Energy-Based Imitation Learning

Minghuan Liu, Tairan He, Minkai Xu, Weinan Zhang[†]
Shanghai Jiaotong University
Shanghai, China
{minghuanliu,whynot,wnzhang}@sjtu.edu.cn,mkxu@apex.sjtu.edu.cn

ABSTRACT

We tackle a common scenario in imitation learning (IL), where agents try to recover the optimal policy from expert demonstrations without further access to the expert or environment reward signals. Except the simple Behavior Cloning (BC) that adopts supervised learning followed by the problem of compounding error, previous solutions like inverse reinforcement learning (IRL) and recent generative adversarial methods involve a bi-level or alternating optimization for updating the reward function and the policy, suffering from high computational cost and training instability. Inspired by recent progress in energy-based model (EBM), in this paper, we propose a simplified IL framework named Energy-Based Imitation Learning (EBIL). Instead of updating the reward and policy iteratively, EBIL breaks out of the traditional IRL paradigm by a simple and flexible two-stage solution: first estimating the expert energy as the surrogate reward function through score matching, then utilizing such a reward for learning the policy by reinforcement learning algorithms. EBIL combines the idea of both EBM and occupancy measure matching, and via theoretic analysis we reveal that EBIL and Max-Entropy IRL (MaxEnt IRL) approaches are two sides of the same coin, and thus EBIL could be an alternative of adversarial IRL methods. Extensive experiments on qualitative and quantitative evaluations indicate that EBIL is able to recover meaningful and interpretative reward signals while achieving effective and comparable performance against existing algorithms on IL benchmarks.

KEYWORDS

Imitation Learning, Inverse Reinforcement Learning, Energy-Based Modeling

1 INTRODUCTION

Imitation learning (IL) [23] allows Reinforcement Learning (RL) agents to learn from demonstrations, without any further access to the expert or explicit rewards. Classic solutions for IL such as behavior cloning (BC) [32] aim to minimize 1-step deviation error along the provided expert trajectories with supervised learning,

---

† Corresponding author.

which requires an extensive collection of expert data and suffers seriously from compounding error caused by covariate shift [34, 35] due to the long-term trajectories mismatch when we just clone each single-step action. As another solution, Inverse Reinforcement Learning (IRL) [1, 13, 29] tries to recover a reward function from the expert and subsequently train an RL policy under that, yet such a bi-level optimization scheme can result in high computational cost. The recent generative adversarial solution [13, 14, 22] takes advantage of GAN [15] to minimize the divergence between the agent's occupancy measure and the expert's while it also inherits the training instability of GAN [6].

Analogous to IL, learning statistical models from given data and generating similar samples has been an important topic in the generative model community. Among them, recent energy-based models (EBMs) have gained much attention because of the simplicity and flexibility in likelihood estimation [10, 41]. In this paper, we propose to leverage the advantages of EBMs to solve IL with a novel but simplified framework called Energy-Based Imitation Learning (EBIL), solving IL in a two-step fashion: first estimates an unnormalized probability density (*a.k.a.* energy) of expert's occupancy measure through score matching, then takes the energy to construct a surrogate reward function as a guidance for the agent to learn the desired policy. We realize that EBIL is high related to MaxEnt IRL, and in detail analyze their relation, which reveals that these two methods are two sides of the same coin, and MaxEnt IRL can be seen as a special form of EBIL because MaxEnt IRL estimates the energy and the policy alternately. Therefore, we can think of EBIL as an simplified alternative of adversarial IRL methods.

In experiments, we first verify the effectiveness of EBIL and the meaningful reward recovered in a simple one-dimensional environment by visualizing the estimated reward and the induced policy; then we evaluate our algorithm on extensive high-dimensional continuous control benchmarks, contains sub-optimal expert demonstrations and optimal demonstrations, showing that EBIL can achieve comparable and stable performance against previous adversarial IRL methods. We also show the functionality for resolving state-only imitation learning by matching the target state marginal distribution, and provide evaluations and ablation study on the recovered EBM.

