Regularizing Adversarial Imitation Learning Using Causal Invariance

Ivan Ovinnikov [1] **Joachim M. Buhmann** [1]

Abstract

Imitation learning methods are used to infer a policy in a Markov decision process from a dataset of expert demonstrations by minimizing a divergence measure between the empirical state occupancy measures of the expert and the policy. The guiding signal to the policy is provided by the discriminator used as part of an adversarial optimization procedure. We observe that this model is prone to absorbing spurious correlations present in the expert data. To alleviate this issue, we propose to use causal invariance as a regularization principle for adversarial training of these models. The regularization objective is applicable in a straightforward manner to existing adversarial imitation frameworks. We demonstrate the efficacy of the regularized formulation in an illustrative two-dimensional setting as well as a number of high-dimensional robot locomotion benchmark tasks.

1. Introduction

The invariant causal prediction principle (Peters et al., 2015) has gained a lot of attention in the recent years. Contemporary methods such as (Arjovsky et al., 2019; Chang et al., 2020; Krueger et al., 2021) propose a representation learning scheme for supervised learning problems which aim to eliminate features which are spuriously correlated with the label. Various instantiations of this principle obtain asymptotically stable label conditionals across interventional settings of the data generating process. The canonical example of deep learning models absorbing such spurious features is the classification of cows and camels. In this example, the model learns the feature encoding of the background as a form of shortcut for classifying the more complex geometry of animal shapes, exploiting a selection bias in the dataset. The model subsequently fails on a test set of images with permuted backgrounds. Analogously, in reinforcement learning, it is desirable to avoid behaviours which would exploit such features. This is particularly relevant when learning from demonstrations, i.e. in the imitation learning setting.

Modern imitation learning methods (Ho & Ermon, 2016) aim to minimize a discrepancy measure between the a finite dataset of expert demonstrations and the trajectories induced by the policy trying to mimic the expert. The discrepancy measure is typically an instance of the family of φ-divergences (Csiszár, 1972) or integral probability metrics (e.g. Wasserstein distance). In both cases, the variational formulation of the density matching problem is chosen for computational purposes whitch has been shown to have strong links to binary classification (Nguyen et al., 2009; Sriperumbudur et al., 2009), a fact widely used in generative adversarial network (GAN) and adversarial imitation methods.

In this work, we observe that the binary classifier used as discriminator in the adversarial optimization scheme is prone to exploiting the spurious correlations present in the mixture of policy and expert trajectory data. This has multiple far-reaching implications for the resulting training procedure. For instance, this could lead to undesired behaviours, similar to the ones associated with reward hacking (Skalse et al., 2022). The exploitation of spurious correlations by a model typically leads to higher empirical performance at training time but will fail at test time. In the context of adversarial training, an overly confident discriminator is known to impede meaningful generator training due to a stale training signal. This issue is typically remedied by regularizing the discriminator in various ways (Gulrajani et al., 2017; Peng et al., 2018). The problem is exacerbated by the fact that the policy will try to optimize the expected density ratio based on spurious features of the discriminator, further contributing to the covariate shift.

To alleviate this issue, we propose to regularize the discriminator using the invariant risk minimization principle (Arjovsky et al., 2019), more specifically, the IRMv1 objective. The application of this regularization technique requires mild assumptions on the problem setting, which are often satisfied in practice, and is easy to implement. To validate our method, we perform an empirical study of the

[*]Equal contribution [1]Department of Computer Science (ETH Zürich), ETH Zürich, Zürich, Switzerland. Correspondence to: Ivan Ovinnikov <ivan.ovinnikov@inf.ethz.ch>.

Published at the ICML 2023 Workshop on Spurious Correlations, Invariance, and Stability. Honolulu, Hawaii, USA.

algorithm performance in both a low-dimensional navigation setting as well as on a number of benchmark tasks from the MuJoCo suite. We observe a consistent improvement in both settings when using the regularized version of common adversarial imitation learning algorithms.

