

Imitation Learning as f -Divergence Minimization

Liyiming Ke¹, Sanjiban Choudhury¹, Matt Barnes¹, Wen Sun², Gilwoo Lee¹,
and Siddhartha Srinivasa¹

¹ Paul G. Allen School of Computer Science & Engineering, University of
Washington. Seattle WA 98105, USA,

{kayke,sanjibac,mbarnes,gilwoo,siddh}@cs.washington.edu,

² The Robotics Institute, Carnegie Mellon University, Pittsburgh PA 15213, USA,
wensun@andrew.cmu.edu

Abstract. We address the problem of imitation learning with multi-modal demonstrations. Instead of attempting to learn all modes, we argue that in many tasks it is sufficient to imitate any one of them. We show that the state-of-the-art methods such as GAIL and behavior cloning, due to their choice of loss function, often incorrectly interpolate between such modes. Our key insight is to minimize the right divergence between the learner and the expert state-action distributions, namely the reverse KL divergence or I-projection. We propose a general imitation learning framework for estimating and minimizing any f -Divergence. By plugging in different divergences, we are able to recover existing algorithms such as Behavior Cloning (Kullback-Leibler), GAIL (Jensen Shannon) and DAGGER (Total Variation). Empirical results show that our approximate I-projection technique is able to imitate multi-modal behaviors more reliably than GAIL and behavior cloning.

Keywords: machine learning, imitation learning, probabilistic reasoning

1 Introduction

We study the problem of imitation learning from demonstrations that have *multiple modes*. This is often the case for tasks with multiple, diverse near-optimal solutions. Here the expert has no clear preference between different choices (e.g. navigating left or right around obstacles [1]). Imperfect human-robot interface also lead to variability in inputs (e.g. kinesthetic demonstrations with robot arms [2]). Experts may also vary in skill, preferences and other latent factors. We argue that in many such settings, it suffices to learn a single mode of the expert demonstrations to solve the task. How do state-of-the-art imitation learning approaches fare when presented with multi-modal inputs?

Consider the example of imitating a racecar driver navigating around an obstacle. The expert sometimes steers left, other times steers right. What happens if we apply behavior cloning [3] on this data? The learner policy (a Gaussian with fixed variance) interpolates between the modes and drives into the obstacle.

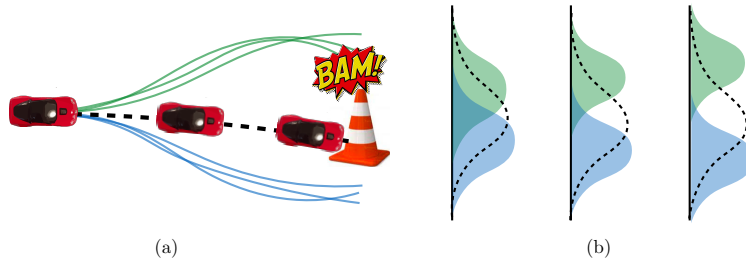


Fig. 1: Behavior cloning fails with multi-modal demonstrations. Experts go left or right around obstacle. Learner interpolates between modes and crashes into obstacle.

Interestingly, this oddity is not restricted to behavior cloning. [4] show that a more sophisticated approach, GAIL [5], also exhibits a similar trend. Their proposed solution, InfoGAIL [4], tries to recover all the latent modes and learn a policy for each one. For demonstrations with several modes, recovering all such policies will be prohibitively slow to converge.

Our key insight is to view imitation learning algorithms as minimizing divergence between the expert and the learner trajectory distributions. Specifically, we examine the family of f -divergences. Since they cannot be minimized exactly, we adopt estimators from [6]. We show that behavior cloning minimizes the Kullback-Leibler (KL) divergence (M-projection), GAIL minimizes the Jensen-Shannon (JS) divergence and DAGGER minimizes the Total Variation (TV). Since both JS and KL divergence exhibit a *mode-covering* behavior, they end up interpolating across modes. On the other hand, the reverse-KL divergence (I-projection) has a *mode-seeking* behavior and elegantly collapses on a subset of modes fairly quickly.

The contributions and organization of the remainder of the paper are:

1. We introduce a unifying framework for imitation learning as minimization of f -divergence between learner and trajectory distributions (Section 3).
2. We propose algorithms for minimizing estimates of any f -divergence. Our framework is able to recover several existing imitation learning algorithms for different divergences. We closely examine reverse KL divergence and propose efficient algorithms for it (Section 4).
3. We argue for using reverse KL to deal with multi-modal inputs (Section 5). We empirically demonstrate that reverse KL collapses to one of the demonstrator modes on both bandit and RL environments, whereas KL and JS unsafely interpolate between the modes (Section 6).

2 Related Work

Imitation learning (IL) has a long-standing history in robotics as a tool to program desired skills and behavior in autonomous machines [7–10]. Even though IL has of late been used to bootstrap reinforcement learning (RL) [11–15], we focus on the original problem where an extrinsic reward is not defined. We ask the

question – “what objective captures the notion of similarity to expert demonstrations?”. Note that this question is orthogonal to other factors such as whether we are model-based / model-free or whether we use a policy / trajectory representation.

IL can be viewed as supervised learning where the learner selects the same action as the expert (referred to as behavior cloning [16]). However small errors lead to large distribution mismatch. This can be somewhat alleviated by interactive learning, such as DAGGER [17]. Although shown to be successful in various applications [1, 18, 19], there are domains where it’s impractical to have on-policy expert labels [20, 21]. More alarmingly, there are counter-examples where the DAGGER objective results in undesirable behaviors [22]. We discuss this further in Appendix C.

Another way is to view IL as recovering a reward (IRL) [23, 24] or Q-value [25] that makes the expert seem optimal. Since this is overly strict, it can be relaxed to value matching which, for linear rewards, further reduces to matching feature expectations [26]. Moment matching naturally leads to maximum entropy formulations [27] which has been used successfully in various applications [2, 28]. Interestingly, our divergence estimators also match moments suggesting a deeper connection.

The degeneracy issues of IRL can be alleviated by a game theoretic framework where an adversary selects a reward function and the learner must compete to do as well as the expert [29, 30]. Hence IRL can be connected to min-max formulations [31] like GANs [32]. GAIL [5], SAM [33] uses this to directly recover policies. AIRL [34], EAIRL [35] uses this to recover rewards. This connection to GANs leads to interesting avenues such as stabilizing min-max games [36], learning from pure observations [37–39] and links to f-divergence minimization [6, 40].

In this paper, we view IL as f -divergence minimization between learner and expert. Our framework encompasses methods that look at specific measures of divergence such as minimizing relative entropy [41] or symmetric cross-entropy [42]. Note that [43] also independently arrives at such connections between f-divergence and IL.³ We particularly focus on multi-modal expert demonstrations which has generally been treated by clustering data and learning on each cluster [44, 45]. InfoGAN [46] formalizes the GAN framework to recover latent clusters which is then extended to IL [4, 47]. MCTE [48] extended maximum entropy formulations with casual Tsallis entropy to learn sparse multi-model policy using sparse mixture density net [49]. [50] studied how choice of divergence affected policy improvement for reinforcement learning. Here, we look at the role of divergence with multi-model expert demonstrations.

3 Problem Formulation

Preliminaries We work with a finite horizon Markov Decision Process (MDP) $\langle \mathcal{S}, \mathcal{A}, P, \rho_0, T \rangle$ where \mathcal{S} is a set of states, \mathcal{A} is a set of actions, and P is the

³ Different from [43], our framework optimizes *trajectory* divergence.

transition dynamics. $\rho_0(s)$ is the initial distribution over states and $T \in \mathbb{N}^+$ is the time horizon. In IL paradigm, the MDP does not include a reward function.

We examine stochastic policies $\pi(a|s) \in [0, 1]$. Let a trajectory be a sequence of state-action pairs $\tau = \{s_0, a_1, s_1, \dots, a_T, s_T\}$. It induces a distribution of trajectories $\rho_\pi(\tau)$ and state $\rho_\pi^t(s)$ as:

$$\begin{aligned}\rho_\pi(\tau) &= \rho_0(s_0) \prod_{t=1}^T \pi(a_t|s_{t-1}) P(s_t|s_{t-1}, a_t) \\ \rho_\pi^t(s) &= \sum_{s', a} \rho_\pi^{t-1}(s') \pi(a|s') P(s'|s, a)\end{aligned}\tag{1}$$

The average state distribution across time $\rho_\pi(s) = \frac{1}{T} \sum_{t=1}^T \rho_\pi^{t-1}(s)$ ⁴.

The f -divergence family Divergences, such as the well known Kullback-Leibler (KL) divergence, measure differences between probability distributions. We consider a broad class of such divergences called *f -divergences* [51, 52]. Given probability distributions $p(x)$ and $q(x)$ over a finite set of random variables X , such that $p(x)$ is absolutely continuous w.r.t $q(x)$, we define the f -divergence:

$$D_f(p, q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right)\tag{2}$$

where $f : \mathbb{R}^+ \rightarrow \mathbb{R}$ is a convex, lower semi-continuous function. Different choices of f recover different divergences, e.g. KL, Jensen Shannon or Total Variation (see [6] for a full list).

Imitation learning as f -divergence minimization Imitation learning is the process by which a learner tries to behave similarly to an expert based on inference from demonstrations or interactions. There are a number of ways to formalize “similarity” (Section 2) – either as a classification problem where learner must select the same action as the expert [17] or as an inverse RL problem where learner recovers a reward to explain expert behavior [23]. Neither of the formulations is error free.

We argue that the metric we actually care about is matching the distribution of trajectories $\rho_{\pi^*}(\tau) \approx \rho_\pi(\tau)$. One such reasonable objective is to minimize the f -divergence between these distributions

$$\hat{\pi} = \arg \min_{\pi \in \Pi} D_f(\rho_{\pi^*}(\tau), \rho_\pi(\tau)) = \arg \min_{\pi \in \Pi} \sum_{\tau} \rho_\pi(\tau) f\left(\frac{\rho_{\pi^*}(\tau)}{\rho_\pi(\tau)}\right)\tag{3}$$

Interestingly, different choice of f -divergence leads to different learned policies (more in Section 5).

⁴ Alternatively $\rho_\pi(s) = \sum_{\tau} \rho_\pi(\tau) \left(\frac{1}{T} \sum_{t=1}^T \mathbb{I}(s_{t-1} = s)\right)$. Refer to Theorem 2 in Appendix D

Since we have only sample access to the expert state-action distribution, the divergence between the expert and the learner has to be estimated. However, we need many samples to accurately estimate the trajectory distribution as the size of the trajectory space grows exponentially with time, i.e. $\mathcal{O}(|\mathcal{S}|^T)$. Instead, we can choose to minimize the divergence between the *average state-action distribution* as the following:

$$\begin{aligned}\hat{\pi} &= \arg \min_{\pi \in \Pi} D_f(\rho_{\pi^*}(s)\pi^*(a|s), \rho_{\pi}(s)\pi(a|s)) \\ &= \arg \min_{\pi \in \Pi} \sum_{s,a} \rho_{\pi}(s)\pi(a|s) f\left(\frac{\rho_{\pi^*}(s)\pi^*(a|s)}{\rho_{\pi}(s)\pi(a|s)}\right)\end{aligned}\quad (4)$$

We show that this lower bounds the original objective, i.e. trajectory distribution divergence.

Theorem 1 (Proof in Appendix A). *Given two policies π and π^* , the f -divergence between trajectory distribution is lower bounded by f -divergence between average state-action distribution.*

$$D_f(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) \geq D_f(\rho_{\pi^*}(s)\pi^*(a|s), \rho_{\pi}(s)\pi(a|s))$$

4 Framework for Divergence Minimization

The key problem is that we don't know the expert policy π^* and only get to observe it. Hence we are unable to compute the divergence exactly and must instead *estimate* it based on *sample* demonstrations. We build an estimator which lower bounds the state-action, and thus, trajectory divergence. The learner then minimizes the estimate.