The remainder of this paper is organized as follows. In Section 3, we first present the formulation of the energy-based imitation learning (EBIL) framework, which is a general and principled two-stage RL framework that models the expert policy with an unnormalized probability function (*i.e.*, energy function). In Section 4, we give a comprehensive discussion about EBIL and classic IRL methods, which are also built upon the energy-based formulation to model the expert trajectories but typically adopt an adversarial (or alternating) training scheme. The discussion allows us to clarify how to

avoid the interactive training of IRL and thus leads to our simplified and principled two-stage algorithm. After that, in Section 5, we practically illustrate how to implement the two-stage energy-based algorithm via score matching method. Lastly, in Section 7, through comprehensive experiments from synthetic domain to continuous control tasks, we demonstrate the interpretability and effectiveness of EBIL over existing algorithms.

2 BACKGROUND

We consider a Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, \rho_0, r, \gamma \rangle$, where $\mathcal{S}$ is the set of states, $\mathcal{A}$ represents the action space of the agent, $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability distribution, $\rho_0 : \mathcal{S} \to [0, 1]$ is the distribution of the initial state s_0, and $\gamma \in [0, 1]$ is the discounted factor. The agent holds its policy $\pi(a|s) : \mathcal{S} \times \mathcal{A} \to [0, 1]$ to make decisions and receive rewards defined as $r : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. For an arbitrary function $f : \langle s, a \rangle \to \mathbb{R}$, we denote the expectation w.r.t. the policy π as $\mathbb{E}_\pi[f(s, a)] \triangleq \mathbb{E}_{s_0 \sim \rho_0, s_t \sim P, a_t \sim \pi} \left[\sum_{t=0}^\infty \gamma^t f(s_t, a_t) \right]$. The objective of Maximum Entropy Reinforcement Learning (MaxEnt RL) is required to find a stochastic policy that can maximize its reward along with the entropy [19, 22] as:

$$\pi^* = \arg\max_\pi \mathbb{E}_\pi \left[r(s, a) \right] + \alpha H(\pi) , \qquad (1)$$

where $H(\pi) \triangleq \mathbb{E}_\pi[-\log \pi(a|s)]$ is the γ-discounted causal entropy [2] and α is the temperature hyperparameter. Throughout this work we denote the occupancy measure $\rho_\pi^{s,a}(s, a)$ or $\rho_\pi^s(s)$ as the density of occurrence of states or state-action pairs[1]:

$$\rho_\pi^{s,a}(s, a) = \sum_{t=0}^\infty \gamma^t P(s_t = s, a_t = a | \pi)$$
$$= \pi(a|s) \sum_{t=0}^\infty \gamma^t P(s_t = s | \pi) = \pi(a|s) \rho_\pi^s(s) , \qquad (2)$$

which allows us to write $\mathbb{E}_\pi[\cdot] = \sum_{s,a} \rho_\pi^{s,a}(s, a)[\cdot] = \mathbb{E}_{(s,a) \sim \rho_\pi^{s,a}}[\cdot]$. For simplicity, we will denote $\rho_\pi^{s,a}$ as ρ_π without further explanation, and $\rho_\pi \in \mathcal{D} \triangleq \{\rho_\pi : \pi \in \Pi\}$.

General Imitation Learning. Imitation learning (IL) [23] studies the task of Learning from Demonstrations (LfD), which aims to learn a policy from expert demonstrations. The expert demonstrations typically consist of the expert trajectories interacted with environments without any reward signals. General IL objective tries to minimize the policy distance or the occupancy measure distance[2]:

$$\pi^* = \arg\min_\pi \mathbb{E}_{s \sim \rho_\pi^s} \left[\ell \left(\pi_E(\cdot|s), \pi(\cdot|s) \right) \right] = \ell \left(\rho_{\pi_E}, \rho_\pi \right) , \quad (3)$$

where ℓ denotes some distance metric. However, as we do not ask the expert agent for further demonstrations, it is always hard to optimize Eq. (3) with only expert trajectories accessible. Thus, Behavior Cloning (BC) [32] provides a straightforward method by

learning the policy in a supervised way, where the objective is represented as a Maximum Likelihood Estimation (MLE):

$$\hat{\pi}^* = \arg\min_\pi \mathbb{E}_{s \sim \rho_{\pi_E}^s} \left[\ell \left(\pi_E(\cdot|s), \pi(\cdot|s) \right) \right] , \qquad (4)$$

which suffers from covariate shift problem [22] for the i.i.d. state assumption.

