies on multiple unseen environments, where we compute the regret with respect to the policy that is optimal in each of the unseen environments. Figure 2 illustrates the experimental result. Each data point represents the evaluation on an unseen environment. The y-axes show the regret value and the x-axes display the distance from each unseen environment to the training environments. The distance is computed as the ℓ^2-norm between the average value of the pairs $(\alpha^e, \mu^e_{X^2})$ in the training environments and the pair $(\alpha^e, \mu^e_{X^2})$ in the unseen test environment. The plot shows that the worst-case behavior of the invariant policy is smaller than the non-invariant one. This empirically supports our result of Theorem 1.

5.2. Learning an Invariant Policy

In practice, we do not know in advance which sets are invariant. We now aim to find an invariant policy from a data set generated under an initial policy π_0 which takes both X^1 and X^2 as input. To do so, we employ the method proposed in Section 4 for testing invariance under distributional shifts. The resulting acceptance rates are shown in Figure 3. Our method yields high acceptance rates for the set $\{X^2\}$, which is the correct invariant set, while the acceptance rates for other sets gradually decrease as the sample size increases. Furthermore, we can see that our testing method is more accurate when the number of training environments increases.

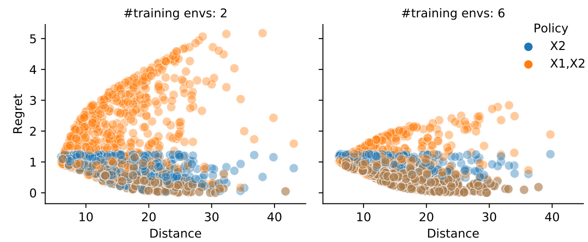

Figure 2. The generalization performance (in terms of regret) of the policy based on an invariant set (X_2) and the policy based on a non-invariant set (X_1, X_2). The left and the right plot show the results when the training environments consist of two and six different environments, respectively. In both cases, the worst-case regret for the invariant policy is evidently smaller than for the non-invariant policy.

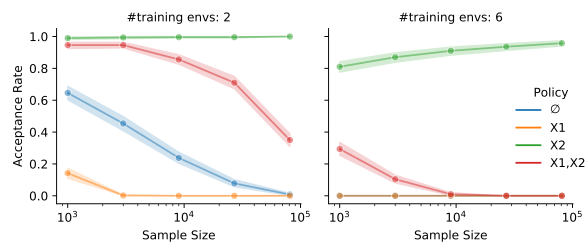

Figure 3. Acceptance rates for the off-policy invariance test. The proposed method correctly accepts the set $\{X^2\}$ with high chance, while gradually rejects other sets.

References

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Bongers, S., Forré, P., Peters, J., and Mooij, J. M. Foundations of structural causal models with cycles and latent variables. *The Annals of Statistics (accepted), arXiv preprint arXiv:1611.06221*, 2016.

Bottou, L., Peters, J., Quiñonero-Candela, J., Charles, D. X., Chickering, D. M., Portugaly, E., Ray, D., Simard, P., and Snelson, E. Counterfactual reasoning and learning systems: The example of computational advertising. *Journal of Machine Learning Research*, 14(65):3207–3260, 2013.

Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. A causal framework for distribution generalization. *arXiv preprint arXiv:2006.07433*, 2020.

Dawid, A. P. Influence diagrams for causal modelling and inference. *International Statistical Review*, 70(2):161–189, 2002.

de Kroon, A. A., Belgrave, D., and Mooij, J. M. Causal discovery for causal bandits utilizing separating sets. *arXiv preprint arXiv:2009.07916*, 2020.

Dudik, M., Langford, J., and Li, L. Doubly robust policy evaluation and learning. In *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 1097–1104, New York, NY, USA, 2011. ACM.

Gretton, A., Borgwardt, K. M., Rasch, M. J., Schölkopf, B., and Smola, A. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):1–35, 2018.

Kallus, N. Balanced policy evaluation and learning. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Lattimore, F., Lattimore, T., and Reid, M. D. Causal bandits: Learning good interventions via causal inference. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems (NeurIPS) 29*. 2016.

Lauritzen, S. L., Dawid, A. P., Larsen, B. N., and Leimer, H.-G. Independence properties of directed Markov fields. *Networks*, 20:491–505, 1990.

Lee, S. and Bareinboim, E. Structural causal bandits: Where to intervene? In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R. (eds.), *Advances in Neural Information Processing Systems (NeurIPS) 31*. 2018.

Magliacane, S., van Ommen, T., Claassen, T., Bongers, S., Versteeg, P., and Mooij, J. M. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Muandet, K., Balduzzi, D., and Schölkopf, B. Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pp. 10–18. PMLR, 2013.

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

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society, Series B (Statistical Methodology) (with discussion)*, 78, 2016.

Peters, J., Janzing, D., and Schölkopf, B. *Elements of causal inference: foundations and learning algorithms*. The MIT Press, 2017.

Pfister, N., Bühlmann, P., and Peters, J. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2018.

Rojas-Carulla, M., Schölkopf, B., Turner, R., and Peters, J. Causal transfer in machine learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pp. 1255–1262. Omnipress, 2012.

Sonar, A., Pacelli, V., and Majumdar, A. Invariant policy optimization: Towards stronger generalization in reinforcement learning. *arXiv preprint arXiv:2006.01096*, 2020.

Spirtes, P., Glymour, C., and Scheines, R. *Causation, Prediction, and Search*. MIT Press, 2nd edition, 2000.

Swaminathan, A. and Joachims, T. Counterfactual risk minimization: Learning from logged bandit feedback. In *Proceedings of the 32nd International Conference on Machine Learning (ICML)*, pp. 814–823, 2015a.

Swaminathan, A. and Joachims, T. The self-normalized estimator for counterfactual learning. In *Advances in*

Neural Information Processing Systems, volume 28, pp. 3231–3239. Curran Associates, Inc., 2015b.

Thams, N., Saengkyongam, S., Pfister, N., and Peters, J. Statistical testing under distributional shifts. *arXiv preprint arXiv:2105.10821*, 2021.

Volpi, R., Namkoong, H., Sener, O., Duchi, J. C., Murino, V., and Savarese, S. Generalizing to unseen domains via adversarial data augmentation. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Yabe, A., Hatano, D., Sumita, H., Ito, S., Kakimura, N., Fukunaga, T., and Kawarabayashi, K.-i. Causal bandits with propagating inference. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

Zhang, A., Lyle, C., Sodhani, S., Filos, A., Kwiatkowska, M., Pineau, J., Gal, Y., and Precup, D. Invariant causal prediction for block MDPs. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119, pp. 11214–11224. PMLR, 2020.

Zhou, Z., Athey, S., and Wager, S. Offline multi-action policy learning: Generalization and optimization. *arXiv preprint arXiv:1810.04778*, 2018.

A. Policy Learning without Unobserved Confounders

This section illustrates a setting in which it is not beneficial to explicitly take into account the environment structure. Here, simply pooling the data from all training environments and applying a standard value-based policy learning algorithm yields a solution to (2). This result sheds light on the role of causality and invariance in contextual bandits and reinforcement learning.