4.1 Variational approximation of divergence

Let's say we want to measure the f -divergence between two distributions $p(x)$ and $q(x)$. Assume they are unknown but we have i.i.d samples, i.e., $x \sim p(x)$ and $x \sim q(x)$. Can we use these to estimate the divergence? [40] show that we can indeed estimate it by expressing $f(\cdot)$ in its *variational form*, i.e. $f(u) = \sup_{t \in \text{dom}_{f^*}} (tu - f^*(t))$, where $f^*(\cdot)$ is the convex conjugate⁵ Plugging this in the expression for f -divergence (2) we have

$$\begin{aligned}D_f(p, q) &= \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right) = \sum_x q(x) \sup_{t \in \text{dom}_{f^*}} \left(t \frac{p(x)}{q(x)} - f^*(t)\right) \\ &\geq \sup_{\phi \in \Phi} \sum_x q(x) \left(\phi(x) \frac{p(x)}{q(x)} - f^*(\phi(x))\right) \\ &\geq \sup_{\phi \in \Phi} \left(\underbrace{\mathbb{E}_{x \sim p(x)} [\phi(x)]}_{\text{sample estimate}} - \underbrace{\mathbb{E}_{x \sim q(x)} [f^*(\phi(x))]}_{\text{sample estimate}} \right)\end{aligned}\quad (5)$$

⁵ For a convex function $f(\cdot)$, the convex conjugate is $f^*(v) = \sup_{u \in \text{dom}_f} (uv - f(u))$. Also $(f^*)^* = f$.

Algorithm 1 f -VIM

-
- 1: Sample trajectories from expert $\tau^* \sim \rho_{\pi^*}$
 - 2: Initialize learner and estimator parameters θ_0, w_0
 - 3: **for** $i = 0$ **to** $N - 1$ **do**
 - 4: Sample trajectories from learner $\tau_i \sim \rho_{\pi_{\theta_i}}$
 - 5: Update estimator

$$w_{i+1} \leftarrow w_i + \eta_w \nabla_w \left(\sum_{(s,a) \in \tau^*} g_f(V_w(s,a)) - \sum_{(s,a) \in \tau_i} f^*(g_f(V_w(s,a))) \right)$$
 - 6: Apply policy gradient

$$\theta_{i+1} \leftarrow \theta_i - \eta_\theta \sum_{(s,a) \sim \tau_i} \nabla_\theta \log \pi_\theta(a|s) Q^{f^*(g_f(V_w))}(s,a)$$

where $Q^{f^*(g_f(V_w))}(s_{t-1}, a_t) = - \sum_{i=t}^T f^*(g_f(V_w(s_{i-1}, a_i)))$
 - 7: **end for**
 - 8: **Return** π_{θ_N}
-

Here $\phi : X \rightarrow \text{dom}_{f^*}$ is a function approximator which we refer to as an *estimator*. The lower bound is both due to Jensen's inequality and the restriction to an estimator class Φ . Intuitively, we convert divergence estimation to a discriminative classification problem between two sample sets.

How should we choose estimator class Φ ? We can find the optimal estimator ϕ^* by taking the variation of the lower bound (5) to get $\phi^*(x) = f' \left(\frac{p(x)}{q(x)} \right)$. Hence Φ should be flexible enough to approximate the subdifferential $f'(\cdot)$ *everywhere*. Can we use neural networks discriminators [32] as our class Φ ? [6] show that to satisfy the range constraints, we can parameterize $\phi(x) = g_f(V_w(x))$ where $V_w : X \rightarrow \mathbb{R}$ is an unconstrained discriminator and $g_f : \mathbb{R} \rightarrow \text{dom}_{f^*}$ is an *activation function*. We plug this in (5) and the result in (4) to arrive at the following problem.

Problem 1 (Variational Imitation (VIM)). Given a divergence $f(\cdot)$, compute a learner π and discriminator V_w as the saddle point of the following optimization

$$\hat{\pi} = \arg \min_{\pi \in \Pi} \max_w \mathbb{E}_{(s,a) \sim \rho_{\pi^*}} [g_f(V_w(s,a))] - \mathbb{E}_{(s,a) \sim \rho_\pi} [f^*(g_f(V_w(s,a)))] \quad (6)$$

where $(s,a) \sim \rho_{\pi^*}$ are sample expert demonstrations, $(s,a) \sim \rho_\pi$ are samples learner rollouts.

We propose the algorithmic framework f -VIM (Algorithm 1) which solves (6) iteratively by updating estimator V_w via supervised learning and learner θ_i via policy gradients. Algorithm 1 is a meta-algorithm. Plugging in different f -divergences (Table 1), we have different algorithms

1. *KL*-VIM: Minimizing forward KL divergence

$$\hat{\pi} = \arg \min_{\pi \in \Pi} \max_w \mathbb{E}_{(s,a) \sim \rho_{\pi^*}} [V_w(s,a)] - \mathbb{E}_{(s,a) \sim \rho_\pi} [\exp(V_w(s,a) - 1)] \quad (7)$$

2. *RKL*-VIM: Minimizing reverse KL divergence (removing constant factors)

$$\hat{\pi} = \arg \min_{\pi \in \Pi} \max_w \mathbb{E}_{(s,a) \sim \rho_{\pi^*}} [-\exp(-V_w(s,a))] + \mathbb{E}_{(s,a) \sim \rho_\pi} [-V_w(s,a)] \quad (8)$$

Table 1: List of f -Divergences used, conjugates, optimal estimators and activation function

Divergence	$f(u)$	$f^*(t)$	$\phi^*(x)$	$g_f(v)$
Kullback-Leibler	$u \log u$	$\exp(t - 1)$	$1 + \log \frac{p(x)}{q(x)}$	v
Reverse KL	$-\log u$	$-1 - \log(-t)$	$-\frac{q(x)}{p(x)}$	$-\exp(v)$
Jensen-Shannon	$-(u+1) \log \frac{1+u}{2} + u \log u$	$-\log(2 - \exp(t))$	$\log \frac{2p(x)}{p(x)+q(x)}$	$-\log(1 + \exp(-v)) + \log(2)$
Total Variation	$\frac{1}{2} u - 1 $	t	$\frac{1}{2} \text{sign}(\frac{p(x)}{q(x)} - 1)$	$\frac{1}{2} \tanh(v)$

3. JS-VIM: Minimizing Jensen-Shannon divergence

$$\hat{\pi} = \arg \min_{\pi \in \Pi} \max_w \mathbb{E}_{(s,a) \sim \rho_{\pi^*}} [\log D_w(s, a)] - \mathbb{E}_{(s,a) \sim \rho_{\pi}} [\log(1 - D_w(s, a))] \quad (9)$$

where $D_w(s, a) = (1 + \exp(-V_w(s, a)))^{-1}$.

4.2 Recovering existing imitation learning algorithms

Various existing IL approaches can be recovered under our framework. We defer the readers to Appendix C for deductions and details.

Behavior Cloning [3] – Kullback-Leibler (KL) divergence. We show that the policy minimizing the KL divergence of trajectory distribution can be $\hat{\pi} = -\mathbb{E}_{s \sim \rho_{\pi^*}, a \sim \pi^*(\cdot|s)} \log(\pi(a|s))$, which is equivalent to behavior cloning with a cross entropy loss for multi-class classification.

Generative Adversarial Imitation Learning (GAIL) [5] – Jensen-Shannon (JS) divergence. We see that JS-VIM (9) is exactly the GAIL optimization (without the entropic regularizer).

Dataset Aggregation (DAGGER) [17] – Total Variation (TV) distance. Using Pinsker’s inequality and the fact that TV is a *distance metric*, we have the following upper bound on TV

$$\begin{aligned} D_{\text{TV}}(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) &\leq T \mathbb{E}_{s \sim \rho_{\pi}(s)} [D_{\text{TV}}(\pi^*(a|s), \pi(a|s))] \\ &\leq T \sqrt{\mathbb{E}_{s \sim \rho_{\pi}(s)} [D_{\text{KL}}(\pi^*(a|s), \pi(a|s))]} \end{aligned}$$

DAGGER solves this non i.i.d problem in an iterative supervised learning manner with an interactive expert. Counter-examples to DAGGER [22] can now be explained as an artifact of this divergence.

4.3 Alternate techniques for Reverse KL minimization via interactive learning

We highlight the Reverse KL divergence which has received relatively less attention in IL literature. *RKL*-VIM (8) has some shortcomings. First, it’s a double lower bound approximation due to Theorem 1 and Equation (5). Secondly, the

optimal estimator is a state-action density ratio which maybe quite complex (Table 1). Finally, the optimization (6) may be slow to converge.

However, assuming access to an *interactive expert*, i.e. we can query an interactive expert for any $\pi^*(a|s)$, we can exploit Reverse KL divergence:

$$\begin{aligned} D_{\text{RKL}}(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) &= T \mathbb{E}_{s \sim \rho_{\pi}} [D_{\text{RKL}}(\pi^*(\cdot|s), \pi(\cdot|s))] \\ &= T \mathbb{E}_{s \sim \rho_{\pi}} \left[\sum_a \pi(a|s) \log \frac{\pi(a|s)}{\pi^*(a|s)} \right] \end{aligned}$$

Hence we can directly minimize action distribution divergence. Since this is on states induced by π , this falls under the regime of *interactive learning* [17] where we query the expert on *states visited by the learner*. We explore two different interactive learning techniques for I-projection, deferring to Appendix D and Appendix E for details.

Variational action divergence minimization. Apply the *RKL-VIM* but on *action divergence*:

$$\hat{\pi} = \arg \min_{\pi \in \Pi} \mathbb{E}_{s \sim \rho_{\pi}} [\mathbb{E}_{a \sim \pi^*(\cdot|s)} [-\exp(V_w(s, a))] + \mathbb{E}_{a \sim \pi(\cdot|s)} [V_w(s, a)]] \quad (10)$$

Unlike *RKL-VIM*, we collect a fresh batch of data from *both* an interactive expert and learner every iteration. We show that this estimator is far easier to approximate than *RKL-VIM* (Appendix D).

Density ratio minimization via no regret online learning. We first upper bound the action divergence:

$$\begin{aligned} D_{\text{RKL}}(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) &= T \mathbb{E}_{s \sim \rho_{\pi}} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} \left[\log \frac{\pi(a|s)}{\pi^*(a|s)} \right] \right] \\ &\leq T \mathbb{E}_{s \sim \rho_{\pi}} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} \left[\frac{\pi(a|s)}{\pi^*(a|s)} - 1 \right] \right] \end{aligned}$$

Given a batch of data from an interactive expert and the learner, we invoke an off-shelf density ratio estimator (DRE) [53] to get $\hat{r}(s, a) \approx \frac{\rho_{\pi}(s)\pi(a|s)}{\rho_{\pi}(s)\pi^*(a|s)} = \frac{\pi(a|s)}{\pi^*(a|s)}$. Since the optimization is a non i.i.d learning problem, we solve it by dataset aggregation. Note this *does not require invoking policy gradients*. In fact, if we choose an expressive enough policy class, this method gives us a global performance guarantee which neither GAIL or any *f-VIM* provides (Appendix E).

5 Multi-modal Trajectory Demonstrations

We now examine multi-modal expert demonstrations. Consider the demonstrations in Fig. 2 which avoid colliding with a tree by turning left or right with equal probability. Depending on the policy class, it may be impossible to achieve zero divergence for *any* choice of *f*-divergence (Fig. 2a), e.g., Π is Gaussian with fixed variance. Then the question becomes, if the globally optimal policy in our policy class achieves non-zero divergence, how should we design our objective to fail elegantly and safely? In this example, one can imagine two reasonable choices: (1)

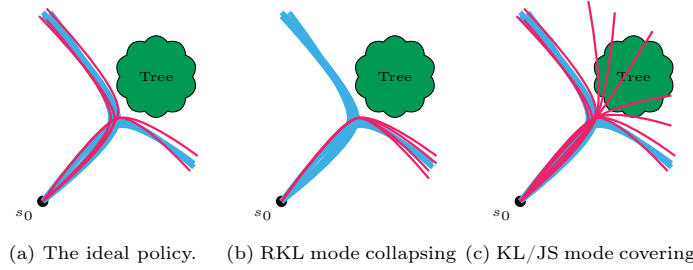


Fig. 2: Illustration of the safety concerns of mode-covering behavior. (a) Expert demonstrations and policy roll-outs are shown in blue and red, respectively. (b) RKL receives only a small penalty for the safe behavior whereas KL receives an infinite penalty. (c) The opposite is true for the unsafe behavior where learner crashes.

replicate one of the modes (mode-collapsing) or (2) cover both the modes plus the region between (mode-covering). We argue that in some imitation learning tasks when the dominant mode is desirable, paradigm (1) is preferable.

Mode-covering in KL. This divergence exhibits strong mode-covering tendencies as in Fig. 2c. Examining the definition of the KL divergence, we see that there is a significant penalty for failing to completely support the demonstration distribution, but no explicit penalty for generating outlier samples. In fact, if $\exists s, a$ s.t. $\rho_{\pi^*}(s, a) > 0, \rho_{\pi}(s, a) = 0$, then the divergence is infinite. However, the opposite does not hold. Thus, the *KL-VIM* optimal policy in Π belongs to the second behavior class – which the agent to frequently crash into the tree.