Maximum-Entropy Inverse Reinforcement Learning. Another branch of methods is Inverse Reinforcement Learning (IRL) [29] that tries to recover the reward function r^* in the environment, with underlying assumption that expert policy is optimal under some reward function r^*. A typically maximum-entropy (MaxEnt) solution [48] models the expert trajectories with a Boltzmann distribution[3]:

$$p_\theta(\tau) = \frac{1}{Z} \exp \left(r_\theta(\tau) \right) , \qquad (5)$$

where $\tau = \{s_1, a_1, \ldots, s_t, a_t\}$ is a trajectory with horizon t, $r_\theta(\tau) = \sum_t \gamma^t r_\theta(s_t, a_t)$ is a parameterized reward function, and the partition function $Z \triangleq \int \exp \left(r_\theta(\tau) \right) d\tau$ is the integral over all trajectories. IRL algorithms suffer from the computational challenge in estimating the partition function Z and the bi-level optimization by alternating between updating the cost function and an optimal policy w.r.t the current cost function with RL.

Derived from MaxEntIRL, Generative Adversarial Imitation Learning (GAIL) [22] shows that the objective of MaxEntIRL is a dual problem of occupancy measure matching, and thus can be solved through generative models such as GAN. Specifically, it shows that the policy learned by RL on the reward recovered by IRL can be characterized by

$$\text{RL} \circ \text{IRL}_\psi(\pi_E) = \arg\min_\pi -H(\pi) + \psi^*(\rho_\pi - \rho_{\pi_E}) , \qquad (6)$$

where ψ is the regularizer, and $f^* : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \overline{\mathbb{R}}$ is the convex conjugate for an arbitrary function $f : \mathbb{R}^{\mathcal{S} \times \mathcal{A}} \to \overline{\mathbb{R}}$ given by $f^*(x) = \sup_{y \in \mathbb{R}^{\mathcal{S} \times \mathcal{A}}} x^T y - f(y)$. Eq. (6) shows that various settings of ψ can be seen as a distance metric leading to various solutions of IL. For example, in [22] they choose a special form of ψ so that the second term becomes minimizing the JS divergence and [14] further shows the second term can be any f-divergence measure of ρ and ρ_E.

Energy-Based Models. The energy term is originally borrowed from statistical physics, where it is employed for describing the distribution of the atoms or molecules. Low-energy states correspond to the high probability of occurring, i.e., the local minima of this function are usually related to the stable stationary states. Such a property is appropriate for modeling the density of a data distribution, where high-density should be assigned with lower energy, and higher energy otherwise. Formally, for a random variable $X \sim p(x)$, an Energy-Based Model (EBM) [27] builds the density of data by estimating the energy function $E(x)$ with sample x as a Boltzmann distribution:

$$p(x) = \frac{1}{Z} \exp(-E(x)) , \qquad (7)$$

---

[1]It is important to note that the definition of occupancy measure is not equivalent to the definition of a normalized distribution since in RL we have to deal with the discounted factor for the expectation w.r.t. the policy $\mathbb{E}_\pi$.

[2]The equivalence of these two objective usually can be easily shown by the one-to-one correspondence of π and ρ and the convexity of ℓ.

[3]Note that Eq. (5) is formulated under the deterministic MDP setting. A general form for stochastic MDP is derived in [47, 48] yet owns similar analysis: the probability of a trajectory is decomposed as the product of conditional probabilities of the states s_t, which can factor out all likelihood ratios since they are not affected by the reward function.