2. Related work

Invariance and causality in reinforcement learning The concept of invariance has been used in a number of works in the reinforcement learning domain. Invariant causal prediction has been utilized in (Zhang et al., 2020) to learn model invariant state abstractions in a multiple MDP setting with a shared latent space. Invariant policy optimization (Sonar et al., 2021) uses the IRM games (Ahuja et al., 2020) formulation to learn policies invariant to certain domain variations. de Haan et al. (2019) tackle the problem of causal confusion in imitation learning by making use of causal structure of demonstrations. The issue of discriminator overfitting to task-irrelevant visual features is addressed in (Zolna et al., 2021). Another example of using causal invariance is presented in (Bica et al., 2021). In contrast to the methods outlined above, our method specifically addresses the issues with spurious correlations *during* the process of adversarial training, which lead to discriminator degeneration.

3. Problem setting

We start by introducing the necessary notation and formalism to describe the problem setting.

MDP We consider environments modelled by a *Markov decision process* $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{T}, \mu, R)$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{T}$ is the family of transition distributions on $\mathcal{S}$ indexed by $\mathcal{S} \times \mathcal{A}$ with $p(s'|s,a)$ describing the probability of transitioning to state s' when taking action a in state s, μ is the initial state distribution, and $R : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the reward function. A *policy* π is a map from states $s \in \mathcal{S}$ to distributions $\pi(\cdot|s)$ over actions, with $\pi(a|s)$ being the probability of taking action a in state s. We denote by $\rho_E = \sum_{s_i \in \mathcal{D}_E} \delta_i(s_i)$ the empirical state occupancy measure of the expert based on a dataset of expert trajectories $\mathcal{D}_E = \{\tau_i\}_{i \leq K}$ where $\tau_i = (s^{(i)}_{1:T}, a^{(i)}_{1:T})$ is a sequence of states and actions of expert i of length T. $\rho_\pi = \sum_{t \leq T} P^\pi_\mu(S_t = s, A_t = a)$ denotes the state occupancy measure induced by the policy π over a finite horizon T for initial measure μ.

Imitation learning methods aim to estimate a policy π_θ parameterized by weights θ , which mimics the expert. To achieve this goal, a distance or divergence measure between the empirical state occupancy measure of the expert ρ_E and the induced state occupancy measure of the policy ρ_π is minimized. More specifically, the divergence measure is typically an instance of the class of φ-divergences (Csiszár, 1972), where the choice the φ-function corresponds to commonly used methods such as GAIL ((Ho & Ermon, 2016)), AIRL ((Fu et al., 2017)) or f-IRL ((Ni et al., 2021)). The adversarial imitation learning (AIL) objective is formulated as follows:

$$\mathcal{L}_{AIL} = \min_\theta \max_\psi \mathbb{E}_{\rho_E}[\log D_\psi(s,a,s')] + \mathbb{E}_{\rho_{\pi_\theta}}[\log(1 - D_\psi(s,a,s'))] - \lambda \mathcal{H}(\pi_\theta)$$

where $D_\psi(s,a,s')$ is the discriminator parametrized by a neural network with parameters ψ, π_θ is the student policy and $\mathcal{H}(\pi_\theta)$ the entropy regularization term. [1]

Invariant causal prediction The principle of invariant causal prediction (Peters et al., 2015; Heinze-Deml et al., 2017) stipulates that for provable out-of-distribution (OOD) generalization in linear regression tasks, the regression coefficients must be stable across interventional settings of the data generating process, indexed by $e \in \mathcal{E}$ where $\mathcal{E}$ denotes the set of datasets sampled from the data generating process. The authors of (Arjovsky et al., 2019) extend this to nonlinear features and introduce a tractable approximation of the bi-level optimization required to identify invariant features Φ. The derived gradient norm penalty $\mathbb{D}(w, \Phi, e) = ||\nabla_{w|w=1.0} \mathcal{L}^e(w \circ \Phi)||^2$ quantifies the violation of the normal equations to measure the optimality of a fixed linear classifier ($w = 1.0$) at each setting e. This leads to the following regularized formulation of the empirical risk minimization (ERM) problem where $\mathcal{E}_{tr} \subseteq \mathcal{E}$ is the set of training environments:

$$\min_{\Phi: \mathcal{X} \to \mathcal{Y}} \sum_{e \in \mathcal{E}_{tr}} \mathcal{L}^e(\Phi) + \lambda ||\nabla_{w|w=1.0} \mathcal{L}^e(w \circ \Phi)||^2 \quad \text{(IRMv1)}$$

In the following section, we will address why this regularization is beneficial for disciminator fitting in adversarial training procedures.