Mode-collapsing in RKL. At the other end of the multi-modal behavior spectrum lies the RKL divergence, which exhibits strong mode-seeking behavior as in Fig. 2b, due to switching the expectation over ρ_{π} with ρ_{π^*} . Note there is no explicit penalty for failing to entirely cover ρ_{π^*} , but an arbitrarily large penalty for generating samples which would be improbable under the demonstrator distribution. This results in always turning left or always turning right around the tree, depending on the initialization and mode mixture. For many tasks, failing in such a manner is predictable and safe, as we have already seen similar trajectories from the demonstrator.

Jensen-Shannon. This divergence may fall into either behavior class, depending on the MDP, the demonstrations, and the optimization initialization. Examining the definition, we see the divergence is symmetric and expectations are taken over both ρ_{π} and ρ_{π^*} . Thus, if either distribution is unsupported (i.e. $\exists s, a$ s.t. $\rho_{\pi^*}(s, a) > 0, \rho_{\pi}(s, a) = 0$ or vice versa) the divergence remains finite. Later, we empirically show that although it is possible to achieve safe mode-collapse with JS on some tasks, this is not always the case.

6 Experiments

6.1 Low dimensional tasks

In this section, we empirically validate the following **Hypotheses**:

- H1** *The globally optimal policy for RKL imitates a subset of the demonstrator modes, whereas JS and KL tend to interpolate between them.*
- H2** *The sample-based estimator for KL and JS underestimates the divergence more than RKL.*
- H3** *The policy gradient optimization landscape for KL and JS with continuously parameterized policies is more susceptible to local minima, compared to RKL.*

We test these hypothesis on two environments. The **Bandit environment** has a single state and three actions, a , b and c . The expert chooses a and b with equal probability as in Fig. 3a. We choose a policy class Π which has 3 policies A , B , and M . A selects a , B selects b and M stochastically selects a , b , or c with probability $(\epsilon_0, \epsilon_0, 1 - 2\epsilon_0)$. The **GridWorld environment** has a 3×3 states (Fig. 3b). There are a start (S) and a terminal (T) state. The center state is undesirable. The environment has control noise ϵ_1 and transition noise ϵ_2 . Fig. 3d shows the expert’s multi-modal demonstration. The policy class Π allows agents to go *up*, *right*, *down*, *left* at each state.

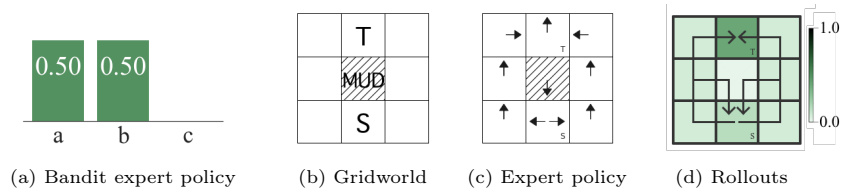


Fig. 3: Bandit and gridworld environment.

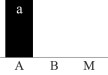

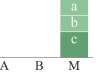
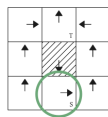
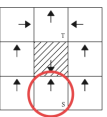
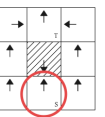
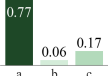


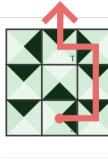


Policy enumeration To test **H1**, we enumerate through all policies in Π , exactly compute their stationary distributions $\rho_\pi(s, a)$, and select the policy with the smallest exact f -divergence, the optimal policy. Our results on the bandit and gridworld (Table 2a and 2b) show that the globally optimal solution to the RKL objective successfully collapses to a single mode (e.g. A and Right, respectively), whereas KL and JS interpolate between the modes (i.e. M and Up, respectively).

Whether the optimal policy is mode-covering or collapsing depends on the *stochasticity in the policy*. In the bandit environment we parameterize this by ϵ_0 and show in Fig 4 how the divergences and resulting optimal policy changes as a function of ϵ_0 . Note that RKL strongly prefers mode collapsing, KL strongly prefers mode covering, and JS is between the two other divergences.

Divergence estimation To test **H2**, we compare the sample-based estimation of f -divergence to the true value in Fig. 5. We highlight the preferred policies under each objective (in the 1 percentile of estimations). For the highlighted group, the estimation is often much lower than the true divergence for KL and JS, perhaps due to the sampling issue discussed in Appendix F.

Policy gradient optimization landscape To test **H3**, we solve for a local-optimal policy using policy gradient for KL -VIM, RKL -VIM and JS -VIM. Though the bandit problem and the gridworld environment have only discrete actions, we

Table 2: Globally optimal policies produced by policy enumeration (2a and 2b), and locally optimal policies produced by policy gradient (2c and 2d). In all cases, the RKL policy tends to collapse to one of the demonstrator modes, whereas the other policies interpolate between the modes, resulting in unsafe behavior.

	Bandit			GridWorld		
	RKL	JS	KL	RKL	JS	KL
H1 global-optima						
	(a)			(b)		
H3 local-optima						
	(c)			(d)		

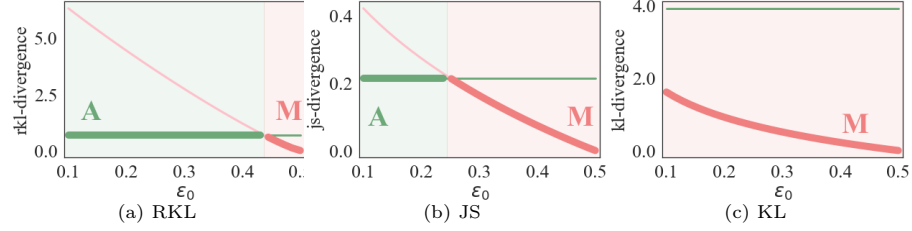


Fig. 4: Divergences and corresponding optimal policy as a function of the control noise ϵ_0 . RKL strongly prefers the mode collapse policy A (except at high control noise), KL strongly prefers the mode covering policy M, and JS is between the two.

consider a continuously parameterized policy class (Appendix G) for use with policy gradient. Table 2c and 2d shows that RKL-VIM empirically produces policies that collapses to a single mode whereas JS and KL-VIM do not.

6.2 High dimensional continuous control task

We tested *RKL-VIM* and *JS-VIM* (GAIL) on a set of high dimensional control tasks in Mujoco. Though our main interest is in multi-modal behavior which occurs frequently in human demonstrations, here we had to generate expert demonstrations using a reinforcement learning policy, which are *single modal*.

The vanilla version of these algorithms were significantly slow to maximize the cumulative reward. Further examination revealed that there were multiple saddle points that the system gets ‘stuck’ in. A reliable way to coax the algo-

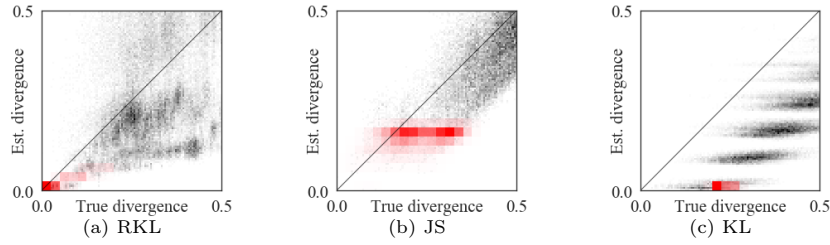


Fig. 5: Comparing f -divergence with the estimated values. Preferred policies under each objective (in the 1 percentile of estimations) are in red. The normalized estimations appear to be typically lower than the normalized true values for JS and KL.

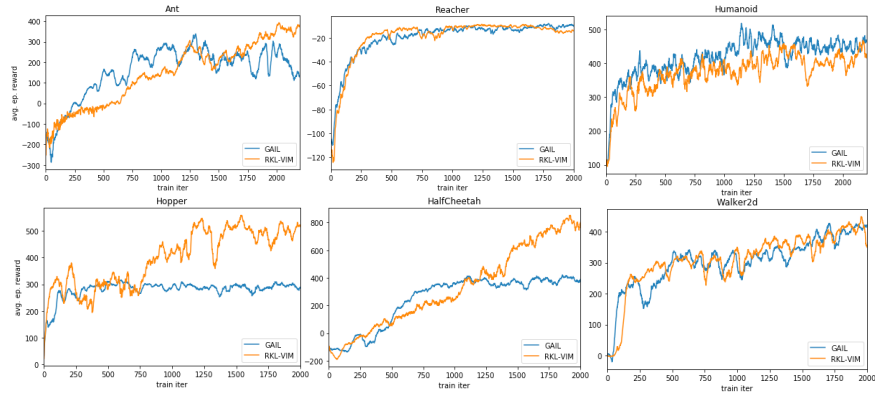


Fig. 6: Training RKL -VIM and JS -VIM (GAIL) on Mujoco environments.

rithm to the desired saddle point was to mix in a small percentage of the true reward along with the discriminator loss. Hence, we augmented the generator loss $\mathbb{E}_{(s,a) \sim \rho_\pi} [(1 - \alpha) - f^*(g_f(V_w(s, a))) + \alpha r(s, a)]$ where $\alpha = 0.2$. This resulted in reliable, albeit different, convergence from both algorithms.

Fig. 6 shows the average episodic reward over training iterations. On Humanoid, Reacher and Walker2d the performance of both algorithms are similar. However on Ant, Hopper and HalfCheetah RKL -VIM converges to a higher value. Further inspection of the discriminator loss reveals that RKL -VIM heavily upweights states that the expert visits over the states that the learner visits. While this makes convergence slightly more sluggish (e.g. Ant), the algorithm terminates with a higher reward.

7 Discussion

We presented an imitation learning framework based on f -divergences, which generalizes existing approaches including behavior cloning (KL), GAIL (JS), and DAGGER (TV). In settings with multi-modal demonstrations, we showed

that RKL divergence safely and efficiently collapses to a subset of the modes, whereas KL and JS often produce unsafe behavior.

Our framework minimizes an *approximate estimation* of the divergence, notably a lower bound (5). KL divergence is the only one we can actually measure (Appendix C). The lower bound (5) is tight if the function approximator $\phi(x)$ has enough capacity to express the function $f'(\frac{p(x)}{q(x)})$. For Reverse KL, $f(u) = -\log u$ and $f'(u) = -\frac{1}{u}$. Hence $f'(\cdot)$ can be unbounded and we may need exponentially large number of samples to correctly estimate $\phi(x)$. On the other hand, deriving a *tight upper bound* on the f -divergence from a finite set of samples is also impossible. e.g. For RKL, without any assumptions about the expert or learner distribution, there is no way to estimate the support accurately given a finite number of samples. Hence we are left only with the choice of ∞ which is vacuous.

There are a few practical remedies that center around a key observation – we care not about measuring the divergence but rather minimizing it. One way to do so is to consider a *noisy* version of divergence minimization as in [54], essentially adding Gaussian noise to both learner and expert to ensure both distributions are absolutely continuous. This upper bounds the magnitude of the divergence. We can think of this as smoothing out the cost function that the policy chooses to minimize. This would help in faster convergence.

We can take these intuitions further and view imitation learning as computing a really good loss - a balance between a loss that maximizes likelihood of expert actions (KL divergence) and a loss that penalizes the learner from visiting states that the expert does not visit. Instead of using estimating the latter term, we can potentially exploit side information. For example, we may already know that the expert does not like to violate obstacle constraints (a fact that we can test from the data). This can then be simply added in as an auxiliary penalty term.

There are a couple interesting directions for future work. One is to unify this framework with maximum entropy moment matching. Given a set of basis function $\phi(x)$, MaxEnt solves for a maximum entropy distribution $q(x)$ such that the moments of the basis functions are matched $\mathbb{E}_{x \sim p(x)} [\phi(x)] = \mathbb{E}_{x \sim q(x)} [\phi(x)]$. Contrast this to (5) where moments of a transformed function are matched. Consequently, MaxEnt *symmetrically* bumps down cost of expert states and bumps up the cost of learner states. In contrast, RKL-VIM (8) *exponentially* bumps down cost of expert and *linearly* bumps up the cost of learner states.

Another interesting direction would be to consider the class of integral probability metrics (IPM). IPMs are metrics that take the form $\sup_{\phi \in \Phi} \mathbb{E}_{x \sim p(x)} [\phi(x)] - \mathbb{E}_{x \sim q(x)} [\phi(x)]$. Unlike f -divergence estimators, these metrics are measurable by definition. Choosing different families of Φ results in MMD, TotalVariation, Earth-movers distance. Preliminary results using such estimators seem promising [55].

Acknowledgements This work was (partially) funded by the National Institute of Health R01 (#R01EB019335), National Science Foundation CPS (#1544797), National Science Foundation NRI (#1637748), the Office of Naval Research, the RCTA, Amazon, and Honda Research Institute USA.