where $Z = \int \exp(-E(x)) \, dx$ is the partition function, which is normally intractable to compute exactly for high-dimensional x. As shown in Fig. 1, the energy function E can be seen as the unnormalized log-density of data which is always optimized to maximize the likelihood of the data. Therefore, we can model the expert demonstrations (state-action pairs) with the energy function, where low energy corresponds to the state-action pairs that the expert mostly perform. Typically, the estimation of the partition function Z is computationally expensive, which requires sampling from the Boltzmann distribution $p(x)$ within the inner loop of learning. However, many researchers have shown easier ways to estimate energy with much more efficiency without estimating the partition function.
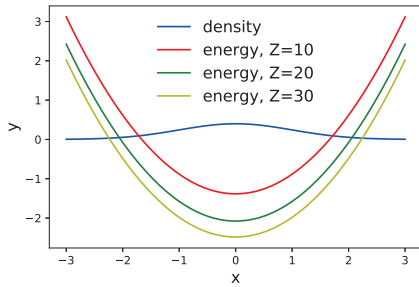
Figure 1: An example of the density and the energy. The blue line represent the density of $x \sim \mathcal{N}(0, 1)$, and the other lines denote the energy of different partition function Z.

3 ENERGY-BASED IMITATION LEARNING

We begin by discussing a specific form of IL objectives. Since the different choices of ψ in Eq. (6) lead to different distance metrics to solve IL, and by any kind of f-divergence $D_f(\rho_{\pi_E} \| \rho_\pi)$ [14] corresponds to one kind of form of ψ, we can let $\psi = \mathbb{E}_{\pi_E}[-1 - \log(r(s, a)) + r(s, a)]$ and we can have a reverse KL divergence objective[4]:

$$\min_\pi \psi^*(\rho_\pi - \rho_{\pi_E}) = D_f(\rho_{\pi_E} \| \rho_\pi) = D_{KL}(\rho_\pi \| \rho_{\pi_E}) \quad (8)$$

Before continuing, we first consider to model the normalized expert occupancy measure ρ_{π_E} with Boltzmann distribution using an EBM $E_{\pi_E}(s, a)$:

$$(1 - \gamma)\rho_{\pi_E}(s, a) = \frac{1}{Z} \exp(-E_{\pi_E}(s, a)) . \quad (9)$$

Then we take Eq. (9) into Eq. (8) and manipulate trivial algebraic deviations[5]:

$$D_{KL}(\rho_\pi \| \rho_{\pi_E}) = \sum_{s,a} \rho_\pi \left(\log \rho_\pi(s, a) - \log \rho_{\pi_E}(s, a) \right)$$
$$= \mathbb{E}_\pi \left[E_{\pi_E}(s, a) \right] - H(\pi) + \text{const} , \quad (10)$$

where E_{π_E} is the EBM of policy π_E. Therefore, a minimization objective on KL divergence $D_{KL}(\rho_\pi \| \rho_{\pi_E})$ is equivalent as:

$$\pi^* = \arg\max_\pi \mathbb{E}_\pi \left[-E_{\pi_E}(s, a) \right] + H(\pi) , \quad (11)$$

which is exactly the objective of the MaxEnt RL (Eq. (1)). It is worth noting that if we remove the entropy term to construct a standard

[4]Full derivations can be found in Appendix C and D of [14] and without loss of generality we replace $c(s, a)$ with $-r(s, a)$.
[5]Full deviations can be found in Appendix A.1.

RL objective, then it will collapse into minimizing the cross entropy of the occupancy measure rather than the reverse KL divergence.

Such an observation essentially shows us a two-stage imitation learning solution. In the first stage called *energy recovery*, we try to estimate the energy of the expert's occupancy measure, and in the second stage we can utilize the recovered energy as a fixed surrogate reward for any *MaxEnt RL* algorithm. It is worth noting that these learning procedures can be fully separated, i.e., the training of the energy and the policy do not have to be alternate! We call such a formulation as Energy-Based Imitation Learning (EBIL) framework which liberates us from designing complex learning algorithms but focus on the estimation of the expert energy $E_{\pi_E}(s, a)$, which is free to take recent advances in energy-based modeling.