Setting 2 (Unconfounded Multi-environment Contextual Bandits)**.** *In this setting, we assume that there is no unobserved confounder in the underlying causal model. More precisely, we modify the model class $S(\pi, e)$ in Setting 1 and consider*

$$\mathcal{S}(\pi, e): \quad \begin{cases} U := s(X, \epsilon_U) \\ X := h_e(X, \epsilon_X) \\ A := g_\pi(X, \epsilon_A) \\ R := f(X, U, A, \epsilon_R). \end{cases}$$

The following theorem shows that in Setting 2 there is a population optimal policy that does not depend on the environment perturbations. In particular, the this optimal policy can be learned from data obtained in any environment subset $\boldsymbol{\mathcal{E}}^{\mathrm{obs}} \subseteq \boldsymbol{\mathcal{E}}$.

Theorem 2. *Assume Setting 2, let $\boldsymbol{\mathcal{E}}^{\mathrm{obs}} \subseteq \boldsymbol{\mathcal{E}}$ be a non-empty subset of observed environments and let $\pi^* \in \Pi$ the policy defined for all $x \in \boldsymbol{\mathcal{X}}$ and all $a \in \mathcal{A}$ by*

$$\pi^*(a \mid x) := \mathbb{1}\big[a = \operatorname*{arg\,max}_{a' \in \mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x, a')\big], \tag{7}$$

where $Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) := \frac{1}{|\boldsymbol{\mathcal{E}}^{\mathrm{obs}}|} \sum_{e \in \boldsymbol{\mathcal{E}}^{\mathrm{obs}}} \mathbb{E}^{\pi_a, e}[R \mid X = x]$, $\mathbb{1}[\cdot]$ denotes the indicator function and π_a is the policy that always selects a. Then,

$$\pi^* \in \operatorname*{arg\,max}_{\pi \in \Pi} V^{\boldsymbol{\mathcal{E}}}(\pi),$$

i.e., π^* is a solution to the maximin problem (2).

Proof. See Appendix D.1. □

The proof of Theorem 2 uses that the conditional expectation of $\mathbb{E}^{\pi_a, e}[R \mid X]$ does not depend on the environment. This, in particular, implies that $Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x, a) = \mathbb{E}^{\pi_a, e}[R \mid X = x]$ for any $e \in \boldsymbol{\mathcal{E}}$. Theorem 2 therefore suggests that we can estimate the optimal policy by pooling the data from training environments and applying a standard value-based policy learning algorithm. Let therefore $\widehat{Q}_n$ be an estimator of the conditional mean $\mathbb{E}^{\pi_a}[R \mid X]$ that is based on n independent observations (X_i, A_i, R_i) from potentially different environments. The following corollary shows that such an approach indeed yields a consistent estimate of the optimal policy given that $\widehat{Q}_n$ is consistent. (In fact, the same arguments would work if instead of pooling, one considers only one environment. But in practice, one would make use of all data available.)

Corollary 1. *Assume Setting 2 and let $\widehat{Q}_n$ be a uniformly consistent estimator of $Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}$, that is, for all $a \in \mathcal{A}$ it holds that*

$$\lim_{n\to\infty} \mathbb{E}_D \left[\sup_{x \in \boldsymbol{\mathcal{X}}} \left| \widehat{Q}_n(x,a) - Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) \right| \right] = 0,$$

where $\mathbb{E}_D$ is an expectation over the n observations (X_i, A_i, R_i) used to estimate $\widehat{Q}_n$. Then, for $\widehat{\pi}_n := \mathbb{1}\big[a = \arg\max_{a' \in \mathcal{A}} \widehat{Q}_n(x,a')\big]$ the robust policy value converges towards its optimal value, that is,

$$\lim_{n\to\infty} \mathbb{E}_D \left[\left| V^{\boldsymbol{\mathcal{E}}}(\widehat{\pi}_n) - V^{\boldsymbol{\mathcal{E}}}(\pi^*) \right| \right] = 0,$$

where π^* is the optimal policy defined in (7)

Proof. See Appendix D.2. □

Whether it is possible to construct a uniformly consistent estimator $\widehat{Q}_n$ depends on the model class that can been assumed in the structural assignment of R, and on the policy used in generating the observations. For example, in the case of additive confounding and noise such as $f(X, U, A, \epsilon_R) = g(X, A) + h(U, \epsilon_R)$ with $g \in \mathcal{F}$ for some function class $\mathcal{F}$) and a policy π that has full support, i.e., $\forall a \in \mathcal{A}, x \in \mathcal{X} : \pi(a \mid x) > 0$, one can consider a least squares estimator of the form

$$\widehat{Q}_n^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}} := \arg\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} (f(X_i, A_i) - R_i)^2.$$

The assumptions of Corollary 1 are then satisfied under further constraints on the function class and noise distributions, such as linear functions, Gaussian noise and bounded domains.

Corollary 1 implies that without hidden confounders, we do not benefit from taking into account the environment structure and considering invariance. However, the following section shows that this is different when hidden confounders do exist.

B. Assumptions

By the Markov condition, which holds in SCMs, a d-separation statement in a graph implies conditional independence (Pearl, 2009; Lauritzen et al., 1990; Peters et al., 2017). Faithfulness (Spirtes et al., 2000) assumes that the converse holds, too.

Assumption 1 (Mean Faithfulness)**.** *For all $S \subseteq \{1, \dots, d\}$ and for all $\pi \in \Pi^S$ it holds that*

$$\begin{aligned} &\forall e, f \in \boldsymbol{\mathcal{E}}, \forall x \in \boldsymbol{\mathcal{X}}^S : \\ &\quad \mathbb{E}^{\pi,e}[R \mid X^S = x] = \mathbb{E}^{\pi,f}[R \mid X^S = x] \\ &\qquad \Longrightarrow \\ &\qquad R \perp\!\!\!\perp_{\mathcal{G}^\pi} e \mid X^S, \end{aligned}$$

where the symbol $\perp\!\!\!\perp_{\mathcal{G}^\pi}$ denotes d-separation in $\mathcal{G}^\pi$.

We now outline the assumptions on the set of environments $\boldsymbol{\mathcal{E}}$ facilitation this result. As we will see in the proof of Theorem 1 the crucial difference between invariant and non-invariant policies is that non-invariant policies use information - related to variables confounded with the reward - that may change across environments. In cases where the environments do not change the system 'too strongly' it can therefore happen that using such information is beneficial across all environments. In practice one might however not know how strong the test environments can change the system in which case such information can become useless or even harmful. Intuitively, this happens, for example, if environments exist where the non-invariant confounded variables no longer contain any information about the reward. ((Christiansen et al., 2020) use confounding removing interventions in the setting of prediction.) Formally, we make the following definition.

Definition 5 (Confounding Removing Environments)**.** *Assume Setting 1. An environment e is said to be a confounding removing environment if there exists $\pi \in \Pi$ such that for all $j \in \{1, \dots, d\}$ for which*

$$\forall S \subseteq \{1, \dots, d\} : R \not\!\perp\!\!\!\perp_{\mathcal{G}^\pi} e \mid X^{S \cup \{j\}}$$

it holds that

$$X^j \perp\!\!\!\perp_{\mathcal{G}^{\pi,e}} U,$$

where $\mathcal{G}^{\pi,e}$ is the graph corresponding to the SCM $\mathcal{S}(\pi, e)$.