Bibliography

- [1] Stéphane Ross, Narek Melik-Barkhudarov, Kumar Shaurya Shankar, Andreas Wendel, Debadeepta Dey, J Andrew Bagnell, and Martial Hebert. Learning monocular reactive uav control in cluttered natural environments. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, 2013.
- [2] Chelsea Finn, Sergey Levine, and Pieter Abbeel. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*, pages 49–58, 2016.
- [3] Dean A Pomerleau. ALVINN: An autonomous land vehicle in a neural network. In D S Touretzky, editor, *Advances in Neural Information Processing Systems 1*, pages 305–313. Morgan-Kaufmann, 1989.
- [4] Yunzhu Li, Jiaming Song, and Stefano Ermon. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, pages 3812–3822, 2017.
- [5] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016.
- [6] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in Neural Information Processing Systems*, pages 271–279, 2016.
- [7] Takayuki Osa, Joni Pajarinen, Gerhard Neumann, J Andrew Bagnell, Pieter Abbeel, and Jan Peters. An algorithmic perspective on imitation learning. *arXiv preprint arXiv:1811.06711*, 2018.
- [8] Brenna D Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. A survey of robot learning from demonstration. *Robotics and autonomous systems*, 57(5): 469–483, 2009.
- [9] Aude G Billard, Sylvain Calinon, and Rüdiger Dillmann. Learning from humans. In *Springer handbook of robotics*, pages 1995–2014. Springer, 2016.
- [10] J. Andrew (Drew) Bagnell. An invitation to imitation. Technical Report CMU-RI-TR-15-08, Carnegie Mellon University, Pittsburgh, PA, March 2015.
- [11] Stéphane Ross and J Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning. *arXiv preprint arXiv:1406.5979*, 2014.
- [12] Wen Sun, Arun Venkatraman, Geoffrey J Gordon, Byron Boots, and J Andrew Bagnell. Deeply aggravated: Differentiable imitation learning for sequential prediction. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 3309–3318. JMLR. org, 2017.
- [13] Wen Sun, J Andrew Bagnell, and Byron Boots. Truncated horizon policy search: Combining reinforcement learning & imitation learning. *arXiv:1805.11240*, 2018.
- [14] Ching-An Cheng, Xinyan Yan, Nolan Wagener, and Byron Boots. Fast policy learning through imitation and reinforcement. *arXiv:1805.10413*, 2018.
- [15] Aravind Rajeswaran, Vikash Kumar, Abhishek Gupta, Giulia Vezzani, John Schulman, Emanuel Todorov, and Sergey Levine. Learning complex dexterous manipulation with deep reinforcement learning and demonstrations. *arXiv preprint arXiv:1709.10087*, 2017.
- [16] Dean A Pomerleau. Alvin: An autonomous land vehicle in a neural network. In *Advances in neural information processing systems*, pages 305–313, 1989.
- [17] Stéphane Ross, Geoffrey J Gordon, and Drew Bagnell. A reduction of imitation learning and structured prediction to no-regret online learning. In *AISTATS*, 2011.
- [18] Beomjoon Kim, Amir-massoud Farahmand, Joelle Pineau, and Doina Precup. Learning from limited demonstrations. In *Advances in Neural Information Processing Systems*, pages 2859–2867, 2013.

- [19] Saurabh Gupta, James Davidson, Sergey Levine, Rahul Sukthankar, and Jitendra Malik. Cognitive mapping and planning for visual navigation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017.
- [20] Michael Laskey, Jonathan Lee, Wesley Hsieh, Richard Liaw, Jeffrey Mahler, Roy Fox, and Ken Goldberg. Iterative noise injection for scalable imitation learning. *arXiv preprint arXiv:1703.09327*, 2017.
- [21] Michael Laskey, Sam Staszak, Wesley Yu-Shu Hsieh, Jeffrey Mahler, Florian T Pokorný, Anca D Dragan, and Ken Goldberg. Shiv: Reducing supervisor burden in dagger using support vectors for efficient learning from demonstrations in high dimensional state spaces. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*, pages 462–469. IEEE, 2016.
- [22] Michael Laskey, Caleb Chuck, Jonathan Lee, Jeffrey Mahler, Sanjay Krishnan, Kevin Jamieson, Anca Dragan, and Ken Goldberg. Comparing human-centric and robot-centric sampling for robot deep learning from demonstrations. In *2017 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2017.
- [23] Nathan D Ratliff, David Silver, and J Andrew Bagnell. Learning to search: Functional gradient techniques for imitation learning. *Autonomous Robots*, 2009.
- [24] Nathan D Ratliff, J Andrew Bagnell, and Martin A Zinkevich. Maximum margin planning. In *International Conference on Machine learning*. ACM, 2006.
- [25] Bilal Piot, Matthieu Geist, and Olivier Pietquin. Bridging the gap between imitation learning and inverse reinforcement learning. *IEEE transactions on neural networks and learning systems*, 28(8):1814–1826, 2017.
- [26] Pieter Abbeel and Andrew Y Ng. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*, page 1. ACM, 2004.
- [27] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. Maximum entropy inverse reinforcement learning. In *AAAI*, 2008.
- [28] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.
- [29] Umar Syed and Robert E Schapire. A game-theoretic approach to apprenticeship learning. In *Advances in neural information processing systems*, 2008.
- [30] Jonathan Ho, Jayesh Gupta, and Stefano Ermon. Model-free imitation learning with policy optimization. In *International Conference on Machine Learning*, 2016.
- [31] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852*, 2016.
- [32] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.
- [33] Lionel Blondé and Alexandros Kalousis. Sample-efficient imitation learning via generative adversarial nets. *arXiv preprint arXiv:1809.02064*, 2018.
- [34] Justin Fu, Katie Luo, and Sergey Levine. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248*, 2017.
- [35] Ahmed H Qureshi and Michael C Yip. Adversarial imitation via variational inverse reinforcement learning. *arXiv preprint arXiv:1809.06404*, 2018.
- [36] Xue Bin Peng, Angjoo Kanazawa, Sam Toyer, Pieter Abbeel, and Sergey Levine. Variational discriminator bottleneck: Improving imitation learning, inverse rl, and gans by constraining information flow. *arXiv preprint arXiv:1810.00821*, 2018.
- [37] Faraz Torabi, Garrett Warnell, and Peter Stone. Generative adversarial imitation from observation. *arXiv preprint arXiv:1807.06158*, 2018.

- [38] Faraz Torabi, Garrett Warnell, and Peter Stone. Behavioral cloning from observation. *arXiv preprint arXiv:1805.01954*, 2018.
- [39] Xue Bin Peng, Angjoo Kanazawa, Jitendra Malik, Pieter Abbeel, and Sergey Levine. Sfr: Reinforcement learning of physical skills from videos. In *SIGGRAPH Asia 2018 Technical Papers*, page 178. ACM, 2018.
- [40] XuanLong Nguyen, Martin J Wainwright, and Michael I Jordan. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.
- [41] Abdeslam Boularias, Jens Kober, and Jan Peters. Relative entropy inverse reinforcement learning. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pages 182–189, 2011.
- [42] Nicholas Rhinehart, Kris M. Kitani, and Paul Vernaza. R2p2: A reparameterized pushforward policy for diverse, precise generative path forecasting. In *The European Conference on Computer Vision (ECCV)*, September 2018.
- [43] Seyed Kamyar Seyed Ghasemipour, Shixiang Gu, and Richard Zemel. Understanding the relation between maximum-entropy inverse reinforcement learning and behaviour cloning. *Workshop ICLR*, 2018.
- [44] Monica Babes, Vukosi Marivate, Kaushik Subramanian, and Michael L Littman. Apprenticeship learning about multiple intentions. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 897–904, 2011.
- [45] Christos Dimitrakakis and Constantin A Rothkopf. Bayesian multitask inverse reinforcement learning. In *European Workshop on Reinforcement Learning*, pages 273–284. Springer, 2011.
- [46] Xi Chen, Yan Duan, Rein Houthoofd, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. In *Advances in neural information processing systems*, pages 2172–2180, 2016.
- [47] Karol Hausman, Yevgen Chebotar, Stefan Schaal, Gaurav Sukhatme, and Joseph J Lim. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets. In *Advances in Neural Information Processing Systems*, pages 1235–1245, 2017.
- [48] Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Maximum causal tsallis entropy imitation learning. In *Advances in Neural Information Processing Systems*, 2018.
- [49] Kyungjae Lee, Sungjoon Choi, and Songhwai Oh. Sparse markov decision processes with causal sparse tsallis entropy regularization for reinforcement learning. *IEEE Robotics and Automation Letters*, 2018.
- [50] Boris Belousov and Jan Peters. f-divergence constrained policy improvement. *arXiv preprint arXiv:1801.00056*, 2017.
- [51] Imre Csiszár and Paul C Shields. *Information theory and statistics: A tutorial*. Now Publishers Inc, 2004.
- [52] Friedrich Liese and Igor Vajda. On divergences and informations in statistics and information theory. *IEEE Transactions on Information Theory*, 2006.
- [53] Takafumi Kanamori, Taiji Suzuki, and Masashi Sugiyama. Statistical analysis of kernel-based least-squares density-ratio estimation. *Machine Learning*, 86(3): 335–367, 2012.
- [54] Mingtian Zhang, Thomas Bird, Raza Habib, Tianlin Xu, and David Barber. Variational f-divergence minimization. *arXiv preprint arXiv:1907.11891*, 2019.
- [55] Wen Sun, Anirudh Vemula, Byron Boots, and J Andrew Bagnell. Provably efficient imitation learning from observation alone. *arXiv preprint arXiv:1905.10948*, 2019.

Appendix for “Imitation Learning as f-Divergence Minimization”

A Lower Bounding f -Divergence of Trajectory Distribution with State Action Distribution

We begin with a lemma that relates f -divergence between two vectors and their sum.

Lemma 1 (Generalized log sum inequality). *Let p_1, \dots, p_n and q_1, \dots, q_n be non-negative numbers. Let $p = \sum_{i=1}^n p_i$ and $q = \sum_{i=1}^n q_i$. Let $f(\cdot)$ be a convex function. We have the following:*

$$\sum_{i=1}^n q_i f\left(\frac{p_i}{q_i}\right) \geq q f\left(\frac{p}{q}\right) \quad (11)$$

Proof.

$$\sum_{i=1}^n q_i f\left(\frac{p_i}{q_i}\right) = q \sum_{i=1}^n \frac{q_i}{q} f\left(\frac{p_i}{q_i}\right) \geq q f\left(\sum_{i=1}^n \frac{q_i}{q} \frac{p_i}{q_i}\right) \geq q f\left(\frac{1}{q} \sum_{i=1}^n p_i\right) \geq q f\left(\frac{p}{q}\right) \quad (12)$$

where (12) is due to Jensen’s inequality since $f(\cdot)$ is convex and $q_i \geq 0$ and $\sum_{i=1}^n \frac{q_i}{q} = 1$. \square

We use Lemma 1 to prove a more general lemma that relates the f -divergence defined over two spaces where one of the space is rich enough in information to explain away the other.

Lemma 2 (Information loss). *Let a and b be two random variables. Let $P(a, b)$ be a joint probability distribution. The marginal distributions are $P(a) = \sum_b P(a, b)$ and $P(b) = \sum_a P(a, b)$. Assume that a **contains all information of** b . This is expressed as follows – given any two probability distribution $P(\cdot)$, $Q(\cdot)$, assume the following equality holds for all a, b :*

$$P(b|a) = Q(b|a) \quad (13)$$

Under these conditions, the following inequality holds:

$$\sum_a Q(a) f\left(\frac{P(a)}{Q(a)}\right) \geq \sum_b Q(b) f\left(\frac{P(b)}{Q(b)}\right) \quad (14)$$

Proof.

$$\sum_a Q(a) f\left(\frac{P(a)}{Q(a)}\right) = \sum_a \left(\sum_b Q(a, b) \right) f\left(\frac{P(a)}{Q(a)}\right) \quad (15)$$

$$= \sum_a \sum_b Q(a, b) f\left(\frac{P(a)}{Q(a)}\right) \quad (16)$$

$$= \sum_b \sum_a Q(a, b) f\left(\frac{P(a, b)/P(b|a)}{Q(a, b)/Q(b|a)}\right) \quad (17)$$

$$= \sum_b \sum_a Q(a, b) f\left(\frac{P(a, b)}{Q(a, b)}\right) \quad (18)$$

$$\geq \sum_b \left(\sum_a Q(a, b) \right) f\left(\frac{\left(\sum_a P(a, b) \right)}{\left(\sum_a Q(a, b) \right)}\right) \quad (19)$$

$$\geq \sum_b Q(b) f\left(\frac{P(b)}{Q(b)}\right) \quad (20)$$

We get (17) by applying $P(a, b) = P(a)P(b|a)$ and $Q(a, b) = Q(a)Q(b|a)$. We get (18) applying the equality constraint from (13). We get (19) from Lemma 1 by setting $p_i = P(a, b)$, $q_i = Q(a, b)$ and summing over all a keeping b fixed. \square

We are now ready to prove Theorem 1 using Lemma 2.

