Imitation Learning as f-Divergence Minimization

Liyiming Ke[1], Sanjiban Choudhury[1], Matt Barnes[1], Wen Sun[2], Gilwoo Lee[1], and Siddhartha Srinivasa[1]

[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

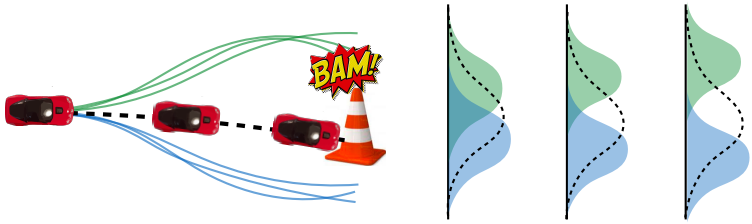
Abstract. We address the problem of imitation learning with multi-modal demonstrations. Instead of attempting to learn all modes, we argue that in many tasks it is sufficient to imitate any one of them. We show that the state-of-the-art methods such as GAIL and behavior cloning, due to their choice of 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.



(a) (b)

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 $\tau = \{s_0, a_1, s_1, \ldots, a_T, s_T\}$. It induces a distribution of trajectories $\rho_\pi(\tau)$ and state $\rho^t_\pi(s)$ as:

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

The average state distribution across time $\rho_\pi(s) = \frac{1}{T} \sum_{t=1}^{T} \rho^{t-1}_\pi(s)$[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:

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

where $f : \mathbb{R}^+ \to \mathbb{R}$ is a convex, lower semi-continuous function. Different choices of f recover different 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

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

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. $\mathcal{O}\left(|\mathcal{S}|^T\right)$. Instead, we can choose to minimize the divergence between the *average state-action distribution* as the following:

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

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.*

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

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 it's *variational form*, i.e. $f(u) = \sup_{t \in \text{dom}_{f^*}}\left(tu - f^*(t)\right)$, where $f^*(\cdot)$ is the convex conjugate [5] Plugging this in the expression for f-divergence (2) we have

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

[5] For a convex function $f(\cdot)$, the convex conjugate is $f^*(v) = \sup_{u \in \text{dom}_f}\left(uv - f(u)\right)$. 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 $\phi : X \to \text{dom}_{f^*}$ 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 $f'(.)$ *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 $V_w : X \to \mathbb{R}$ is an unconstrained discriminator and $g_f : \mathbb{R} \to \text{dom}_{f^*}$ 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

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

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

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

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

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

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	$\frac{1}{2}\lvert u-1\rvert$	t	$\frac{1}{2}\text{sign}(\frac{p(x)}{q(x)} - 1)$	$\frac{1}{2}\tanh(v)$

3. *JS*–VIM: Minimizing Jensen-Shannon divergence

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

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:

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

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 techinques for I-projection, deferring to Appendix D and Appendix E for details.

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

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

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).

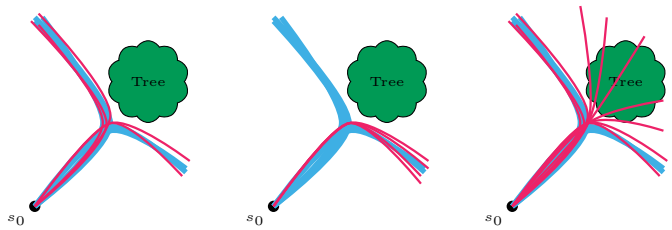
Density ratio minimization via no regret online learning. We first upper bound the action divergence:

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

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)



(a) The ideal policy. (b) RKL mode collapsing (c) KL/JS mode covering

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 are 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

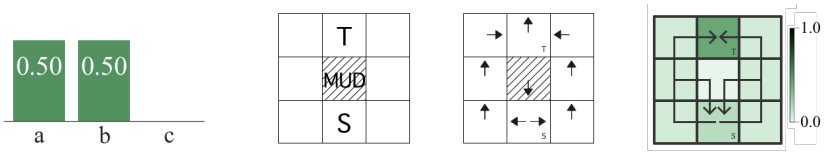
In this section, we empirically validate the following **Hypotheses**:

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.



(a) Bandit expert policy (b) Gridworld (c) Expert policy (d) Rollouts

Fig. 3: Bandit and gridworld environment.