4. Spurious correlations in adversarial imitation learning

We will now describe the mechanisms by which spurious correlations lead to issues in the process of adversarial training. At every adversarial optimization round, a number of discriminator gradient updates is performed. In particular, the discriminator at round k is updated using the concatenated transition samples from the expert dataset and the policy buffer, $\mathcal{D}_k = (\mathcal{D}_E, \mathcal{D}_\pi)$. Let us assume a decomposition of the feature space $\mathbf{x} \in \mathcal{X}$ into two subsets, $\mathbf{x}_C$ and $\mathbf{x}_{NC}$, which denote the causal

[1] In the case of AIRL, ψ denotes the joint set of structured discriminator parameters of functions g_ξ and h_ϕ

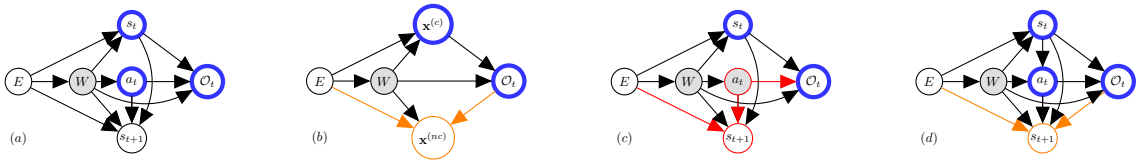

Figure 1: (a) Probabilistic graphical model of a transition under influence of the index variable E and latent variable W. The stable conditional is highlighted in blue. (b) General setting where $\mathcal{O}_t$ depends on causal $\mathbf{x}^{(c)}$ and non-causal $\mathbf{x}^{(nc)}$ features of the transition. (c) Spurious correlations assuming wrong edge orientation $\mathcal{O}_t \rightarrow s_{t+1}$. (d) Spurious correlations assuming state-only formulation.

and spurious[2] features respectively. By definition, spurious features do not generalize outside of the training set. The discriminator parametrizes the density ratio $D(\mathbf{x}) = D(\mathbf{x}_C, \mathbf{x}_{NC}) = \hat{\rho}_E(\mathbf{x}_C, \mathbf{x}_{NC})/\hat{\rho}_\pi(\mathbf{x}_C, \mathbf{x}_{NC})$, where $\hat{\rho}_E(\mathbf{x}_C, \mathbf{x}_{NC})$ and $\hat{\rho}_\pi(\mathbf{x}_C, \mathbf{x}_{NC})$ denote the density estimators of the expert and policy respectively. Suppose in the cow-camel classification example, the camera operator is given agency to move the camera to frame different parts of the scene. The guiding signal is provided by the negative log likelihood ratio of features present in subsets of the cow and camel images, meaning deviations from unity ratio are penalized. If the ratio is easier to estimate using the $\mathbf{x}_{NC}$ variable of the joint distribution $\hat{p}(\mathbf{x}_C, \mathbf{x}_{NC})$, the camera might end up focusing on parts of the scene which do not contain objects of interest. Analogously, in the imitation learning case, the policy occupancy measure will converge to a part of the state space, which might not describe meaningful behaviours.