The existence of confounding removing environments implies that at least in some of the environments it is impossible to benefit from a non-invariant policy. However, this is not necessarily sufficient to ensure that one can not benefit in the worst-case. Therefore we make the following additional assumption.

Assumption 2 (Strong Environments)**.** *For all $e \in \boldsymbol{\mathcal{E}}$, there exists $f \in \boldsymbol{\mathcal{E}}$ such that f is a confounding removing environment and $\mathbb{P}_X^{\pi,e} = \mathbb{P}_X^{\pi,f}$.*

C. Testing Invariance under Distributional Shifts

To test the hypothesis $H_0(S, \pi^S)$ for some set $S \subseteq \{1, \dots, d\}$, we apply the off-policy test from (Thams et al., 2021), which proposes to draw a target sample from π^S by resampling the offline data that were drawn from π^0, and to test invariance in the target sample.

More formally, we assume that for every $e \in \boldsymbol{\mathcal{E}}^{\mathrm{obs}}$ we observe a data set D^e consisting of n_e observations[1] $D_i^e = (X_i^e, A_i^e, R_i^e, \pi^0(A_i^e|X_i^e))$. For all $e \in \boldsymbol{\mathcal{E}}^{\mathrm{obs}}$ and all $i \in \{1, \dots, n_e\}$ define the relative weights as

$$r(D_i^e) := \frac{\pi^S(A_i^e|X_i^e)}{\pi^0(A_i^e|X_i^e)},$$

where $\pi^S \in \Pi^S$ is a target policy. Then, for all $e \in \boldsymbol{\mathcal{E}}^{\mathrm{obs}}$, we draw a weighted resample $D^{e,\pi^S} := (D_{i_1}^e, \dots, D_{i_{m_e}}^e)$ of size m_e from D^e with weights

$$w_{i_1,\dots,i_{m_e}}^e := \begin{cases} \frac{\prod_{\ell=1}^{m_e} r(D_{i_\ell}^e)}{\sum_{(j_1,\dots,j_{m_e}) \text{ distinct}} \prod_{\ell=1}^{m_e} r(D_{j_\ell}^e)} & (i_1, \dots, i_{m_e}) \text{ distinct} \\ 0 & \text{otherwise.} \end{cases} \tag{8}$$

We then apply an invariance test to the resampled data $D^{e_1,\pi^S}, \dots, D^{e_L,\pi^S}$. An invariance test φ^S is a function that takes data from environments $e_1, \dots, e_L$, each of size m_{e_i}, and tests whether S is invariant

$$\varphi^S : \mathcal{D}^{m_{e_1}+\dots+m_{e_L}} \to \{0, 1\}.$$

Here, $\varphi = 1$ indicates that we reject the hypothesis of invariance, and such a test is said to have pointwise asymptotic level if for all S invariant and all $\pi \in \Pi^S$ it holds that

$$\limsup_{n_{e_1},\dots,n_{e_L} \to \infty} \mathbb{P}^\pi(\varphi^S(D^{e_1}, \dots, D^{e_L}) = 1) \leq \alpha.$$

We detail a concrete test φ below, but we first state that the overall procedure has asymptotic level if the test φ has asymptotic level under π. For simplicity, we assume that $n_{e_1} = \dots = n_{e_L} =: n$ and $m_{e_1} = \dots = m_{e_L} =: m$. The following result follows directly from (Thams et al., 2021).

Proposition 3. *Suppose that for each environment $e_1, \dots, e_L$, we observe a data set D^e consisting of n observations $D_i^e = (X_i^e, A_i^e, R_i^e, \pi_i^0(A_i^e \mid X_i^e))$. Let $m = o(\sqrt{n})$ and assume that for all $e \in \boldsymbol{\mathcal{E}}$, $\mathbb{E}^{\pi^0}[r(D_i^e)^2] < \infty$. Consider $\pi^S \in \Pi^S$ and for all e, let $D^{e,\pi} := (D_{i_1}^e, \dots, D_{i_m}^e)$ be a resample of D^e drawn with weights given by (8). Let φ be a hypothesis test for invariance of the conditional expectation $\mathbb{E}^{\pi,e}[R \mid X^S]$ that has asymptotic level $\alpha \in (0, 1)$ when φ is applied to data sampled from π. Applying φ to the resampled data yields pointwise asymptotic level, that is,*

$$\limsup_{n \to \infty} \mathbb{P}^{\pi^0}(\varphi(D^{e_1,\pi}, \dots, D^{e_L,\pi}) = 1) \leq \alpha$$

if S is invariant.

[1] It is even possible to allow a different initial policy π_i^0 at each observation i. One then needs to define the relative weights as $r(D_i^e) := \pi^S(A_i^e|X_i^e)/\pi_i^0(A_i^e|X_i^e)$.

In other words, we can test whether X^S is invariant by resampling the data and applying an invariance test on the resampled data set. Proposition 3 states that this procedure holds level if the sample size goes to infinity.

C.1. Choice of Target Test

We now detail a test φ to test invariance in the target sample. We first pool data from all environments into one data set and estimate the conditional $\mathbb{E}[R \mid X^S]$ using any prediction method (such as linear regression or a neural network). We then test whether the residuals $R - \mathbb{E}[R \mid X^S]$ are equally distributed over the environments $e \in \boldsymbol{\mathcal{E}}$, i.e., we split the sample back into L groups (corresponding to the environments) and test whether the residuals in these groups are equally distributed (see also (Peters et al., 2016), for example). We then define φ to be the composition of these operations, that is, φ returns 1 if the test for equal distribution of the residuals is rejected.

In the simulations in Section 5 below, we use an F-test to test whether the residuals have the same mean across environments; this test holds pointwise asymptotic level for all $\alpha \in (0, 1)$ (see Proposition 3). To obtain power against more alternatives, one could also use non-parametric tests, such as two-sample kernel tests using the maximum mean discrepancy (Gretton et al., 2012) and then correct for the multiple testing using Bonferroni-corrections (see also (Rojas-Carulla et al., 2018), for example).