Proof of Theorem 1. Let random variable a belong to the space of trajectories τ . Let random variable b belong to the space of state action pairs $z = (s, a)$. Note that for any joint distribution $P(z, \tau)$ and $Q(z, \tau)$, the following is true

$$P(z|\tau) = Q(z|\tau) \quad (21)$$

This is because a trajectory τ contains all information about z , i.e. $\tau = \{s_0, a_1, s_1, \dots\}$. Upon applying Lemma 2 we have the inequality

$$\sum_{\tau} Q(\tau) f\left(\frac{P(\tau)}{Q(\tau)}\right) \geq \sum_z Q(z) f\left(\frac{P(z)}{Q(z)}\right) = \sum_{(s,a)} \rho_{\pi}(s, a) f\left(\frac{\rho_{\pi^*}(s, a)}{\rho_{\pi}(s, a)}\right) \quad (22)$$

\square

The bound is reasonable as it merely states that information gets lost when temporal information is discarded. Note that the theorem also extends to state distributions, i.e.

Corollary 1. *Divergence between trajectory distribution is lower bounded by state distribution.*

$$D_f(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) \geq D_f(\rho_{\pi^*}(s), \rho_{\pi}(s))$$

How tight is this lower bound? We examine the gap

Corollary 2. *The gap between the two divergences is*

$$D_f(P(\tau), Q(\tau)) - D_f(P(z), Q(z)) = \sum_z P(z) D_f(P(\tau|z), Q(\tau|z))$$

Proof.

$$\begin{aligned} \sum_{\tau} P(z, \tau) f\left(\frac{P(z, \tau)}{Q(z, \tau)}\right) - P(z) f\left(\frac{P(z)}{Q(z)}\right) &= P(z) \sum_{\tau} P(\tau|z) f\left(\frac{P(\tau|z)P(z)}{Q(\tau|z)Q(z)}\right) - P(z) f\left(\frac{P(z)}{Q(z)}\right) \\ &= P(z) \left(\sum_{\tau} P(\tau|z) f\left(\frac{P(\tau|z)P(z)}{Q(\tau|z)Q(z)}\right) - f\left(\frac{P(z)}{Q(z)}\right) \right) \\ &= P(z) \left(\sum_{\tau} P(\tau|z) f\left(\frac{P(\tau|z)P(z)}{Q(\tau|z)Q(z)}\right) - \sum_{\tau} P(\tau|z) f\left(\frac{P(z)}{Q(z)}\right) \right) \\ &= P(z) \cdot D_f(P(\tau|z), Q(\tau|z)) \end{aligned}$$

where we use $\sum_{\tau} P(\tau|z) = 1$. □

Let \mathcal{A} be the set of trajectories that contain z , i.e., $\mathcal{A} = \{\tau | P(z|\tau) = Q(z|\tau) > 0\}$. The gap is the conditional f-divergence of $\tau \in \mathcal{A}$ scaled by $P(z)$. The gap comes from whether we treat $\tau \in \mathcal{A}$ as separate events (in the case of trajectories) or as the same event (in the case of z).

B Relating f -Divergence of Trajectory Distribution with Expected Action Distribution

In this section we explore the relation of divergences between induced trajectory distribution and induced action distribution. We begin with a general lemma

Lemma 3. *Given a policy π and a general feature function $\phi(s, a)$, the expected feature counts along induced trajectories is the same as expected feature counts on induced state action distribution*

$$\sum_{\tau} \rho_{\pi}(\tau) \left(\sum_{t=1}^T \phi(s_{t-1}, a_t) \right) = T \sum_s \rho_{\pi}(s) \sum_a \pi(a|s) \phi(s, a) \quad (23)$$

Proof. Expanding the LHS we have

$$\sum_{\tau} \rho_{\pi}(\tau) \left(\sum_{t=1}^T \phi(s_{t-1}, a_t) \right) \quad (24)$$

$$= \sum_{t=1}^T \sum_{\tau} \rho_{\pi} \phi(s_{t-1}, a_t) \quad (25)$$

$$= \sum_{t=1}^T \sum_{s_{t-1}} \rho_{\pi}^{t-1}(s_{t-1}) \sum_{a_t} \pi(a_t|s_{t-1}) \phi(s_{t-1}, a_t) \sum_{s_{t+1}} P(s_t|s_{t-1}, a_t) \cdots \sum_{s_T} P(s_T|s_{T-1}, a_T) \quad (26)$$

$$= \sum_{t=1}^T \sum_{s_{t-1}} \rho_{\pi}^{t-1}(s_{t-1}) \sum_{a_t} \pi(a_t|s_{t-1}) \phi(s_{t-1}, a_t) \quad (27)$$

$$= \sum_{t=1}^T \sum_s \rho_{\pi}^{t-1}(s) \sum_a \pi(a|s) \phi(s, a) \quad (28)$$

$$= T \sum_s \rho_{\pi}(s) \sum_a \pi(a|s) \phi(s, a) \quad (29)$$

where (27) is due to marginalizing out the future, (28) is due to the fact that state space is same across time and (29) results from applying average state distribution definition. \square

We can use this lemma to get several useful equalities such as the average state visitation frequency

Theorem 2. *Given a policy π , if we do a tally count of states visited by induced trajectories, we recover the average state visitation frequency.*

$$\sum_{\tau} \rho_{\pi}(\tau) \left(\sum_{t=1}^T \frac{1}{T} \mathbb{I}(s_{t-1} = z) \right) = \rho_{\pi}(z) \quad (30)$$

Proof. Apply Lemma 3 with $\phi(s, a) = \mathbb{I}(s = z)$ \square

Unfortunately Lemma 3 does not hold for f -divergences in general. But we can analyze a subclass of f -divergences that satisfy the following triangle inequality:

$$D_f(p, q) \leq D_f(p, r) + D_f(r, q) \quad (31)$$

Examples of such divergences are Total Variation distance or Squared Hellinger distance.

We now show that for such divergences (which are actually distances), we can *upper bound* the f -divergence. Contrast this to the *lower bound* discussed in Appendix A. The upper bound is attractive because the trajectory divergence is the term that we actually care about bounding.

Also note the implications of the upper bound – we now need expert labels on states collected by the learner $s \sim \rho_\pi$. Hence we need an interactive expert that we can query from arbitrary states.

Theorem 3 (Upper bound). *Given two policies π and π^* , and f -divergences that satisfy the triangle inequality, divergence between the trajectory distribution is upper bounded by the expected divergence between the action distribution on states induced by π .*

$$D_f(\rho_{\pi^*}(\tau), \rho_\pi(\tau)) \leq T \mathbb{E}_{s \sim \rho_\pi} [D_f(\pi^*(a|s), \pi(a|s))]$$

Proof of Theorem 3. We will introduce some notations to aid in explaining the proof. Let a trajectory segment $\tau_{t,T}$ be

$$\tau_{t,T} = \{s_t, a_{t+1}, s_{t+1}, \dots, a_T, s_T\} \quad (32)$$

Recall that the probability of a trajectory induced by policy π is

$$\rho_\pi(\tau_{0,T}) = \rho_0(s_0) \prod_{t=1}^T \pi(a_t|s_{t-1}) P(s_t|s_{t-1}, a_t) \quad (33)$$

We also introduce a non-stationary policy $\bar{\pi}$ that executes π and then π^* thereafter. Hence, the probability of a trajectory induced by $\bar{\pi}$ is

$$\rho_{\bar{\pi}}(\tau_{0,T}) = \rho_0(s_0) \pi(a_1|s_0) P(s_1|s_0, a_1) \prod_{t=2}^T \pi^*(a_t|s_{t-1}) P(s_t|s_{t-1}, a_t) \quad (34)$$

Let us consider the divergence between distributions $\rho_{\pi^*}(\tau_{0,T})$ and $\rho_{\pi}(\tau_{0,T})$ and apply the triangle inequality (31) with respect to $\rho_{\bar{\pi}}(\tau_{0,T})$

$$D_f(\rho_{\pi^*}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \quad (35)$$

$$\leq D_f(\rho_{\pi^*}(\tau_{0,T}), \rho_{\bar{\pi}}(\tau_{0,T})) + D_f(\rho_{\bar{\pi}}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \quad (36)$$

$$\leq \sum_{\tau_{0,T}} \rho_{\bar{\pi}}(\tau_{0,T}) f\left(\frac{\rho_{\pi^*}(\tau_{0,T})}{\rho_{\bar{\pi}}(\tau_{0,T})}\right) + D_f(\rho_{\bar{\pi}}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \quad (37)$$

$$\leq \sum_{\tau_{0,T}} \rho_{\bar{\pi}}(\tau_{0,T}) f\left(\frac{\rho_0(s_0)\pi^*(a_1|s_0)P(s_1|s_0, a_1) \prod_{t=2}^T \pi^*(a_t|s_{t-1})P(s_t|s_{t-1}, a_t)}{\rho_0(s_0)\pi(a_1|s_0)P(s_1|s_0, a_1) \prod_{t=2}^T \pi^*(a_t|s_{t-1})P(s_t|s_{t-1}, a_t)}\right) \quad (38)$$

$$+ D_f(\rho_{\bar{\pi}}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T}))$$

$$\leq \sum_{\tau_{0,T}} \rho_{\bar{\pi}}(\tau_{0,T}) f\left(\frac{\pi^*(a_1|s_0)}{\pi(a_1|s_0)}\right) + D_f(\rho_{\bar{\pi}}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \quad (39)$$

$$\leq \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) f\left(\frac{\pi^*(a_1|s_0)}{\pi(a_1|s_0)}\right) \sum_{s_1} P(s_1|s_0, a_1) \sum_{a_1} \cdots + D_f(\rho_{\bar{\pi}}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \quad (40)$$

$$\leq \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) f\left(\frac{\pi^*(a_1|s_0)}{\pi(a_1|s_0)}\right) + D_f(\rho_{\bar{\pi}}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \quad (41)$$

$$\leq \sum_s \rho_0(s) \sum_a \pi(a|s) f\left(\frac{\pi^*(a|s)}{\pi(a|s)}\right) + D_f(\rho_{\bar{\pi}}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \quad (42)$$

$$\leq \mathbb{E}_{s \sim \rho_0(s)} [D_f(\pi^*(a|s), \pi(a|s))] + D_f(\rho_{\bar{\pi}}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \quad (43)$$

Expanding the second term we have

$$D_f(\rho_{\bar{\pi}}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \quad (44)$$

$$= \sum_{\tau_{0,T}} \rho_{\pi}(\tau_{0,T}) f \left(\frac{\rho_0(s_0) \pi(a_1|s_0) P(s_1|s_0, a_1) \prod_{t=2}^T \pi^*(a_t|s_{t-1}) P(s_t|s_{t-1}, a_t)}{\rho_0(s_0) \pi(a_1|s_0) P(s_1|s_0, a_1) \prod_{t=2}^T \pi(a_t|s_{t-1}) P(s_t|s_{t-1}, a_t)} \right) \quad (45)$$

$$= \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) \sum_{s_1} P(s_1|s_0, a_1) \dots f \left(\frac{P(s_1|s_0, a_1) \prod_{t=2}^T \pi^*(a_t|s_{t-1}) P(s_t|s_{t-1}, a_t)}{P(s_1|s_0, a_1) \prod_{t=2}^T \pi(a_t|s_{t-1}) P(s_t|s_{t-1}, a_t)} \right) \quad (46)$$

$$= \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) \sum_{\tau_{1,T}} \rho_{\pi}(\tau_{1,T}) f \left(\frac{\rho_{\pi^*}(\tau_{1,T})}{\rho_{\pi}(\tau_{1,T})} \right) \quad (47)$$

$$= \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) D_f(\rho_{\pi^*}(\tau_{1,T}), \rho_{\pi}(\tau_{1,T})) \quad (48)$$