Except the way that utilizing the energy as the reward for reinforcement learning, one may notice that there is a simpler way to recover the expert policy directly via Gibbs distribution:

$$\pi^*(a|s) = \frac{\rho_{\pi_E}(s, a)}{\sum_{a'} \rho_{\pi_E}(s, a')} = \frac{\frac{1}{Z} \exp(-E_{\pi_E}(s, a))}{\frac{1}{Z} \sum_{a'} \exp\left(-E_{\pi_E}(s, a')\right)}$$
$$= \frac{\exp(-E_{\pi_E}(s, a))}{\sum_{a'} \exp\left(-E_{\pi_E}(s, a')\right)} . \quad (12)$$

This indicates that if we accurately estimate the energy of the expert's occupancy measure, we have recovered the expert policy. However, this may be hard to generalize to continuous or high-dimensional action space but it worth to try in simple discrete domains.

State-only Imitation Learning. It is also easy to derive an energy based algorithm for state-only imitation learning [14, 28, 43], where agents can only get the states (or observations) of the expert, by optimizing the following objective:

$$\pi^* = \arg\min_\pi D_{KL}(\rho(s) \| \rho_{\pi_E}(s)) . \quad (13)$$

Modeling the normalized state occupancy using Boltzmann distribution and following similar deviations, we can get an equivalent objective as Eq. (11) by replacing the energy of state-action $E(s, a)$ with the energy of states $E(s)$.

In practice, we train the EBM from expert demonstrations as the reward function. Specifically, instead of directly using the estimated energy function $E(s, a)$, we can construct a surrogate reward function $\hat{r}(s, a) = h(-E_{\pi_E}(s, a))$, where $h(x)$ is a monotonically increasing linear function, which do not change the optimality in Eq. (11) and can be specified differently for various environments. The step-by-step EBIL algorithm is presented in Algo. 1, which is simple and straightforward.

4 RELATION TO INVERSE REINFORCEMENT LEARNING

As shown in Eq. (5), MaxEnt IRL [48] also models the trajectory distribution in an energy form, which reminds us to analyze the relation between EBIL and IRL. Recent remarkable works focus on solving the IRL problem in an adversarial style [12, 13, 22] motivated by the progress in Generative Adversarial Nets (GANs) [15], which alternately optimize the reward function and a maximum-entropy policy corresponding to that reward. In this section, we aims to understand why previous solutions need a alternating training

between the policy and the reward, and what to do if we want a two-stage algorithm. We conclude that MaxEnt IRL can be seen as a special form of EBIL which alternately estimates the energy and EBIL can be a simplified alternative to adversarial IRL methods.

Related theoretical connections among IRL, GANs have been thoroughly discussed in work by Finn et al. [11], which can be bridged by EBMs. We borrow the insight from Finn et al. [11] and figure out the following questions:

Q1: Is the adversarial (or alternating) style necessary for IRL?
Q2: If unnecessary, how can we avoid adversarial training?
Q3: Is the adversarial style superior than other methods?

To answer Q1, we need to first identify the reason where the alternating training comes from. As revealed by Finn et al. [11], the connections between MaxEnt IRL and GAN is derived from Guided Cost Learning (GCL) [12], a sampling based method of IRL. Typically, to solve the cost function in Eq. (5), a maximum likelihood loss function is utilized:

$$\mathcal{L}_{\text{cost}}(\theta) = \mathbb{E}_{\tau \sim p}[-\log p_\theta(\tau)] = \mathbb{E}_{\tau \sim p}[c_\theta(\tau)] + \log Z . \quad (14)$$

Notice that such a loss function needs to estimate the partition function Z, which is always hard for high-dimensional data. To that end, Finn et al. [11] proposed to train a new sampling distribution $q(\tau)$ and estimated the partition function Z via importance sampling $\mathbb{E}_{\tau \sim q}\left[\frac{\exp(-c_\theta(\tau))}{q(\tau)}\right]$. In fact, the sampling distribution $p(\tau)$ corresponds to the agent policy, which is optimized to minimize the KL divergence between $q(\tau)$ and the $p_\theta(\tau)$:

$$\mathcal{L}_{\text{sampler}}(q) = D_{\text{KL}}(q(\tau)\|p_\theta(\tau)) = \mathbb{E}_{\tau \sim q}[c_\theta(\tau) + \log q(\tau)] . \quad (15)$$