D. Proofs

D.1. Proof of Theorem 2

Proof. We begin by showing that the model class in Setting 2 satisfies an invariance property. Let $e \in \boldsymbol{\mathcal{E}}$, $a \in \mathcal{A}$ and $x \in \boldsymbol{\mathcal{X}}$, then by using the explicit SCM structure from Setting 2, it holds that

$$\begin{aligned} &\mathbb{E}^{\pi_a,e}\left[R \mid X = x\right] \\ &\quad = \mathbb{E}^{\pi_a,e}\left[f(X, s(X, \epsilon_U), A, \epsilon_R) \mid X = x\right]. \end{aligned}$$

Since we assume there is no hidden confounding, it holds that $\epsilon_U \perp\!\!\!\perp X$ which implies that

$$\begin{aligned} &\mathbb{E}^{\pi_a,e}\left[f(X, s(X, \epsilon_U), A, \epsilon_R) \mid X = x\right] \\ &\quad = \mathbb{E}_{\epsilon_U,\epsilon_R}\left[f(x, s(x, \epsilon_U), a, \epsilon_R)\right] \end{aligned}$$

and hence $\mathbb{E}^{\pi_a,e}\left[R \mid X = x\right]$ does not depend on the environment. This in particular implies that for all $e \in \boldsymbol{\mathcal{E}}$, all $x \in \boldsymbol{\mathcal{X}}$ and all $a \in \mathcal{A}$, it holds that

$$\begin{aligned} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x, a) &= \frac{1}{|\boldsymbol{\mathcal{E}}^{\mathrm{obs}}|} \sum_{f \in \boldsymbol{\mathcal{E}}^{\mathrm{obs}}} \mathbb{E}^{\pi_a,f}[R \mid X = x] \\ &= \mathbb{E}^{\pi_a,e}[R \mid X = x]. \end{aligned} \tag{9}$$

We thus have for any policy $\pi \in \Pi$ and $x \in \boldsymbol{\mathcal{X}}$ that,

$$\max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) \geq \mathbb{E}^{\pi,e}[R \mid X=x]. \tag{10}$$

Next, take the expectation over X on both sides to get

$$\begin{aligned}\mathbb{E}^{e}\big[\max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(X,a)\big] &\geq \mathbb{E}^{e}\big[\mathbb{E}^{\pi,e}[R \mid X]\big]\\ &= \mathbb{E}^{\pi,e}\big[R\big].\end{aligned}$$

Finally, taking the infimum over $e \in \boldsymbol{\mathcal{E}}$ leads to

$$\inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^{e}\big[\max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(X,a)\big] \geq \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^{\pi,e}\big[R\big]. \tag{11}$$

By definition, the policy

$$\pi^*(a \mid x) \coloneqq \mathbb{1}\big[a = \arg\max_{a'\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a')\big]$$

satisfies $\mathbb{E}^{\pi^*,e}\big[R\big] = \mathbb{E}^{e}\big[\max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(X,a)\big]$. Therefore (11), implies that π^* is an optimal policy, which completes the proof of Theorem 2. □

D.2. Proof of Corollary 1

Proof. We begin by defining for all $n \in \mathbb{N}$ the term

$$c(n) = \max_{a\in\mathcal{A}} \sup_{x\in\mathcal{X}} |Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) - \widehat{Q}_n(x,a)|.$$

As $\mathcal{A}$ is assumed to be finite and because $\widehat{Q}_n$ is assumed to be uniformly convergent it holds that

$$\lim_{n\to\infty} \mathbb{E}_D[c(n)] = 0. \tag{12}$$

Moreover, as shown in (9), in the proof of Theorem 2, we know that for all $e \in \boldsymbol{\mathcal{E}}$, all $a \in \mathcal{A}$ and all $x \in \mathcal{X}$ it holds that

$$Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) = \mathbb{E}^{\pi_a,e}[R \mid X=x].$$

This implies that for all $x \in \boldsymbol{\mathcal{X}}$ and all $e \in \boldsymbol{\mathcal{E}}$ it holds that

$$\begin{aligned}&\mathbb{E}^{\widehat{\pi}_n,e}[R \mid X=x]\\ &= \sum_{a\in\mathcal{A}} \mathbb{E}^{\pi_a,e}[R \mid X=x]\widehat{\pi}_n(a|x)\\ &= \sum_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a)\widehat{\pi}_n(a|x)\\ &= \sum_{a\in\mathcal{A}} \widehat{Q}_n(x,a)\widehat{\pi}_n(a|x)\\ &\qquad + \sum_{a\in\mathcal{A}} (Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) - \widehat{Q}_n(x,a))\widehat{\pi}_n(a|x).\end{aligned} \tag{13}$$

Next, observe that

$$\begin{aligned}&\left|\sum_{a\in\mathcal{A}} (Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) - \widehat{Q}_n(x,a))\widehat{\pi}_n(a|x)\right|\\ &\leq \sum_{a\in\mathcal{A}} \left|Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) - \widehat{Q}_n(x,a)\right|\widehat{\pi}_n(a|x)\\ &\leq c(n)\end{aligned} \tag{14}$$

and

$$\begin{aligned}&\sum_{a\in\mathcal{A}} \widehat{Q}_n(x,a)\widehat{\pi}_n(a|x)\\ &= \max_{a\in\mathcal{A}} \widehat{Q}_n(x,a)\\ &= \max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a)\\ &\qquad + (\max_{a\in\mathcal{A}} \widehat{Q}_n(x,a) - \max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a)).\end{aligned} \tag{15}$$

Using (14), (15) and (13) together with the triangle inequality leads to

$$\begin{aligned}&\left|\mathbb{E}^{\widehat{\pi}_n,e}[R \mid X=x] - \max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a)\right|\\ &= \Big|\max_{a\in\mathcal{A}} \widehat{Q}_n(x,a) - \max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a)\\ &\qquad + \sum_{a\in\mathcal{A}} (Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) - \widehat{Q}_n(x,a))\widehat{\pi}_n(a|x)\Big|\\ &\leq 2c(n).\end{aligned}$$

This in particular implies that for all $e \in \boldsymbol{\mathcal{E}}$ and all $x \in \boldsymbol{\mathcal{X}}$ it holds that

$$\max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) - 2c(n) \leq \mathbb{E}^{\widehat{\pi}_n,e}[R \mid X=x]$$

and that

$$\mathbb{E}^{\widehat{\pi}_n,e}[R \mid X=x] \leq \max_{a\in\mathcal{A}} Q^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(x,a) + 2c(n).$$

Taking the expectation over X and the infimum over $\boldsymbol{\mathcal{E}}$ in both inequalities leads to

$$V^{\boldsymbol{\mathcal{E}}}(\pi^*) - 2c(n) \leq V^{\boldsymbol{\mathcal{E}}}(\widehat{\pi}_n) \leq V^{\boldsymbol{\mathcal{E}}}(\pi^*) + 2c(n).$$

Finally, we use (12) and get that

$$\lim_{n\to\infty} \mathbb{E}_D\left[|V^{\boldsymbol{\mathcal{E}}}(\widehat{\pi}_n) - V^{\boldsymbol{\mathcal{E}}}(\pi^*)|\right] \leq \lim_{n\to\infty} \mathbb{E}_D[4c(n)] = 0.$$

This completes the proof of Theorem 1. □

D.3. Proof of Proposition 1

Proof. First, we show in the following Lemma that the opposite of the mean faithfulness holds true.

Lemma 1 (Extended Markov Property)**.** *For any fixed policy $\pi \in \Pi$, it holds for all subsets $S \subseteq \{1,\dots,d\}$ and for all $Z \in \{U,R\}$ that*

$$\begin{gathered}Z \perp\!\!\!\perp_{\mathcal{G}^\pi} e \mid X^S\\ \Longrightarrow\\ \forall e,f \in \boldsymbol{\mathcal{E}}, \forall x \in \boldsymbol{\mathcal{X}}^S :\\ \mathbb{E}^{\pi,e}[Z \mid X^S = x] = \mathbb{E}^{\pi,f}[Z \mid X^S = x],\end{gathered}$$

where the symbol $\perp\!\!\!\perp_{\mathcal{G}^\pi}$ denotes d-separation in the graph $\mathcal{G}^\pi$.