We can apply triangle inequality again with respect to $\rho_{\bar{\pi}}(\tau_{1,T})$ to get

$$\begin{aligned} & \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) D_f(\rho_{\pi^*}(\tau_{1,T}), \rho_{\pi}(\tau_{1,T})) \quad (49) \\ & \leq \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) [D_f(\rho_{\pi^*}(\tau_{1,T}), \rho_{\bar{\pi}}(\tau_{1,T})) + D_f(\rho_{\bar{\pi}}(\tau_{1,T}), \rho_{\pi}(\tau_{1,T}))] \quad (50) \end{aligned}$$

$$\leq \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) \left[\sum_{s_1} P(s_1|s_0, a_1) \sum_{a_2} \pi(a_2|s_1) f \left(\frac{\pi^*(a_2|s_1)}{\pi(a_2|s_1)} \right) + D_f(\rho_{\bar{\pi}}(\tau_{1,T}), \rho_{\pi}(\tau_{1,T})) \right] \quad (51)$$

$$\leq \sum_s \rho_{\pi}^0(s) \sum_a \pi(a|s) f \left(\frac{\pi^*(a|s)}{\pi(a|s)} \right) + \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) D_f(\rho_{\bar{\pi}}(\tau_{1,T}), \rho_{\pi}(\tau_{1,T})) \quad (52)$$

$$\begin{aligned} & \leq \mathbb{E}_{s \sim \rho_{\pi}^0(s)} [D_f(\pi^*(a|s), \pi(a|s))] \quad (53) \\ & \quad + \sum_{s_0} \rho_0(s_0) \sum_{a_1} \pi(a_1|s_0) \sum_{s_1} P(s_1|s_0, a_1) \sum_{a_2} \pi(a_2|s_1) D_f(\rho_{\pi^*}(\tau_{2,T}), \rho_{\pi}(\tau_{2,T})) \end{aligned}$$

Again if we continue to expand $D_f(\rho_{\pi^*}(\tau_{2,T}), \rho_{\pi}(\tau_{2,T}))$ and add to (43) we have

$$D_f(\rho_{\pi^*}(\tau_{0,T}), \rho_{\pi}(\tau_{0,T})) \leq \sum_{t=0}^{T-1} \mathbb{E}_{s \sim \rho_{\pi}^t(s)} [D_f(\pi^*(a|s), \pi(a|s))] \quad (54)$$

$$\leq T \mathbb{E}_{s \sim \rho_{\pi}(s)} [D_f(\pi^*(a|s), \pi(a|s))] \quad (55)$$

where (55) follows from $\rho_\pi(s) = \frac{1}{T} \sum_{t=0}^{T-1} \rho_\pi^t(s)$

□

C Existing algorithms as different f-divergence minimization

Behavior Cloning – Kullback-Leibler (KL) divergence. If we use KL divergence $f(u) = u \log(u)$ in our framework’s trajectory matching problem:

$$\begin{aligned}
D_{KL}(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) &= \sum_{\tau} \rho_{\pi^*}(\tau) \log\left(\frac{\rho_{\pi^*}}{\rho_{\pi}}\right) = \sum_{\tau} \rho_{\pi^*}(\tau) \log\left(\prod_t \frac{\pi^*(a_t|s_{t-1})}{\pi(a_t|s_{t-1})}\right) \\
&= \sum_{\tau} \rho_{\pi^*}(\tau) \sum_t \log\left(\frac{\pi^*(a_t|s_{t-1})}{\pi(a_t|s_{t-1})}\right) \\
&= \mathbb{E}_{s \sim \rho_{\pi^*}, a \sim \pi^*} [\log \pi^*(a|s) - \log \pi(a|s)] \\
\hat{\pi} &= \min D_{KL}(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) \\
&= \max \mathbb{E}_{s \sim \rho_{\pi^*}, a \sim \pi^*} (\cdot|s) \log(\pi(a|s))
\end{aligned} \tag{56}$$

Note that this is exactly the behavior cloning [3] objective, which tries to minimize a classification loss under the expert’s state-action distribution. The loss used in (56) is the cross entropy loss for multi-class classification. This optimization is also referred to as *M-projection*. A benefit of this method is that it does not rely on any interactions with the environment; data is provided by the expert.

It’s well known that behavior cloning often leads to covariant shift problem in practice [17]. One explanation is that supervised learning errors compound exponentially in time. We can also view this a side-effect of M-projection which can lead to situations where $\pi(a|s) > 0$, $\pi^*(a|s) = 0$.

Generative Adversarial Imitation Learning (GAIL) [5] – Jensen-Shannon (JS) divergence. Plugging in the JS divergence $f(u) = -(u+1) \log \frac{1+u}{2} + u \log u$ in (6) we have

$$\hat{\pi} = \arg \min_{\pi \in \Pi} \max_w \mathbb{E}_{(s,a) \sim \rho_{\pi^*}} [\log D_w(s, a)] - \mathbb{E}_{(s,a) \sim \rho_{\pi}} [\log(1 - D_w(s, a))] \tag{57}$$

this matches the GAIL objective (without the entropic regularizer). Note that this is minimizing an estimate of the lower bound of JS divergence. While this requires a more expensive minimax optimization procedure, but at least in practice GAIL appears to outperform behavior cloning on a range of simulated environments.

Dataset Aggregation (DAGGER) [17] – Total Variation (TV) distance. If we choose $f(u) = \frac{1}{2} |u - 1|$ in (2), we get the total variation distance $D_{TV}(p, q) = \frac{1}{2} \sum_x |p(x) - q(x)|$. TV satisfies the triangle inequalities and hence can be shown to satisfy the following:

Theorem 4. *The Total Variation distance between trajectory distributions is upper bounded by the expected distance between the action distribution on states induced by π .*

$$D_{\text{TV}}(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) \leq T \mathbb{E}_{s \sim \rho_{\pi}(s)} [D_{\text{TV}}(\pi^*(a|s), \pi(a|s))] \leq T \sqrt{\mathbb{E}_{s \sim \rho_{\pi}(s)} [D_{\text{KL}}(\pi^*(a|s), \pi(a|s))]} \quad (58)$$

Proof. We first apply Theorem 3 on total variation distance. Then by Cauchy-Schwartz inequality we have

$$\mathbb{E}_{s \sim \rho_{\pi}(s)} [D_{\text{TV}}(\pi^*(a|s), \pi(a|s))] \leq \sqrt{\mathbb{E}_{s \sim \rho_{\pi}(s)} [(D_{\text{TV}}(\pi^*(a|s), \pi(a|s)))^2]} \quad (59)$$

Finally by Pinsker's inequality we have

$$(D_{\text{TV}}(\pi^*(a|s), \pi(a|s)))^2 \leq D_{\text{KL}}(\pi^*(a|s), \pi(a|s)) \quad (60)$$

Putting all inequalities together we have the proof. \square

DAGGER solves the following non i.i.d learning problem

$$\begin{aligned} \hat{\pi} &= \arg \min_{\pi \in \Pi} \mathbb{E}_{s \sim \rho_{\pi}(s), a \sim \pi^*(a|s)} [\ell(s, a)] \\ &= \arg \min_{\pi \in \Pi} \mathbb{E}_{s \sim \rho_{\pi}(s), a \sim \pi^*(a|s)} [-\log(\pi(a|s))] \\ &= \arg \min_{\pi \in \Pi} \mathbb{E}_{s \sim \rho_{\pi}(s), a \sim \pi^*(a|s)} [D_{\text{KL}}(\pi^*(a|s), \pi(a|s))] \end{aligned} \quad (61)$$

DAGGER reduces this to an iterative supervised learning problem. Every iteration a classification algorithm is called. Let $\epsilon_N = \min_{\pi \in \Pi} \frac{1}{N} \sum_{i=1}^N \mathbb{E}_{s \sim \rho_{\pi_i}(s)} [D_{\text{KL}}(\pi^*(a|s), \pi(a|s))]$. Let γ_N be the average regret which goes to zero asymptotically, i.e. $\lim_{N \rightarrow \infty} \gamma_N = 0$. DAGGER guarantees that there exists a learnt policy π that satisfies the following bound (infinite sample case):

$$\mathbb{E}_{s \sim \rho_{\pi}(s)} [D_{\text{KL}}(\pi^*(a|s), \pi(a|s))] \leq \epsilon_N + \gamma_N + \mathcal{O}\left(\frac{\log T}{N}\right) \quad (62)$$

Putting all together we have a bound on total variation distance $T \sqrt{\epsilon_N + \gamma_N + \mathcal{O}\left(\frac{\log T}{N}\right)}$.

We highlight some undesirable behaviors in DAGGER. Consider the example shown in Fig. 7. The expert stays on the track and never visits bad states $s \in \mathcal{S}_{\text{bad}}$. The learner, on the other hand, immediately drifts off into \mathcal{S}_{bad} . Moreover, for all $s \in \mathcal{S}_{\text{bad}}$, the learner can perfectly imitate the expert. In other words, $\ell(s, \pi) = 0$ for these states. In fact, it is likely that for certain policy classes, this is the optimal solution! At the very least, DAGGER is susceptible to learn such policies as is shown in the counter example in [22].

This phenomenon can also be explained from the lens of Total Variation distance. TV measures $D_{\text{TV}}(p, q) = \frac{1}{2} \sum_x |p(x) - q(x)|$. This distance does not penalize the learner going off the track as much as RKL. In this case, RKL would have a very high penalization for reasons mentioned in Section 5.

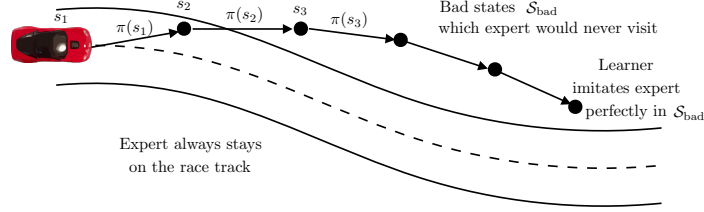


Fig. 7: Problem with the DAGGER formulation. The expert policy stays perfectly on the race track. The learner immediately goes off the race track and visits bad states.

D Reverse KL Divergence via Interactive Variational Imitation

RKL-VIM is an approximation of I-projection that minimizes the lower bound of Reverse KL divergence. There are few things we don't like about it. First, it's a double lower bound – moving to state-action divergence (Theorem 1) and then variational lower bound (5). Second, the optimal state action divergence estimator may require a complex function class. Finally, the min-max optimization (6) may be slow to converge. Interestingly, Reverse KL has a special structure that we can exploit to do even better if we have an interactive expert!

Theorem 5.

$$D_{\text{RKL}}(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) = T \sum_s \rho_{\pi}(s) D_{\text{RKL}}(\pi^*(a|s), \pi(a|s)) \quad (63)$$

which means

$$\sum_{\tau} \rho_{\pi}(\tau) \log \left(\frac{\rho_{\pi}(\tau)}{\rho_{\pi^*}(\tau)} \right) = T \sum_s \rho_{\pi}(s) \sum_a \pi(a|s) \log \left(\frac{\pi(a|s)}{\pi^*(a|s)} \right) \quad (64)$$

Proof. Applying lemma 3 with $\phi(s, a) = \log \left(\frac{\pi(a|s)}{\pi^*(a|s)} \right)$ □

Note that different from Theorem 3, we have strict equality in the above equation. Hence we can directly minimize action distribution divergence. Since this is on states induced by π , this falls under the regime of *interactive learning* [17]. In this regime, we need to query the expert on *states visited by the learner*. Note that this may not always be convenient - the expert is required during training. However, as we will see, it does lead to better estimators.

We now explore variational imitation approaches similar to *RKL-VIM* for minimizing action divergence. We get the following update rule:

$$\hat{\pi} = \arg \min_{\pi \in \Pi} \mathbb{E}_{s \sim \rho_{\pi}} [\mathbb{E}_{a \sim \pi^*(\cdot|s)} [-\exp(V_w(s, a))] + \mathbb{E}_{a \sim \pi(\cdot|s)} [V_w(s, a)]] \quad (65)$$

We call this *interactive variational imitation* (*RKL-iVIM*). The algorithm is described in Algorithm 2. Unlike *RKL-VIM*, we collect a fresh batch of data

from *both* the expert and the learner every iteration. The optimal estimator in this case is $V_{w^*}(s, a) = \log \left(\frac{\pi(a|s)}{\pi^*(a|s)} \right)$. This can be far simpler to estimate than the optimal estimator for state-action divergence $V_{w^*}(s, a) = \log \left(\frac{\rho_{\pi^*}(s)\pi(a|s)}{\rho_{\pi^*}(s)\pi^*(a|s)} \right)$.