Spurious correlations in model of transitions Figure 1 describes our setting from a probabilistic graphical model point of view. We consider various settings in which a non-causal information path corresponding to spurious correlations is formed in the structural causal model (SCM) of transitions (Fig. 1a). Fig. 1b illustrates the most general setting where an arbitrary transition input (s, a, s') is partitioned into the causal transition feature components $\mathbf{x}^{(c)}$ and $\mathbf{x}^{(nc)} : (s, a, s') = (\mathbf{x}^{(c)}, \mathbf{x}^{(nc)})$, whereby conditioning on the $\mathbf{x}^{(nc)}$ collider introduces a spurious correlation path. In Fig. 1c we can observe the scenario where we do not condition the discriminator $D(s)$ on the action. By conditioning on the collider node s_{t+1} and not observing the action node a_t, a path is formed between the setting index E and the optimality variable $\mathcal{O}_t$, resulting in the violation of their conditional independence relationship. A third scenario can be observed in Fig. 1d. This scenario requires the assumption that the orientation of the edge from node $\mathcal{O}_t$ to node s_{t+1} is temporally causal, meaning that the optimality of a state at time t is a causal parent of the next state. In this case, observing the collider node s_{t+1} implies the following conditional independence relationship: $E \not\perp\!\!\!\perp O_t | s_{t+1}$.

[2]Here, spuriousness is defined w.r.t. the output label

Algorithm 1 Causally invariant adversarial imitation learning (CIAIL)

Input: Expert trajectories $\mathcal{D}_E^e$ from settings $e \in \mathcal{E}$
Initialize *actor-critic* π_θ, V_ϑ or *soft actor-critic* $\pi_\theta, Q_\varsigma^{(j)}, V_\vartheta$ and discriminator D_ψ
for $t = 1$ **to** N_{rounds} **do**
  Collect trajectory buffer $\mathcal{D}_\pi = \{\tau_i\}_{i \leq |\mathcal{D}_\pi|}$ by executing the policy π_θ
  Update $D_\psi(s, a)$ via binary logistic regression by maximizing $\mathcal{L}(\psi, e)$ using tuple $\mathcal{D}_t = (\mathcal{D}_E, \mathcal{D}_\pi)$:

$$\mathcal{L}(\psi; e) = \mathcal{L}_{\text{BCE}}(\psi; e) + \lambda ||\nabla_{\omega|\omega=1.0} \mathcal{L}_{\text{BCE}}(\psi; e)||^2$$

  Compute $\log D_\psi(s, a, s')\ \forall (s, a, s') \in \mathcal{D}$
  1. (On-policy CIAIL): Update $(\pi_\theta, V_\vartheta)$ using a constrained policy gradient method (e.g. PPO) using r_ψ as reward
  2. (Off-policy CIAIL): Update $(\pi_\theta, Q_\varsigma^{(j)}, V_\vartheta)$ using an off-policy methods (e.g. SAC) using r_ψ as reward
end for

In order to apply this intuition in the desired context, we must make the following assumption which has implications on the necessary data and training procedure specifics.

Assumption 4.1. *The data samples in the discriminator training tuple at round k $\mathcal{D}_k$ stem from different* settings $e \in \mathcal{E}_{tr}$.

The assumption is motivated by the fact that in the IRMv1 formulation, no explicit environment specification is necessary to perform the optimization and obtain invariant features. For the problem we are considering, this assumption is satisfied in two cases. The first case necessitates a *varied* set of expert demonstrations $\mathcal{D}_E^e$ where e denotes the setting index. This scenario is quite common as the expert demonstrations from multiple sources are typically pooled into one dataset. The second case corresponds to a setting where the policy set of transitions $\mathcal{D}_\pi$ contains transitions gathered over multiple policy optimization episodes. This setting corresponds to off-policy reinforcement learning methods such as Soft Actor-Critic (SAC) (Haarnoja et al., 2018), where the replay buffer contains rollouts of policies from previous

Table 1: Policy rollout results using ground truth reward for 2d navigation environment (`MovePoint`) with varying regularization strength and number of discriminator updates. Expert reference: -23665.025$_{\pm 2264.521}$