Next, fix $S \subseteq \{1, \dots, p\}$, and let $\pi \in \Pi^S$. Assume π is an invariant policy, i.e., satisfies (3). By mean faithfulness, we have that $R \perp\!\!\!\perp_{\mathcal{G}^\pi} e \mid X^S$. Furthermore, since for all $\pi' \in \Pi^S$ the graphs $\mathcal{G}_{\pi'}$ and $\mathcal{G}_\pi$ are identical, this implies that the same d-separation also holds in $\mathcal{G}_{\pi'}$. Consequently, by Lemma 1, we then have for all $\pi' \in \Pi^S$ it holds that

$$\forall e, f \in \boldsymbol{\mathcal{E}}, \forall x \in \boldsymbol{\mathcal{X}}^S : \mathbb{E}^{\pi,e}[Z \mid X^S = x] = \mathbb{E}^{\pi,f}[Z \mid X^S = x],$$

which concludes the proof of Proposition 1. $\square$

D.3.1. Proof of Lemma 1

Proof. Lemma 1 corresponds to a global Markov property in the augmented graph (including the non-random environment index). Such results are well-established and heavily used in settings in which $\boldsymbol{\mathcal{E}}$ is finite, for example in influence diagrams (Dawid, 2002). The result, however, also holds for more general $\boldsymbol{\mathcal{E}}$.

To prove this, we will first fix $\pi \in \Pi$, $S \subseteq \{1, \dots, d\}$ and $Z \in \{U, R\}$. Furthermore, let $e \in \boldsymbol{\mathcal{E}}$, let Σ be the discrete σ-algebra on $\boldsymbol{\mathcal{E}}$ and let $\nu_e : \Sigma \to [0,1]$ be a probability measure that puts non-zero mass on $\{e\}$. Adding a random variable E with distribution ν_e the class of SCMs $(\mathcal{S}(\pi, f))_{f \in \boldsymbol{\mathcal{E}}}$ induces a joint distribution over (E, X, U, A, R) that is globally Markov with respect to the graph $\mathcal{G}^\pi$. This implies that the d-separation $Z \perp\!\!\!\perp_{\mathcal{G}^\pi} E \mid X^S$ (which is implied by $Z \perp\!\!\!\perp_{\mathcal{G}^\pi} e \mid X^S$) implies that the joint distribution (E, X, U, A, R) satisfies the following conditional independence

$$Z \perp\!\!\!\perp E \mid X^S.$$

By definition this implies for all $x \in \boldsymbol{\mathcal{X}}^S$ and all $f \in \boldsymbol{\mathcal{E}}$ with $\nu_e(f) > 0$ that

$$\mathbb{E}^f[Z \,|\, X^S = x] = \mathbb{E}[Z \,|\, X^S = x, E = f] = \mathbb{E}[Z \,|\, X^S = x],$$

where the expectations without superscript are taken with respect to the joint distribution including E. The function $w(x) \coloneqq \mathbb{E}[Z, \,|\, X^S = x]$ therefore no longer depends on the environment nor on ν_e. Since $\nu_e(e) > 0$ this in particular implies that for all $x \in \boldsymbol{\mathcal{X}}^S$ it holds that

$$\mathbb{E}^e[Z \,|\, X^S = x] = w(x).$$

As this construction works for all $e \in \boldsymbol{\mathcal{E}}$, this completes the proof of Lemma 1. $\square$

D.4. Proof of Proposition 2

Proof. As before, we define $\pi_a(a' \mid x) \coloneqq \mathbb{1}\big[a' = a\big]$ as the policy that always selects the action a and $\mathbf{S}_{\text{inv}}$ as the collection of all invariant sets.

For any set $S \in \mathbf{S}_{\text{inv}}$ it holds by definition of invariant sets that Q_S^e does not depend on e, we will therefore drop the superscript e in these cases. For all $S \in \mathbf{S}_{\text{inv}}$, all policies $\pi^S \in \Pi^S$, all $x \in \boldsymbol{\mathcal{X}}$ and all $e \in \boldsymbol{\mathcal{E}}$ it further holds that

$$\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^S = x] \geq \mathbb{E}^{\pi^S, e}[R \mid X^S = x],$$

see Equation (10). Taking the expectation over X^S on both sides yields

$$\begin{aligned} \mathbb{E}^e\big[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^S]\big] &\geq \mathbb{E}^e\big[\mathbb{E}^{\pi^S, e}[R \mid X^S]\big] \\ &= \mathbb{E}^{\pi^S, e}[R]. \end{aligned} \tag{16}$$

Next, we define the invariant blanket S_{IB} as the smallest set $S_{IB} \subseteq \mathbf{S}_{\text{inv}}$ that satisfies

$$\forall S \in \mathbf{S}_{\text{inv}} \backslash S_{IB} : \quad \mathbb{E}^{\pi_a}[R \mid X^S, X^{S_{IB} \backslash S}] = \mathbb{E}^{\pi_a}[R \mid X^{S_{IB}}].$$

Then we have that

$$\begin{aligned} &\mathbb{E}^e\big[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^S]\big] \\ &= \mathbb{E}^e\big[\max_{a \in \mathcal{A}} \mathbb{E}^e_{X^{S_{IB} \backslash S}}\big[\mathbb{E}^{\pi_a, e}[R \mid X^S, X^{S_{IB} \backslash S}]\big]\big], \end{aligned}$$

by Jensen's inequality,

$$\leq \mathbb{E}^e\big[\mathbb{E}^e_{X^{S_{IB} \backslash S}}\big[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^S, X^{S_{IB} \backslash S}]\big]\big]$$

by the definition of the invariant blanket S_{IB},

$$\begin{aligned} &= \mathbb{E}^e\big[\mathbb{E}^e_{X^{S_{IB} \backslash S}}\big[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^{S_{IB}}]\big]\big] \\ &= \mathbb{E}^e\big[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^{S_{IB}}]\big]. \end{aligned}$$

Combining with (16), we have

$$\mathbb{E}^e\big[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^{S_{IB}}]\big] \geq \mathbb{E}^{\pi^S, e}[R]$$

Taking the infimum over $e \in \boldsymbol{\mathcal{E}}$ leads to

$$\inf_{e \in \boldsymbol{\mathcal{E}}} \mathbb{E}^e\big[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^{S_{IB}}]\big] \geq \inf_{e \in \boldsymbol{\mathcal{E}}} \mathbb{E}^{\pi^S, e}\big[R\big] = V^{\boldsymbol{\mathcal{E}}}(\pi^S). \tag{17}$$

Furthermore, for

$$\pi^*(a \mid x) \coloneqq \mathbb{1}\big[a = \operatorname*{arg\,max}_{a' \in \mathcal{A}} \mathbb{E}^{\pi_{a'}}\big[R \mid X^{S_{IB}} = x\big]\big]$$

we have

$$\begin{aligned} V^{\boldsymbol{\mathcal{E}}}(\pi^*) &= \inf_{e \in \boldsymbol{\mathcal{E}}} \mathbb{E}^{\pi^*, e}\big[R\big] \\ &= \inf_{e \in \boldsymbol{\mathcal{E}}} \mathbb{E}^e\big[\mathbb{E}^{\pi^*, e}[R \mid X^{S_{IB}}]\big] \\ &= \inf_{e \in \boldsymbol{\mathcal{E}}} \mathbb{E}^e\big[\max_{a \in \mathcal{A}} \mathbb{E}^{\pi_a, e}[R \mid X^{S_{IB}}]\big] \end{aligned}$$