Algorithm 2 *RKL- i VIM*

- 1: Initialize learner and estimator parameters θ_0, w_0
 - 2: **for** $i = 0$ **to** $N - 1$ **do**
 - 3: Sample state-action pairs from learner $\tau_i \sim \rho_{\pi_{\theta_i}}$
 - 4: Query expert on all trajectories τ_i to get $a^* \sim \pi^*(\cdot|s)$.
 - 5: Update estimator $w_{i+1} \leftarrow w_i + \eta_w \nabla_w (\mathbb{E}_{s \sim \tau_i} [\mathbb{E}_{a \sim a^*} [-\exp(V_w(s, a))] + \mathbb{E}_{a \sim \tau_i} [V_w(s, a)])$
 - 6: Update policy (using policy gradient) $\theta_{i+1} \leftarrow \theta_i - \eta_{\theta} \mathbb{E}_{(s,a) \sim \tau_i} [\nabla_{\theta} \log \pi_{\theta}(a|s) Q^{V_w}(s, a)]$
 where $Q^{V_w}(s_{t-1}, a_t) = \sum_{i=t}^T V_w(s_{i-1}, a_i)$
 - 7: **end for**
 - 8: **Return** π_{θ_N}
-

E Density Ratio Minimization for Reverse KL via No Regret Online Learning

We continue the argument made in D for better algorithms for reverse KL minimization. Instead of variational lower bound, we can *upper bound* the action divergence as follows:

$$\frac{1}{T} D_{\text{RKL}}(\rho_{\pi^*}(\tau), \rho_{\pi}(\tau)) = \mathbb{E}_{s \sim \rho_{\pi}} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} \left[\log \frac{\pi(a|s)}{\pi^*(a|s)} \right] \right] \leq \mathbb{E}_{s \sim \rho_{\pi}} \left[\mathbb{E}_{a \sim \pi(\cdot|s)} \left[\frac{\pi(a|s)}{\pi^*(a|s)} - 1 \right] \right]$$

where we use the fact that $\log(x) \leq x$ for any $x > 0$. To estimate the conditional density ratio $\pi(a|s)/\pi^*(a|s)$, we can leverage an off-shelf density ratio estimator (DRE) ⁶ as follows. Rather than directly estimating $\pi(a|s)/\pi^*(a|s)$ for all s , we notice that $\pi(a|s)/\pi^*(a|s) = (\rho_{\pi}(s)\pi(a|s)) / (\rho_{\pi}(s)\pi^*(a|s))$. We know how to sample (s, a) from $\rho_{\pi}(s)\pi(a|s)$, and under the interactive setting, we can also sample (s, a) from $\rho_{\pi}(s)\pi^*(a|s)$ by first sampling $s \sim \rho_{\pi}(\cdot)$ and then sample action $a \sim \pi^*(\cdot|s)$, i.e., query expert at state s . Given a dataset $D = \{s, a\} \sim \rho_{\pi}\pi$, and a dataset $D^* = \{s, a^*\} \sim \rho_{\pi}\pi^*$, DRE takes the two datasets and returns an estimator: $\hat{r} = \text{DRE}(D, D^*)$ such that $\hat{r}(s, a) \approx \frac{\rho_{\pi}(s)\pi(a|s)}{\rho_{\pi}(s)\pi^*(a|s)} = \frac{\pi(a|s)}{\pi^*(a|s)}$. Hence, by just using a classic DRE, we can form a conditional density ratio estimator via leveraging the interactive expert.

With the above trick to estimate conditional density ratio estimator, now we are ready to introduce our algorithm (Alg. 3). Our algorithm takes a density ratio estimator (DRE) and a cost-sensitive classifier as input.⁷ At the n -th iteration, it uses the current policy π_n to generate states s , and then collect a dataset $\{s, a\}$ with $a \sim \pi_n(\cdot|s)$, and another dataset $\{s, a^*\}$ with $a^* \sim \pi^*(\cdot|s)$. It then uses DRE to learn a conditional density ratio estimator $\hat{r}_n(s, a) \approx \pi_n(a|s)/\pi^*(a|s)$ for any $(s, a) \in \mathcal{S} \times \mathcal{A}$. It then performs data aggregation by aggregating newly generated state cost vector pairs $\{s, \hat{r}_n(\cdot, s)\}$ to the cost-sensitive classification dataset \mathcal{D} . We then update the policy to π_{n+1} via performing cost-sensitive classification on \mathcal{D} .

Below we provide an agnostic analysis of the performance of the returned policy from Alg. 3. Our analysis is reduction based, in a sense that the performance of the learned policy depends on performance of the off-shelf DRE, the performance of the cost sensitive classifier, and the no-regret learning rate. Similar

⁶ Given two distributions $p(z) \in \Delta(\mathcal{Z})$ and $q(z) \in \Delta(\mathcal{Z})$ over a finite set \mathcal{Z} , the Density Ratio Estimator (DRE) aims to compute an estimator $\hat{r} : \mathcal{Z} \rightarrow \mathbb{R}^+$ such that, $\hat{r}(z) \approx p(z)/q(z)$ (we assume q has no smaller support than p on \mathcal{Z} , i.e., $q(z) = 0$ implies $p(z) = 0$), with access only to two sets of samples $\{z_i\}_{i=1}^N \sim p$ and $\{z'_i\}_{i=1}^N \sim q$. In this work, we treat DRE as a black box that takes two datasets as input, and returns a corresponding ratio estimator: $\hat{r} = \text{DRE}(\{z_i\}_{i=1}^N, \{z'_i\}_{i=1}^N)$. We further assume that DRE achieves small prediction error: $\mathbb{E}_{z \sim q} [|\hat{r}(z) - p(z)/q(z)|] \leq \delta \in \mathbb{R}^+$.

⁷ A cost sensitive classifier \mathcal{C} takes a dataset $\{s, c\}$ with $c \in \mathbb{R}^{|\mathcal{A}|}$ as input, and outputs a classifier that minimizes the classification cost: $\pi = \arg \min_{\pi \in \Pi} \sum_i c_{\pi(s)}$, where we use c_a to denote the entry in the cost vector c that corresponds to action a .

Algorithm 3 Interactive DRE Minimization

```

1: Input: Density Ratio Estimator (DRE)  $\mathcal{R}$ , expert  $\pi^*$ , Cost sensitive classifier  $\mathcal{C}$ 
2: Initialize learner  $\pi_0$ , dataset  $\mathcal{D} = \emptyset$ 
3: for  $n = 0$  to  $N - 1$  do
4:    $s_0 \sim \rho_0$ 
5:   Initialize  $D = \emptyset$ ,  $D^* = \emptyset$ 
6:   for  $e = 0$  to  $E$  do
7:     for  $t = 0$  to  $T - 1$  do
8:       Query expert:  $a_t^* \sim \pi^*(\cdot|s_t)$ 
9:       Execute action  $a_t \sim \pi_n(\cdot|s_t)$  and receive  $s_{t+1}$ 
10:       $D = D \cup \{s_t, a_t\}$ ,  $D^* = D^* \cup \{s_t, a_t^*\}$ 
11:    end for
12:  end for
13:   $\hat{r}_n = \mathcal{R}(D, D^*)$  ▷ Density Ratio Estimation
14:   $\mathcal{D} = \mathcal{D} \cup \{s, \hat{r}_n(\cdot, s)\}_{(s,a) \in D}$  ▷ Data Aggregation
15:   $\pi_{n+1} = \mathcal{C}(\mathcal{D})$  ▷ Cost sensitive classification
16: end for
17: Return  $\pi_{\theta_N}$ 

```

to DAgger [17], note that Alg. 3 can be understood as running Follow-The-Leader (FTL) no-regret online learner on the sequence of loss functions $\hat{\ell}_n(\pi) \triangleq \mathbb{E}_{s \sim \rho_{\pi_n}} [\mathbb{E}_{a \sim \pi(\cdot|s)} [\hat{r}_n(a, s) - 1]]$ for $n \in [0, N - 1]$, which approximate $\ell_n(\pi) \triangleq \mathbb{E}_{s \sim \rho_{\pi_n}} [\mathbb{E}_{a \sim \pi(\cdot|s)} [r_n(s, a) - 1]]$. We denote $\epsilon_{\text{class}} = \min_{\pi \in \Pi} \frac{1}{N} \sum_{i=0}^{N-1} \ell_n(\pi)$ as the minimal cost sensitive classification error one could achieve in hindsight. Note the ϵ_{class} represents the richness of our policy class Π . If Π is rich enough such that $\pi^* \in \Pi$, we have $\epsilon_{\text{class}} = 0$. In general, ϵ_{class} decreases when we increase the representation power of policy class Π . Without loss of generality, we also assume the following black-box DRE oracle \mathcal{R} performance:

$$\max_{n \in [N]} \mathbb{E}_{s \sim \rho_{\pi_n}} [\mathbb{E}_{a \sim \pi^*(\cdot|s)} |\hat{r}_n(s, a) - r_n(s, a)|] \leq \gamma, \quad (66)$$

with $r_n(s, a) = \rho_{\pi_n}(s)\pi_n(a|s)/(\rho_{\pi_n}(s)\pi_n^*(a|s))$ being the true ratio. Note that this is the standard performance guarantee one can get from the theoretical foundation of Density Ratio Estimation. In Appendix E.2 we give an example of DRE with its finite sample performance guarantee in the form of Eq. 66. We also assume that the expert has non-zero probability of trying any action at any state, i.e., $\min_{s,a} \pi^*(a|s) \geq c \in [0, 1]$. We ignore sample complexity here and simply focus on analyzing the quality of the learned policy under the assumption that every round we can draw enough samples to accurately estimate all expectations. Finite sample analysis can be done via standard concentration inequalities.

Theorem 6. *Run Alg. 3 for N iterations. Then there exists a policy $\pi \in \{\pi_0, \dots, \pi_{N-1}\}$ such that*

$$D_{RKL}(\rho_{\pi^*}, \rho_{\pi}) \leq T \left(\left(1 + \frac{1}{c}\right) \gamma + \epsilon_{\text{class}} + \frac{o(N)}{N} \right).$$

The detailed proof is deferred to Appendix E. By definition of $o(N)$, we have $\lim_{N \rightarrow \infty} o(N)/N \rightarrow 0$. The above theorem indicates that as $N \rightarrow \infty$, the inverse KL divergence is upper bounded by $(1 + 1/c)\gamma + \epsilon_{\text{class}}$. Increasing the representation power of Π will decrease ϵ_{class} and any performance improvement the density ratio estimation community could make on DRE can be immediately transferred to a performance improvement of Alg. 3.

E.1 Proof of Theorem 6

Denote the ideal loss function at iteration n as

$$\ell_n(\pi) = \mathbb{E}_{s \sim \rho_{\pi_n}} [\mathbb{E}_{a \sim \pi(\cdot|s)} [r_n(s, a)]] , \quad (67)$$

where $r_n(s, a) = \frac{\pi_n(a|s)}{\pi^*(a|s)} - 1$ (note we include the constant -1 for analysis simplicity). Note that we basically treat $r_n(s, a)$ as a cost of π classifying to action a at state s .

Of course, we do not have a perfect $r_n(s, a)$, as we cannot access π^* 's likelihood. Instead we rely on an off-shelf density ratio estimation (DRE) solver to approximate r_n by \hat{r}_n . We simply assume that the returned \hat{r}_n has the following performance guarantee:

$$\mathbb{E}_{s \sim \rho_{\pi_i}} [\mathbb{E}_{s \sim \pi^*(\cdot|s)} |\hat{r}_i(s, a) - r_i(s, a)|] \leq \gamma. \quad (68)$$

Note that this performance guarantee is well analyzed and is directly delivered by existing off-shelf density ratio estimation algorithms (e.g., [40, 53]). Such γ in general depends on the richness of the function approximator we use to approximate r , and the number of samples we draw from $\rho_{\pi_n}\pi_n$ and $\rho_{\pi_n}\pi^*$, and can be analyzed using classic learning theory tools. In realizable setting (i.e., our hypothesis class contains $r(s, a)$) and in the limit where we have infinitely many samples, γ will be zero. The authors in [40] analysis γ with respect to the number of samples under the realizable assumption.

With \hat{r}_n , let us define $\hat{\ell}_n(\pi)$ that approximates $\ell_n(\pi)$ as follows:

$$\hat{\ell}_n(\pi) = \mathbb{E}_{s \sim \rho_{\pi_n}} [\mathbb{E}_{a \sim \pi(\cdot|s)} [\hat{r}_n(s, a)]] , \quad (69)$$

where we simply replace r_n by \hat{r}_n . Now we bound the difference between $\ell_n(\pi^*)$ and $\hat{\ell}_n(\pi^*)$ using γ (the reason we use π^* inside ℓ and $\hat{\ell}$ will be clear later):

$$\begin{aligned} |\ell_n(\pi^*) - \hat{\ell}_n(\pi^*)| &= |\mathbb{E}_{s \sim \rho_{\pi_n}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} (r_n(s, a) - \hat{r}_n(s, a))| \\ &\leq \mathbb{E}_{s \sim \rho_{\pi_n}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} |r_n(s, a) - \hat{r}_n(s, a)| \leq \gamma, \end{aligned} \quad (70)$$

where we simply applied Jensen inequality.