One may notice that these two optimization problems depend on each other, and thus lead to an alternating training scheme. However, a serious problem comes from the high variance of the importance sampling ratio and Finn et al. [12] applied a mixture distribution to alleviate this problem. Another different solution comes from Ho and Ermon [22], which utilizes the typical unconstrained form of the discriminator without using the generator density, but the optimization of the discriminator is corresponding to the optimization of the cost function in Eq. (14), as Finn et al. [11] proved. Therefore, we understand that the alternating IL algorithm suffers from the interdependence between the cost (reward) function and the policy due to the choice of maximum likelihood objective for the cost function. In this point of view, MaxEnt IRL can be seen as a special form of EBIL which estimates the energy alternately according to the following proposition.

Proposition 1. *Suppose that we have recovered the optimal cost function $\hat{c}^*$, then minimizing the distance between trajectories is equivalent to minimizing the distance of occupancy measures:*

$$\arg\min_\pi D_{KL}(p(\tau)\|p(\tau_E)) = \arg\min_\pi D_{KL}(\rho(s, a)\|\rho_{\pi_E}(s, a)) . \quad (16)$$

The proof is straightforward and we leave it in Appendix A.2 for detailed checking.

Therefore, the answer to Q1 is definitely not and with sufficient analysis we get our conclusion to Q2: we can avoid alternating training if we can decouple the dependence between the optimization of the reward $r(s, a)$ and the policy $\pi(a|s)$. Specifically, if we can *learn the energy $E_{\pi_E}(s, a)$ without estimating the partition function Z*, or

estimate the partition without using a learnable sampling policy, we are free to change the alternative training style into a two-stage procedure. Therefore, EBIL and MaxEnt IRL are actually two sides of the same coin, and EBIL can be thought as an simplified alternative to adversarial IRL methods, which benefits from the flexibility and simplicity of EBMs.

For Q3, we suggest the readers to refer to researches on probabilistic modeling [17, 27]. Although many recent works [4, 15] depend on an adversarially learned sampler, there are no theoretical guarantees showing that such an alternating learning style is more advanced than other streams of methods. Therefore, we can not theoretically get an answer to Q3 so leave the answer of Q3 in our quantitative and qualitative experiments where we compare our two-stage EBIL algorithms with those former adversarial IRL methods on various tasks.

5 EXPERT ENERGY ESTIMATION VIA SCORE MATCHING

As described above, we desire a two-stage energy-based imitation learning algorithm, where the estimation of the energy function can be decoupled from learning the policy. In particular, the estimation of energy can either be done without estimating the partition function, or learned with an additional policy to approximate the partition, which is referred as adversarial IRL methods. However, although various related work lie in the domain of energy based statistic modeling, they are not easy to be used for EBIL since many of them do not estimate the energy value $E(x)$ itself. For example, although the branch of score matching methods [42, 44] defines the score as the gradient of the energy, i.e., $\nabla_x \log p(x) = -\nabla_x E(x)$, estimate $-\nabla_x E(x)$ is enough for statistic modeling. To that end, in this paper, we refer to a recent denoising score matching work (e.g., DEEN [38]) that directly estimates the energy value through deep neural network in a differentiable framework.

Formally, let the random variable $X = (s, a) \sim \rho_{\pi_E}(s, a)$. Let the random variable Y be the noisy observation of X that $y \sim x + N(0, \sigma^2 I)$, i.e., y is derived from samples x by adding with white Gaussian noise $\xi \sim N(0, \sigma^2 I)$. The empirical Bayes least square estimator, i.e., the optimal denoising function $g(y)$ for the white Gaussian noise, is solved as

$$g(y) = y + \sigma^2 \nabla_y \log p(y) . \quad (17)$$

Solving such a problem, we can get the score function $\nabla_y \log p(y)$. But remember we need the energy function instead, therefore a parameterized energy function $E_\theta(y)$ is modeled by a neural network explicitly. As shown in [38], such a DEEN framework can be trained by minimizing the following objective:

$$\arg\min_\theta \sum_{x_i \in X, y_i \in Y} \left\| x_i - y_i + \sigma^2 \frac{\partial E_\theta(y = y_i)}{\partial y} \right\|^2 , \quad (18)$$

which ensures the relation of score function $\partial E_\theta(y)/\partial y$ shown in Eq. (17). It is worth noting that the EBM estimates the energy of the noisy samples. This can be seen as a Parzen window estimation of $p(x)$ with variance σ^2 as the smoothing parameter [37, 44]. A trivial problem here is that Eq. (18) requires the samples (state-action pairs) to be continuous so that the gradient can be accurately computed. Practically, such a problem can be solved via state/action

embedding or using other energy estimation methods, *e.g.*, Noise Contrastive Estimation [17].