And because the inequality (17) holds for all $S \in \mathbf{S}_{\text{inv}}$ and all $\pi^S \in \Pi^S$, this implies

$$\pi^* \in \operatorname*{arg\,max}_{\pi \in \Pi_{\text{inv}}} V^{\boldsymbol{\mathcal{E}}}(\pi).$$

By taking the infimum over $\boldsymbol{\mathcal{E}}^{\text{obs}}$ (instead of $\boldsymbol{\mathcal{E}}$) before (17), we also get that $\pi^*(a \mid x) \in \arg\max_{\pi \in \Pi_{\text{inv}}} V^{\boldsymbol{\mathcal{E}}^{\text{obs}}}(\pi)$. This completes the proof of Proposition 2. $\square$

D.5. Proof of Theorem 1

Proof. To simplify notation we assume $X^k \in \mathrm{DE}(X^k)$. We begin by defining the stable set

$$S_{NI} \coloneqq \{1,\dots,d\}\setminus\{j \in \{1,\dots\} \mid \exists k \in \mathrm{CI} : j \in \mathrm{DE}(X^k)\},$$

where CI are confounded and directly intervened on nodes (i.e., for $k \in \mathrm{CI}$ there exists $\ell \in \{1,\dots,p\}$ such that $e \to X^k \leftarrow U^\ell \to R$). Furthermore, let us denote $\pi_a(a' \mid x) \coloneqq \mathbb{1}\big[a' = a\big]$ as the policy that always selects the action a. Now we provide the following lemmas which are the main parts of our proof.

Lemma 2 (Stable sets and Invariance)**.** *Assume Setting 1 and Assumption 1. If an invariant set exists, it holds that the stable set S_{NI} is an invariant set.*

Lemma 3 (Lower bound on $V^{\boldsymbol{\mathcal{E}}}(\bar{\pi}^*)$)**.** *Assume Setting 1 and Assumption 1. Let $\bar{\pi}^* \in \arg\max_{\pi\in\Pi_{\mathrm{inv}}} V^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(\pi)$. Then it holds that*

$$V^{\boldsymbol{\mathcal{E}}}(\bar{\pi}^*) \geq \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}}}\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a}\left[R \mid X^{S_{NI}}\right]\right].$$

Lemma 4 (Upper bound on $V^{\boldsymbol{\mathcal{E}}}(\underline{\pi}^*)$)**.** *Assume Setting 1 and Assumptions 1 and 2. Let $\pi \in \Pi\setminus\Pi_{\mathrm{inv}}$ be an arbitrary non-invariant policy. Then it holds that it holds that*

$$V^{\boldsymbol{\mathcal{E}}}(\pi) \leq \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}}}\left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a}\left[R \mid X^{S_{NI}}\right]\right].$$

Then, based on these Lemmas, the proof proceeds as follows: Let $\bar{\pi}^* \in \arg\max_{\pi\in\Pi_{\mathrm{inv}}} V^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(\pi)$ and $\underline{\pi}^* \in \arg\max_{\pi\in\Pi\setminus\Pi_{\mathrm{inv}}} V^{\boldsymbol{\mathcal{E}}^{\mathrm{obs}}}(\pi)$, then by Lemma 3 and Lemma 4, we have

$$V^{\boldsymbol{\mathcal{E}}}(\bar{\pi}^*) \geq V^{\boldsymbol{\mathcal{E}}}(\underline{\pi}^*),$$

which concludes the proof of Theorem 1. □

D.5.1. Proof of Lemma 2

Proof. Recall our definition of the stable set

$$S_{NI} \coloneqq \{1,\dots,d\}\setminus\{j \in \{1,\dots\} \mid \exists k \in \mathrm{CI} : j \in \mathrm{DE}(X^k)\},$$

where CI are confounded and directly intervened on nodes (i.e., for $k \in \mathrm{CI}$ there exists $\ell \in \{1,\dots,p\}$ such that $e \to X^k \leftarrow U^\ell \to R$).

Claim 1. *If an invariant set exists, it holds that $S_R \subseteq S_{NI}$.*

Proof. We prove the claim by contrapositive. Assume that there exists $j \in S_R$ such that $j \notin S_{NI}$. This implies that there exist $k \in \{1,\dots,d\}$ and $\ell \in \{1,\dots,p\}$ such that $e \to X^k \leftarrow U^\ell \to R$ in $\mathcal{G}^\pi$ and $j \in \mathrm{DE}(X^k)$. Now, we will show that there is no invariant set. Let $S \subseteq \{1,\dots,d\}$ be an arbitrary set. There are two possibilities,

(a) $j \in S$: Since $j \in \mathrm{DE}(X^k)$, the path $e \to X^k \leftarrow U^\ell \to R$ is d-connected given S, and therefore $R \not\perp\!\!\!\perp_{\mathcal{G}^\pi} e \mid X^S$. By Assumption 1, this directly implies that S is not an invariant set.

(b) $j \notin S$: Since $j \in S_R$ is a parent of R, S can only d-separates e from R if $\mathrm{PA}(X^j) \in S$. However, because $j \in \mathrm{DE}(X^k)$, the path $e \to X^k \leftarrow U^\ell \to R$ is d-connected given $\mathrm{PA}(X^j)$. We thus have that $R \not\perp\!\!\!\perp_{\mathcal{G}^\pi} e \mid X^S$, which by Assumption 1 implies that S is not an invariant set.

The two possibilities conclude that there is no invariant set which completes the proof (by contrapositive). □

Now, we prove the main result. By Lemma 1, it suffices to show that any path ρ from e to R is d-separated given S_{NI}. First, by Claim 1, we have that $S_R \subseteq S_{NI}$. Therefore, if the path ρ enters R through a parent it will be d-separated given $X^{S_{NI}}$. So assume ρ enters R through a hidden variable U^ℓ. Then, there are two cases:

(i) $\rho: \quad e \to \cdots X^k \to X^j \leftarrow U^\ell \to R$

(ii) $\rho: \quad e \to \cdots X^k \leftarrow X^j \leftarrow U^\ell \to R$

In case (i), since X^j is a collider, ρ can only be d-connected given $X^{\mathrm{SB}_e(R)}$ if $k \notin S_{NI}$ and $\mathrm{DE}(X^j) \cap S_{NI} \neq \varnothing$. Consider two cases:

(a) $k \in S_{NI}$: This directly implies that ρ is d-separated given $X^{S_{NI}}$.

(b) $k \notin S_{NI}$: By definition of S_{NI} it holds that $\mathrm{DE}(X^k) \cap S_{NI} = \varnothing$. Hence, ρ is d-separated given $X^{S_{NI}}$.