Note that at this stage, we can see that the Alg. 3 is simply using FTL on a sequence of loss functions $\ell_n(\pi)$ for $n \in [N]$. The no-regret property from FTL immediately gives us the following inequality:

$$\sum_{i=1}^N \hat{\ell}_i(\pi_i) - \min_{\pi \in \Pi} \sum_{i=1}^N \hat{\ell}_i(\pi) \leq o(N). \quad (71)$$

Let us examine the second term on the LHS of the above inequality: $\min_{\pi \in \Pi} \sum_{i=1}^N \hat{\ell}_i(\pi)$:

$$\min_{\pi \in \Pi} \frac{1}{N} \sum_{i=1}^N \hat{\ell}_i(\pi) \leq \frac{1}{N} \sum_{i=1}^N \hat{\ell}_i(\pi^*) \leq \frac{1}{N} \sum_{i=1}^N \ell_i(\pi^*) + \gamma = \epsilon_{\text{class}} + \gamma, \quad (72)$$

where we used inequality 70, and the fact that $\ell_i(\pi^*) = \mathbb{E}_{s \sim d_{\pi_i}} \mathbb{E}_{a \sim \pi^*(\cdot|s)} r_i(s, a) = 0$ (recall we define $r_i(s, a) = \pi_i(a|s)/\pi^*(a|s) - 1$). Hence, we get there exists a least one policy $\hat{\pi}$ among $\{\pi_i\}_{i=1}^N$, such that:

$$\mathbb{E}_{s \sim \rho_{\hat{\pi}}} [\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)} [r(s, a)]] \leq \gamma + o(N)/N, \quad (73)$$

where we denote $r(s, a)$ as the ratio approximator of $\hat{\pi}(a|s)/\pi^*(a|s)$.

We now link $\hat{\ell}_i(\pi_i)$ and $\ell_i(\pi_i)$ as follows:

$$\begin{aligned} |\hat{\ell}_i(\pi_i) - \ell_i(\pi_i)| &\leq \mathbb{E}_{s \sim \rho_{\pi_i}} \mathbb{E}_{a \sim \pi_i} |\hat{r}_i(s, a) - r_i(s, a)| \\ &\leq \left(\max_{s, a} \frac{\pi_i(a|s)}{\pi^*(a|s)} \right) \mathbb{E}_{s \sim \rho_{\pi_i}} \mathbb{E}_{a \sim \pi^*(a|s)} |\hat{r}_i(s, a) - r_i(s, a)| \leq \left(\max_{s, a} \frac{\pi_i(a|s)}{\pi^*(a|s)} \right) \gamma \leq \frac{1}{c} \gamma, \end{aligned} \quad (74)$$

where we assumed $\min_{s, a} \pi^*(a|s) \geq c$, which is a necessary assumption to make $D_{KL}(\rho_{\pi}, \rho_{\pi^*})$ well defined.

Put everything together, we get:

$$D_{RKL}(\rho_{\hat{\pi}}, \rho_{\pi^*}) \leq T \mathbb{E}_{s \sim d_{\hat{\pi}}} [\mathbb{E}_{a \sim \hat{\pi}(\cdot|s)} [r(s, a)]] \leq T \left(\left(1 + \frac{1}{c} \right) \gamma + \epsilon_{\text{class}} + o(N)/N \right). \quad (75)$$

Note the linear dependency on T is not improvable and shows up in the original DAgger as well.

E.2 An Example of Density Ratio Estimation and Its Finite Sample Analysis

We consider the algorithm proposed by Nguyen et al. [40] for density ratio estimation. We also provide finite sample analysis, which is original missing in [40].

We consider the following general density ratio estimation problem. Given a finite set of elements \mathcal{Z} , and two probability distribution over \mathcal{Z} , $p \in \Delta(\mathcal{Z})$ and $q \in \Delta(\mathcal{Z})$. We are interested in estimating the ratio $r(z) = p(z)/q(z)$. The first assumption we use in this section is that $q(z)$ is lower bounded by a constant for any z :

Assumption E1 (Boundness). *There exists a small positive constant $c \in \mathbb{R}^+$, such that for any $z \in \mathcal{Z}$, we always have $q(z) \geq c$ and $p(z) \geq c$.*

Essentially we assume that q has full support over \mathcal{Z} . The above assumption ensures that the ratio is well defined for all $z \in \mathcal{Z}$, and $p(z)/q(z) \in [c, 1/c]$.

Let us assume that we are equipped with a set of function approximators $\mathcal{G} = \{g : \mathcal{Z} \rightarrow [-1/c, 1/c]\}$ with $a, b \in \mathbb{R}^+$ two positive constants. The second assumption is the realizable assumption:

Assumption E2 (Realizability). *We assume that $r(z) \triangleq p(z)/q(z) \in \mathcal{G}$, and \mathcal{G} is discrete, i.e., \mathcal{G} contains finitely many function approximators.*

Note that for analysis simplicity we assume \mathcal{G} is discrete. As we will show later that $|\mathcal{G}|$ is only going to appear inside a log term, hence \mathcal{G} could contain large number of function approximators. Also, our analysis below uses standard uniform convergence analysis with standard concentration inequality (i.e., Hoeffding's inequality), it is standard to relax the above assumption to continuous \mathcal{G} where $\log(|\mathcal{G}|)$ will be replaced by complexity terms such as Rademacher complexity.

Given two sets of samples, $\{x_i\}_{i=1}^N \sim q$, and $\{y_i\}_{i=1}^N \sim p$, we perform the following optimization:

$$\hat{g} = \arg \min_{g \in \mathcal{G}} \frac{1}{N} \sum_{i=1}^N (g(x_i))^2 - \frac{2}{N} \sum_{i=1}^N g(y_i). \quad (76)$$

One example is that when \mathcal{G} belongs to some Reproducing Kernel Hilbert Space \mathcal{H} with some kernel $k : \mathcal{Z} \times \mathcal{Z} \rightarrow \mathbb{R}^+$, then the above optimization has closed form solution.

We are interested in bounding the following risk:

$$\mathbb{E}_{z \sim q} |\hat{g}(z) - r(z)|. \quad (77)$$

Theorem 7. *Fix $\delta \in (0, 1)$. Under assumption E1 and assumption E2, with probability at least $1 - \delta$, DRE in Eq 76 returns an ratio estimator, such that:*

$$\mathbb{E}_{z \sim q} |\hat{g}(z) - r(z)| \leq \frac{2}{c} \sqrt{\log(2|\mathcal{G}|/\delta)} N^{-1/4}.$$

Below we prove several useful lemmas that will lead us to Theorem 7.

Lemma 4. *For any $g \in \mathcal{G}$, we have:*

$$\mathbb{E}_{z \sim q} |g(z) - r(z)| \leq \sqrt{\mathbb{E}_{x \sim q} (g(x) - r(x))^2}.$$

The proof of the above simply uses Jensen's inequality with the fact that \sqrt{x} is a concave function.

Define the true risk as $\ell(g)$:

$$\ell(g) = \mathbb{E}_{z \sim q} [g(z)^2] - 2\mathbb{E}_{z \sim p} [g(z)].$$

Note that by realizability assumption, $r = \arg \min_{g \in \mathcal{G}} \ell(g)$. For a fixed g , denote $v_i(g) = (g(x_i))^2 - 2g(y_i)$. Note that $\mathbb{E}_i[v_i(g)] = \ell(g)$. Also note that $|v_i(g)| \leq 1/c^2$.

Lemma 5 (Uniform Convergence). *With probability at least $1 - \delta$, for all $g \in \mathcal{G}$, we have:*

$$\left| \frac{1}{N} \sum_{i=1}^N v_i(g) - \ell(g) \right| \leq \frac{2}{c^2} \sqrt{\frac{\log(2|\mathcal{G}|/\delta)}{N}}.$$

The above lemma can be easily proved by first applying Hoeffding's inequality on $\sum_{i=1}^N v_i/N$ for a fixed g , then applying union bound over all $g \in \mathcal{G}$.

Lemma 6. *For any $g \in \mathcal{G}$, we have:*

$$\mathbb{E}_{z \sim q}(g(z) - r(z))^2 = \mathbb{E}_{z \sim q}(g(z))^2 - 2\mathbb{E}_{z \sim p}g(z) + \mathbb{E}_{z \sim p}r(z).$$

Proof. From $\mathbb{E}_{z \sim q}(g(z) - r(z))^2$, we complete the square:

$$\begin{aligned} \mathbb{E}_{z \sim q}(g(z) - r(z))^2 &= \mathbb{E}_{z \sim q}g(z)^2 + \mathbb{E}_{z \sim q}r(z)^2 - 2\mathbb{E}_{z \sim q}g(z)r(z) \\ &= \mathbb{E}_{z \sim q}g(z)^2 + \mathbb{E}_{z \sim p}r(z) - 2\mathbb{E}_{z \sim p}g(z). \end{aligned}$$

where we use the fact that that $r(z) = p(z)/q(z)$, and $\mathbb{E}_{z \sim q}r(z)g(z) = \mathbb{E}_{z \sim q}g(z)p(z)/q(z) = \mathbb{E}_{z \sim p}g(z)$. \square

Now we are ready to prove the main theorem.

Proof of Theorem 7. We are going to condition on the event that Lemma 5 being hold. Denote $C_N = (2/c^2)\sqrt{\log(2|\mathcal{G}|/\delta)/N}$, based on Lemma 5, we have:

$$\ell(\hat{g}) \leq \sum_{i=1}^N v_i(\hat{g}) + C_N \leq \sum_{i=1}^N v_i(r) + C_N \leq \ell(r) + 2C_N,$$

where the first and last inequality uses Lemma 5, while the second inequality uses the fact that \hat{g} is the minimizer of $\sum_{i=1}^N v_i(g)$, and the fact that \mathcal{G} is realizable.

Based on Lemma 6, we have:

$$\mathbb{E}_{z \sim q}(\hat{g}(z) - r(z))^2 = \ell(\hat{g}) + \mathbb{E}_{z \sim p}r(z) \leq \ell(r) + 2C_N + \mathbb{E}_{z \sim p}r(z) = \mathbb{E}_{z \sim q}(r(z) - r(z))^2 + 2C_N = 2C_N.$$

Now use Lemma 4, we have:

$$\mathbb{E}_{z \sim q}|\hat{g}(z) - r(z)| \leq \sqrt{\mathbb{E}_{z \sim q}(\hat{g}(z) - r(z))^2} \leq \sqrt{2C_N}.$$

Hence we prove the theorem. \square

F Divide-by-zero Issues with KL and JS

Recall the definition of the f -divergence

$$D_f(p, q) = \sum_x q(x) f\left(\frac{p(x)}{q(x)}\right). \quad (78)$$

The core divide-by-zero issue is best illustrated by considering the setting where we have samples from q yet can exactly evaluate $p(x)$ and $q(x)$. At first glance, it may appear the following estimator

$$\mathbb{E}_{x \sim q} f\left(\frac{p(x)}{q(x)}\right). \quad (79)$$

is not unreasonable. However, the issue here is if $\exists x$ s.t. $p(x) > 0, q(x) = 0$, then depending on f , this divergence may be infinite (in f -divergences, $0 * \infty = \infty$). Yet the estimator will never sample the location where $q(x) = 0$ and thus fail to realize the infinite divergence. This issue is particularly pronounced in KL and JS due to their respective f functions.

G Experimental Details

Environmental Noise For the bandit environment we set control noise $\epsilon_0 = 0.28$, unless otherwise specified. For the grid world we tested with the control noise $\epsilon_1 = 0.14$ and the transitional noise $\epsilon_2 = 0.15$.

Parameterization of policies Both bandit and grid world environment have discrete actions. To transform the discrete action space into a continuous policy space for policy gradient, we consider the following settings. Bandit’s policy is parameterized by one real number θ . Given θ , one can construct a vector $V = [\cos(-\theta - \frac{\pi}{4}), \cos(\theta), \cos(-\theta + \frac{\pi}{4})]$. The probability of executing the discrete actions a, b, c is: $\text{softmax}[A(V + 1)]$ where we set $A = 2.5$ in our experiment. For the grid world, the policy is a matrix θ of size N where N is the number of states. For state i one can construct a vector $V_i = [\cos(0 - \theta), \cos(\pi/2 - \theta), \cos(\pi - \theta), \cos(-\pi/2 - \theta)]$. The probability of executing discrete actions $UP, RIGHT, DOWN, LEFT$ at state i is: $\text{softmax}[A(V_i + 1)]$ where we set $A = 2.5$.