$n_{updates}$	irm: $\lambda = 0.01$	irm: $\lambda = 0.1$	irm: $\lambda = 1.0$	irm: $\lambda = 10.0$	erm
GAIL					
1	-24747.54$_{\pm 4386.43}$	-25702.6$_{\pm 3534.82}$	-24732.45$_{\pm 3270.27}$	**-22836.3**$_{\pm 3240.09}$	-26703.01$_{\pm 4607.2}$
2	-27169.45$_{\pm 4005.41}$	-28277.35$_{\pm 4534.81}$	-32413.52$_{\pm 4023.55}$	-27741.34$_{\pm 4692.81}$	**-22320.1**$_{\pm 2181.62}$
5	**-25339.01**$_{\pm 3560.01}$	-28124.89$_{\pm 5341.83}$	-33132.36$_{\pm 6205.38}$	-28396.54$_{\pm 2183.32}$	-30267.07$_{\pm 4701.83}$
10	-36217.57$_{\pm 5066.6}$	-34263.65$_{\pm 6112.41}$	**-29385.81**$_{\pm 3815.97}$	-34222.82$_{\pm 5145.31}$	-33012.71$_{\pm 5759.67}$
AIRL					
1	**-29958.88**$_{\pm 4681.55}$	-34682.4$_{\pm 4199.57}$	-33810.05$_{\pm 5590.15}$	-30720.46$_{\pm 4385.96}$	-30003.58$_{\pm 2758.92}$
2	-42563.59$_{\pm 5112.57}$	**-30877.76**$_{\pm 4797.0}$	-31796.58$_{\pm 4120.83}$	-45376.83$_{\pm 9748.12}$	-34177.21$_{\pm 6388.29}$
5	-34297.02$_{\pm 6927.63}$	-39084.49$_{\pm 7758.24}$	**-33692.61**$_{\pm 6301.28}$	-43255.18$_{\pm 6108.74}$	-42262.5$_{\pm 6552.44}$
10	-33756.08$_{\pm 5623.63}$	-40828.81$_{\pm 6786.65}$	-38408.9$_{\pm 6991.46}$	**-30690.02**$_{\pm 4484.04}$	-35218.11$_{\pm 3468.91}$

optimization rounds.

Algorithm We outline the proposed algorithm in 1. The algorithm introduces two novel aspects to the adversarial imitation learning pipeline. The first is a straightforward application of the IRMv1 penalty to the discriminator binary cross-entropy loss. This can be applied to both the *on-policy* and the *off-policy* formulations of the algorithm. The on-policy formulation utilizes the Proximal Policy Optimization (PPO) (Schulman et al., 2017) algorithm for policy training. The introduction of an off-policy algorithm, Soft Actor-Critic (SAC) (Haarnoja et al., 2018) is the second addition. The use of an off-policy algorithm has previously been explored in (Kostrikov et al., 2018) for the purposes of sample efficiency. In our case, the off-policy formulation is one of the scenarios which satisfies 4.1.

5. Experiments

In order to evaluate our method empirically, we propose to conduct two different experiments. In both experiments, we compare the performance of the proposed regularization applied to two well-established adversarial imitation learning baselines: GAIL (Ho & Ermon, 2016) and AIRL (Fu et al., 2017).

5.1. 2d goal navigation

The first task consists of a simple two-dimensional goal navigation problem (`MoveP-v0`) defined on states given by the concatenation of Cartesian coordinates of the agent and the target and imbued with a discrete action space corresponding to the movement directions. The 10 expert trajectories used in training are obtained by sampling a policy trained on the ground truth reward defined as the Euclidean distance to target. In order to simulate the fact that experts stem from different settings, we introduce an intermediate goal which varies across the experts. Here, we limit our evaluation to the on-policy version of algorithm 1. The regularization coefficient λ is varied in the range $\lambda \in \{0.01, 0.1, 1.0, 10.0\}$ and the number of discriminator updates is varied in the range $n \in \{1, 2, 5, 10\}$. The results are summarized in Table 1. We can observe that applying the causal invariance penalty has a consistent positive effect when evaluating the rollout performance of the policies.