We have therefore shown that in case (i) the path ρ is d-separated given $X^{S_{NI}}$. Next, consider the case (ii) and let X^r be the collider closest to X^j on ρ. The path ρ can only be d-connected given $X^{S_{NI}}$ if $j \notin S_{NI}$ and $\mathrm{DE}(X^r) \cap S_{NI} \neq \varnothing$. Consider the two cases:

(a) $j \in S_{NI}$: This directly implies that ρ is d-separated given $X^{S_{NI}}$.

(b) $j \notin S_{NI}$: Since $\mathrm{DE}(X^r) \subseteq \mathrm{DE}(X^j)$, this implies that $\mathrm{DE}(X^r) \cap S_{NI} = \varnothing$. Hence, the path ρ is d-separated given $X^{S_{NI}}$.

From the above two cases, we have shown that the path ρ is d-separated by $X^{S_{NI}}$, which proves that S_{NI} is an invariant set.

□

D.5.2. PROOF OF LEMMA 3

Proof. Let Π_{inv} be the set of all invariant policies defined in (4). First, from Proposition 2, we have that

$$\forall \pi_{\text{inv}} \in \Pi_{\text{inv}} : V^{\boldsymbol{\mathcal{E}}}(\bar{\pi}^*) \geq V^{\boldsymbol{\mathcal{E}}}(\pi_{\text{inv}}).$$

Because S_{NI} is an invariant set (by Lemma 2), the above inequality directly implies

$$V^{\boldsymbol{\mathcal{E}}}(\bar{\pi}^*) \geq \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}}} \left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a}\left[R \mid X^{S_{NI}}\right]\right],$$

which concludes the proof of Lemma 3. □

D.5.3. PROOF OF LEMMA 4

Proof. Recall our definition of the stable set

$$S_{NI} \coloneqq \{1,\dots,d\}\setminus\{j\in\{1,\dots\} \mid \exists k\in \text{CI} : j\in \text{DE}(X^k)\},$$

where CI are confounded and directly intervened on nodes (i.e., for $k \in \text{CI}$ there exists $\ell \in \{1,\dots,p\}$ such that $e \to X^k \leftarrow U^\ell \to R$). To begin with, we provide the following claim as part of the main proof.

Claim 2. *If $e \in \boldsymbol{\mathcal{E}}$ is a confounding removing environment it holds that $\forall j \in \{1,\dots,d\} \setminus S_{NI} : R \perp\!\!\!\perp_{\mathcal{G}^\pi} X^j \mid X^{S_{NI}}$.*

Proof. Let $j \in \{1,\dots,x\} \setminus S_{NI}$, then by definition there exist $k \in \{1,\dots,d\}$ and $\ell \in \{1,\dots,p\}$ such that $e \to X^k \leftarrow U^\ell \to R$ in $\mathcal{G}^\pi$ and $j \in \text{DE}(X^k)$. Since X^k is a collider this implies that for all $S \subseteq \{1,\dots,d\}$ it holds that $R \not\perp\!\!\!\perp_{\mathcal{G}^\pi} e \mid X^{S\cup\{j\}}$. Therefore, by the definition of a confounding removing environment we know that $X^j \perp\!\!\!\perp_{\mathcal{G}^{\pi,e}} U$. Additionally, by Claim 1, we also know that $S_R \subseteq S_{NI}$. Thus, R and X^j can only be d-connected given $X^{S_{NI}}$ if there exists $u \in \{1,\dots,d\}$ and $v \in \{1,\dots,p\}$ such that $X^j \to X^u \leftarrow U^v \to R$ and $S_{NI} \cap \text{DE}(X^u) \neq \varnothing$. However, since $u \in \text{DE}(X^j)$, this implies that $u \in \text{DE}(X^k)$. Then, by the definition of S_{NI} we have that $S_{NI} \cap \text{DE}(X^u) \neq \varnothing$ which then concludes that $R \perp\!\!\!\perp_{\mathcal{G}^\pi} X^j \mid X^{S_{NI}}$ and completes the proof. □

Now, we are ready to prove the main result. Let $\pi \in \Pi \setminus \Pi_{\text{inv}}$ be an arbitrary non-invariant policy, and X^S be the variables that the policy π depends on, i.e., there exists $\pi^S : \boldsymbol{\mathcal{X}}^S \to \Delta(\mathcal{A})$ such that for all $x \in \boldsymbol{\mathcal{X}}$ it holds that $\pi(\cdot|x) = \pi^S(\cdot|x^S)$. We have

$$\begin{aligned} &V^{\boldsymbol{\mathcal{E}}}(\pi)\\ &= \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^{\pi,e}\left[R\right], \end{aligned}$$

by the tower property of conditional expectation,

$$\begin{aligned} &= \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}},X^{S\setminus S_{NI}}} \left[\mathbb{E}^{\pi,e}\left[R \mid X^{S_{NI}}, X^{S\setminus S_{NI}}\right]\right]\\ &= \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}},X^{S\setminus S_{NI}}} \Bigg[\int \mathbb{E}^{\pi_a,e}\left[R \mid X^{S_{NI}}, X^{S\setminus S_{NI}}\right] \pi(a|X^S)\,\mu(\mathrm{d}a)\Bigg]. \end{aligned}$$

Now, we use Assumption 2. For each $e \in \boldsymbol{\mathcal{E}}$ we choose a confounding removing environment $f(e)$ such that $\mathbb{P}^{\pi,f(e)}_X = \mathbb{P}^{\pi,e}_X$. We then have

$$\begin{aligned} &V^{\boldsymbol{\mathcal{E}}}(\pi)\\ &\leq \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}},X^{S\setminus S_{NI}}} \Bigg[\int \mathbb{E}^{\pi_a,f(e)}\left[R \mid X^{S_{NI}}, X^{S\setminus S_{NI}}\right] \pi(a|X^S)\,\mu(\mathrm{d}a)\Bigg]. \end{aligned}$$

Next, we use the result from Claim 2. We have that $\forall j \in \{1,\dots,d\} \setminus S_{NI} : R \perp\!\!\!\perp_{\mathcal{G}^\pi} X^j \mid X^{S_{NI}}$. Then, by the Markov property, we get

$$\begin{aligned} &V^{\boldsymbol{\mathcal{E}}}(\pi)\\ &\leq \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}},X^{S\setminus S_{NI}}} \Bigg[\int \mathbb{E}^{\pi_a,f(e)}\left[R \mid X^{S_{NI}}\right] \pi(a|X^S)\,\mu(\mathrm{d}a)\Bigg], \end{aligned}$$

we can then omit $f(e)$ since S_{NI} is an invariant set (Lemma 2),

$$\begin{aligned} &= \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}},X^{S\setminus S_{NI}}} \Bigg[\int \mathbb{E}^{\pi_a}\left[R \mid X^{S_{NI}}\right] \pi(a|X^S)\,\mu(\mathrm{d}a)\Bigg]\\ &= \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}}} \Bigg[\int \mathbb{E}^{\pi_a}\left[R \mid X^{S_{NI}}\right] \mathbb{E}^e_{X^{S\setminus S_{NI}}}\left[\pi(a|X^S)\right]\mu(\mathrm{d}a)\Bigg], \end{aligned}$$

letting $\tilde{\pi}(a \mid X^{S_{NI}}) \coloneqq \mathbb{E}_{X^{S\setminus S_{NI}}}[\pi(a|X^S)]$,

$$\begin{aligned} &= \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}}} \left[\int \mathbb{E}^{\pi_a}\left[R \mid X^{S_{NI}}\right]\tilde{\pi}(a \mid X^{S_{NI}})\,\mu(\mathrm{d}a)\right]\\ &\leq \inf_{e\in\boldsymbol{\mathcal{E}}} \mathbb{E}^e_{X^{S_{NI}}} \left[\max_{a\in\mathcal{A}} \mathbb{E}^{\pi_a}\left[R \mid X^{S_{NI}}\right]\right]. \end{aligned}$$

□

D.6. Proof of Theorem 3

Proof. We only show that our setting with environments can be cast as that in (Thams et al., 2021), which has no reference to environments.