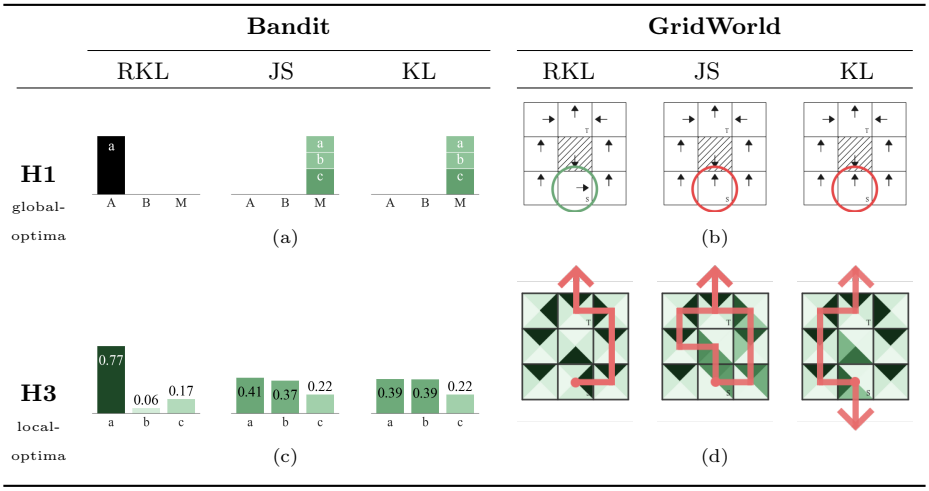
Policy enumeration To test **H1**, we enumerate through all policies in Π, exactly compute their stationary distributions $\rho_\pi(s, a)$, and select the policy with the smallest exact f-divergence, the optimal policy. Our results on the bandit and gridworld (Table 2a and 2b) show that the globally 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.



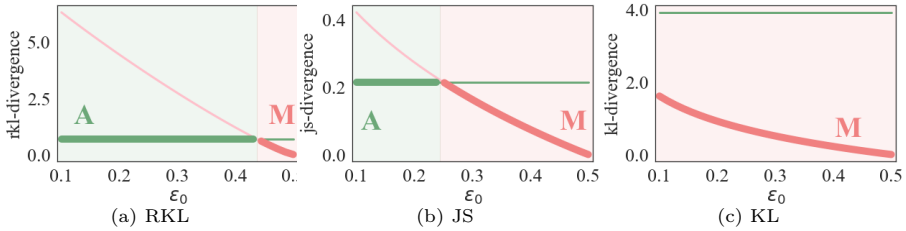


Fig. 4: Divergences and corresponding optimal policy as a function of the control noise ϵ_0. RKL strongly prefers the mode collapse policy A (except at high control noise), KL strongly prefers the mode covering policy M, and JS is between the two.

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-

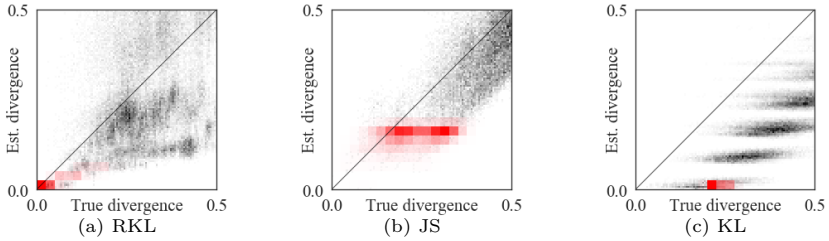


Fig. 5: Comparing f-divergence with the estimated values. Preferred policies under each objective (in the 1 percentile of estimations) are in red. The normalized estimations appear to be typically lower than the normalized true values for JS and KL.

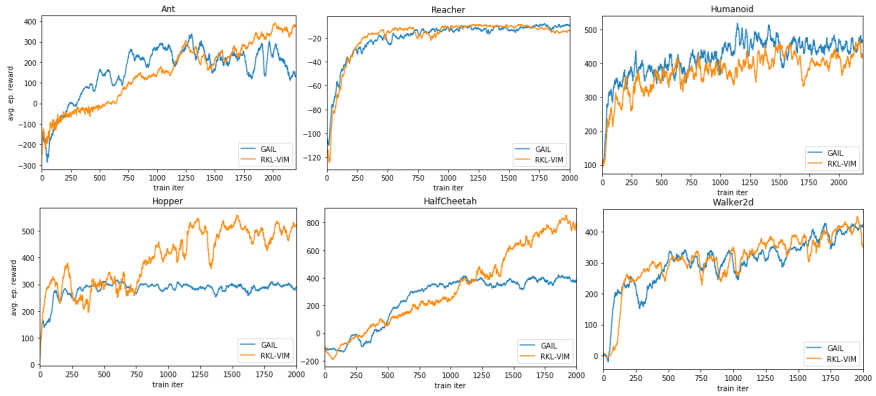


Fig. 6: Training *RKL*–VIM and *JS*–VIM (GAIL) on Mujoco environments.

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 $\mathbb{E}_{(s,a)\sim\rho_\pi}\left[(1-\alpha) - f^*(g_f(V_w(s,a))) + \alpha r(s,a)\right]$ 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 $f'(.)$ 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] Stephane 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 aggrevated: 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. Alvinn: 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 Pokorny, 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. Sfv: 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 Houthooft, 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.