5.2. MuJoCo robot locomotion

The second setting is a subset of MuJoCo robot locomotion tasks. Here, we evaluate both the on-policy and the off-policy formulations of the presented algorithm. In Table 2, we can observe that for both policy learning algorithms, regularizing the discriminator significantly improved the cumulative ground truth reward metric obtained by rolling out the learned policies. In particular, we observe a dramatic improvement for the case where the off-policy algorithm (SAC) is used for policy optimization, which validates our assumption 4.1. Our algorithm also favorably compares to an existing gradient penalty regularization method which is based on the convex combination of inputs (mixup) (Carratino et al., 2020) denoted by the GP suffix in Table 2.

6. Discussion

In this work, we have presented a novel algorithm which introduces a causal invariance regularization objective to adversarial imitation learning. We have observed its efficacy in a number of settings and described scenarios which benefits from its application. Future work includes extending these preliminary results to the image domain and a more in depth comparison to existing regularization techniques some of which have recently been interpreted through a causal lens. While the emprical evaluation seems to indicate a strong benefit of the method, a more thorough theoretical analysis of the distribution shift of the discriminator input would be beneficial. Furthermore, a stronger link between

Table 2: Policy rollout results using ground truth reward for MuJoCo environments. The `GP` suffix corresponds to the gradient penalty regularization with regularization coefficient $\lambda_{GP} = 50.0$ and `IRM` suffix to the IRM regularization with coefficient $\lambda_{IRM} = 50.0$. The algorithms were trained on 10 expert trajectories with the following recorded rollout performance: Ant-v3: $4303.532_{\pm 1553.060}$, HalfCheetah-v3: $9018.685_{\pm 125.446}$, Hopper-v3: $1709.923_{\pm 859.010}$, Walker2d-v3: $3984.531_{\pm 64.259}$

Environment	GAIL-ERM	GAIL-GP	GAIL-IRM	AIRL-ERM	AIRL-GP	AIRL-IRM
SAC						
Ant-v3	$2163.28_{\pm 1835.18}$	$3292.43_{\pm 1365.91}$	$\mathbf{4291.28}_{\pm 1243.44}$	$361.35_{\pm 218.897}$	$9.69_{\pm 3.89}$	$\mathbf{2010.191}_{\pm 2170.729}$
HalfCheetah-v3	$941.69_{\pm 382.93}$	$1983.39_{\pm 382.93}$	$\mathbf{2352.82}_{\pm 733.15}$	$2666.90_{\pm 515.74}$	$2849.03_{\pm 367.74}$	$\mathbf{3450.245}_{\pm 1465.968}$
Hopper-v3	$3079.70_{\pm 951.30}$	$2819.37_{\pm 983.72}$	$\mathbf{3315.53}_{\pm 956.21}$	$3581.49_{\pm 39.04}$	$728.96_{\pm 324.067}$	$\mathbf{3770.767}_{\pm 61.337}$
Walker2d-v3	$3128.51_{\pm 1452.83}$	$3076.22_{\pm 1275.81}$	$\mathbf{3705.04}_{\pm 1068.60}$	$2355.69_{\pm 646.41}$	$1538.38_{\pm 1070.435}$	$\mathbf{4213.480}_{\pm 58.596}$
PPO						
Ant-v3	$18.483_{\pm 12.318}$	$-28.1_{\pm 100.89}$	$\mathbf{26.326}_{\pm 20.265}$	$48.650423_{\pm 9.492}$	$8.43_{\pm 11.92}$	$\mathbf{73.427}_{\pm 21.424}$
HalfCheetah-v3	$2577.173_{\pm 1324.064}$	$\mathbf{4117.92}_{\pm 1214.57}$	$2788.065_{\pm 1015.200}$	$571.990_{\pm 223.634}$	$52.01_{\pm 137.43}$	$\mathbf{976.267}_{\pm 1365.386}$
Hopper-v3	$2859.980_{\pm 1114.946}$	$2708.60_{\pm 976.71}$	$\mathbf{3173.536}_{\pm 923.694}$	$170.987_{\pm 60.603}$	$45.01_{\pm 40.98}$	$\mathbf{1421.980}_{\pm 1523.009}$
Walker2d-v3	$2648.653_{\pm 1128.649}$	$1945.99_{\pm 781.85}$	$\mathbf{3443.572}_{\pm 1032.796}$	$24.773_{\pm 5.793}$	$2.91_{\pm 3.74}$	$\mathbf{1361.887}_{\pm 1642.685}$

spurious correlations and reward hacking behaviours should be established.