First we randomly permute the rows of each dataset D^e to obtain a set $\tilde{D}^e$. Then construct an auxiliary data set $D^{\mathcal{E}}$, where each observation D_i of $D^{\mathcal{E}}$ is the concatenation of the i'th observation (after permutation) from all environments, $D_i := (\tilde{D}_i^{e_1}, \ldots, \tilde{D}_i^{e_L})$.

We can now apply the resampling methodology from (Thams et al., 2021) to draw a sequence $(D_{i_1}, \ldots, D_{i_m})$ with weights given by (8) where

$$r(D_i) := \frac{\pi(A_i^{e_1} \mid X_i^{e_1})}{\pi^0(A_i^{e_1} \mid X_i^{e_1})} \cdots \frac{\pi(A_i^{e_L} \mid X_i^{e_L})}{\pi^0(A_i^{e_L} \mid X_i^{e_L})}$$

Because the observations are independent, both within and between environments, the probability of the sequence $(D_{i_1}, \ldots, D_{i_m}) = ((D_{i_1}^{e_1}, \ldots, D_{i_1}^{e_L}), \ldots, (D_{i_m}^{e_1}, \ldots, D_{i_m}^{e_L}))$ is equal to the probability of drawing first m observations from e_1, $(D_{i_1}^{e_1}, \ldots, D_{i_m}^{e_1})$, and then m from e_2 etc. $\square$

E. Connection to Random Environments

It is possible to define multi-environment contextual bandits using random environments.

Setting 3 (Random Environment Contextual Bandits)**.** *Let $X = (X^1, \ldots, X^d) \in \boldsymbol{\mathcal{X}} = \mathcal{X}^1 \times \ldots \times \mathcal{X}^d$, $U = (U^1, \ldots, U^p) \in \boldsymbol{\mathcal{U}} = \mathcal{U}^1 \times \ldots \times \mathcal{U}^p$, $A \in \mathcal{A} = \{a^1, \ldots, a^k\}$, $R \in \mathbb{R}$, $E \in \boldsymbol{\mathcal{E}}$. For any $\pi \in \{\pi : \boldsymbol{\mathcal{X}} \to \Delta(\mathcal{A})\}$, let g_π denote the function that ensures, for all $x \in \boldsymbol{\mathcal{X}}$, $g_\pi(x, \epsilon_A)$ equals $\pi(x)$ in distribution for a uniformly distributed ϵ_A. Now, consider functions s, h, and f, a factorizing distribution $\mathbb{P}_\epsilon = \mathbb{P}_{\epsilon_E} \times \mathbb{P}_{\epsilon_U} \times \mathbb{P}_{\epsilon_X} \times \mathbb{P}_{\epsilon_A} \times \mathbb{P}_{\epsilon_R}$ whose ϵ_A component is uniform, and a structural causal model $\mathcal{S}(\pi)$ given by*

$$\mathcal{S}(\pi): \quad \begin{cases} E := \epsilon_E \\ U := s(X, \epsilon_U) \\ X := h(X, U, E, \epsilon_X) \\ A := g_\pi(X, \epsilon_A) \\ R := f(X, U, A, \epsilon_R). \end{cases}$$

Assume further that for all π, the SCM induces a unique distribution over (E, X, U, A, R), which we denote by $\mathbb{P}^\pi$. The structure of the SCM $\mathcal{S}(\pi, e)$ can be also visualized by a graph G^π which is constructed in a similar way to the graph in Setting 1, except that the environment becomes one of the variable nodes in this graph.

Remark 1. Setting 3 is a special case of Setting 1 in the following sense: Assume, starting from Setting 3, for all $i \in \{1, \ldots, n\}$ that $(X_i, U_i, A_i, R_i, E_i)$, are independent and distributed according to $\mathbb{P}_{X,U,A,R,E}^{\pi_i}$. Then, defining $h_e(\cdot, \cdot) := h(\cdot, e, \cdot)$, we have that, for all $i \in \{1, \ldots, n\}$, (X_i, U_i, A_i, R_i), are independent and distributed according to $\mathbb{P}_{X,U,A,R}^{\pi_i, E_i}$, using Setting 1.

F. Simulation Details

F.1. Data Generating Process

We generate data from the following SCM $\mathcal{S}(\pi, e)$:

$$\begin{aligned} U &:= \mu_U + \epsilon_U \\ X^2 &:= \mu_{X^2}^e + \epsilon_{X^2} \\ X^1 &:= \alpha^e U + \epsilon_{X^1} \\ A \mid X^1, X^2 &\sim \pi(A \mid X^1, X^2) \\ R &:= \beta_{A,1} X^2 + \beta_{A,2} U + \epsilon_R, \end{aligned}$$

where $\epsilon_U, \epsilon_{X^2}, \epsilon_{X^1}, \epsilon_R \sim \mathcal{N}(0, 1)$, A takes values in the space $\{a_1, \ldots, a_L\}$. In our experiments, we consider 3 possible actions ($L = 3$) and randomly draw the parameters $\mu_U, \beta_{a_1,1}, \ldots, \beta_{a_3,1}, \beta_{a_1,2}, \ldots, \beta_{a_3,2}$ from $\mathcal{N}(0, 1)$, while the environment-specific parameters $\mu_{X^2}^e, \alpha^e$ are drawn from $\mathcal{N}(0, 4)$. These parameters are then fixed across all experiment runs.

F.2. Policy Optimization Algorithm

In Section 5.1, we consider the policy of the form

$$\pi^S(a \mid x) := \mathbb{1}\big[a = \arg\max_{a' \in \mathcal{A}} \hat{f}^S(x, a')\big],$$

where $\hat{f}^S$ is an estimator of the conditional mean $\mathbb{E}^{\pi_a}(R \mid X^S)$ and π_a is the policy that always selects a.

Given observations $\{(X_i^1, X_i^2, A_i, R_i, e_i)\}_{i=1}^n$ from an initial policy π_0, we find an importance-weighted regressor

$$\hat{f}^S := \arg\min_{f \in \mathcal{F}^S} \sum_{i=1}^n \frac{1}{\pi_0(A_i \mid X_i)} (f(A_i, X_i^S) - R_i)^2, \quad (18)$$

where $\mathcal{F}^S = \{\boldsymbol{\mathcal{X}}^S \times \mathcal{A} \to \mathbb{R}\}$ is a class of functions. In our experiment, we consider a linear function class. Solving the objective (18) then reduces to fitting a weighted linear regression.