References

Ahuja, K., Shanmugam, K., Varshney, K., and Dhurandhar, A. Invariant risk minimization games. *arXiv preprint arXiv:2002.04692*, 2020.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. Invariant risk minimization, 2019.

Bica, I., Jarrett, D., and van der Schaar, M. Invariant causal imitation learning for generalizable policies. *Advances in Neural Information Processing Systems*, 34:3952–3964, 2021.

Carratino, L., Cissé, M., Jenatton, R., and Vert, J.-P. On mixup regularization. *arXiv preprint arXiv:2006.06049*, 2020.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. Invariant rationalization. In *International Conference on Machine Learning*, pp. 1448–1458. PMLR, 2020.

Csiszár, I. A class of measures of informativity of observation channels. *Periodica Mathematica Hungarica*, 2(1-4): 191–213, 1972.

de Haan, P., Jayaraman, D., and Levine, S. Causal confusion in imitation learning. *arXiv preprint arXiv:1905.11979*, 2019.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. Improved training of wasserstein gans. *Advances in neural information processing systems*, 30, 2017.

Haarnoja, T., Zhou, A., Hartikainen, K., Tucker, G., Ha, S., Tan, J., Kumar, V., Zhu, H., Gupta, A., Abbeel, P., et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

Heinze-Deml, C., Peters, J., and Meinshausen, N. Invariant causal prediction for nonlinear models, 2017.

Ho, J. and Ermon, S. Generative adversarial imitation learning. *arXiv preprint arXiv:1606.03476*, 2016.

Kostrikov, I., Agrawal, K. K., Dwibedi, D., Levine, S., and Tompson, J. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *arXiv preprint arXiv:1809.02925*, 2018.

Krueger, D., Caballero, E., Jacobsen, J.-H., Zhang, A., Binas, J., Zhang, D., Le Priol, R., and Courville, A. Out-of-distribution generalization via risk extrapolation (rex). In *International Conference on Machine Learning*, pp. 5815–5826. PMLR, 2021.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. On surrogate loss functions and f-divergences. 2009.

Ni, T., Sikchi, H., Wang, Y., Gupta, T., Lee, L., and Eysenbach, B. f-irl: Inverse reinforcement learning via state marginal matching. In *Conference on Robot Learning*, pp. 529–551. PMLR, 2021.

Peng, X. B., Kanazawa, A., Toyer, S., Abbeel, P., and Levine, S. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*, 2018.

Peters, J., Bühlmann, P., and Meinshausen, N. Causal inference using invariant prediction: identification and confidence intervals. *arXiv preprint arXiv:1501.01332*, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Skalse, J., Howe, N. H. R., Krasheninnikov, D., and Krueger, D. Defining and characterizing reward hacking, 2022.

Sonar, A., Pacelli, V., and Majumdar, A. Invariant policy optimization: Towards stronger generalization in reinforcement learning. In *Learning for Dynamics and Control*, pp. 21–33. PMLR, 2021.

Sriperumbudur, B. K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G. R. On integral probability metrics,\phi-divergences and binary classification. *arXiv preprint arXiv:0901.2698*, 2009.

Zhang, A., Lyle, C., Sodhani, S., Filos, A., Kwiatkowska, M., Pineau, J., Gal, Y., and Precup, D. Invariant causal prediction for block mdps. In *International Conference on Machine Learning*, pp. 11214–11224. PMLR, 2020.

Zolna, K., Reed, S., Novikov, A., Colmenarejo, S. G., Budden, D., Cabi, S., Denil, M., de Freitas, N., and Wang, Z. Task-relevant adversarial imitation learning. In *Conference on Robot Learning*, pp. 247–263. PMLR, 2021.