In practice, we learn the EBM of expert data from offline demonstrations and construct the reward function, which will be fixed until the end to help agent learn its policy with a normal RL procedure. Specifically, we construct the surrogate reward function $\hat{r}(s, a)$ as follows:

$$\hat{r}(s, a) = h(-E_{\pi_E}(s, a)) , \qquad (19)$$

where $h(x)$ is a monotonically increasing linear function, which can be specified for different environments. Formally, the overall EBIL algorithm is presented in Algo. 1.

Algorithm 1 Energy-Based Imitation Learning

1: **Input:** Expert demonstration data $\tau_E = \{(s_i, a_i)\}_{i=1}^N$, parameterized energy-based model E_ϕ, parameterized policy π_θ;
2: **for** $k = 0, 1, 2, \ldots$ **do**
3:     Optimize ϕ with the objective in Eq. (18).
4: **end for**
    Compute the surrogate reward function $\hat{r}$ via Eq. (19).
5: **for** $k = 0, 1, 2, \ldots$ **do**
6:     Update θ with a normal RL procedure using the surrogate reward function $\hat{r}$.
7: **end for**
8: **return** π

---

6 RELATED WORK

6.1 Imitation Learning

As extensions for the traditional solution as inverse reinforcement learning [1, 13, 29], generative adversarial algorithms have been raised up since years ago. Tracing back to GAIL, which models the imitation learning as an occupancy measure matching problem, and takes a GAN-form objective to optimize both the reward and the policy [22]. After that, Adversarial Inverse Reinforcement Learning (AIRL) simplifies the idea of Finn et al. [11] and use a disentangled discriminator to recover the reward function in an energy-form [13]. However, as we show in experiments, they do not actually recover the energy. Based on previous works, Ke et al. [24] and Ghasemipour et al. [14] concurrently proposed to unify the adversarial learning algorithms with f-divergence. Like Nowozin et al. [31], they claimed that any f-divergence can be used to construct a generative adversarial imitation learning algorithm. Specifically, Ghasemipour et al. [14] proposed FAIRL, which adopts the forward KL as the distance metric.

Instead of seeking to alternatively update the policy and the reward function as in IRL and GAIL, many recent works of IL aim to learn a fixed reward function directly from expert demonstrations and then apply a normal reinforcement learning procedure with that reward function. This idea can be found inherently in Generative Moment Matching Imitation Learning (GMMIL) [25] that utilizes the maximum mean discrepancy as the distance metric to guide training. Recently, Wang et al. [45] proposed Random Expert Distillation (RED), which employs the idea of Random Network Distillation[7] to compute the reward function by the loss of fitting a random neural network or an auto-encoder. Reddy et al.

[33] applied constant rewards by setting positive rewards for the expert state-actions and zero rewards for other ones, which is optimized with the off-policy Soft Q-Learning (SQL) algorithm [20]. In addition, Disagreement-Regularized Imitation Learning (DRIL) [5] constructs the reward function using the disagreement in their predictions of an ensemble of policies trained on the demonstration data, which is optimized together with a supervised behavioral cloning cost. These works resembles the idea of EBIL, where the reward function can be seen as approximated energy functions. For example, RED [45] estimates the reward using an auto-encoder, which is exactly the way of energy modeling in EBGAN [46]. In addition, Wang et al. [45] and Brantley et al. [5] also utilized the prediction errors, which are low on demonstration data but high on data that is out of the demonstration (similar to energy). The method of Reddy et al. [33] is more straightforward, which simply sets the energy of the expert as 0 and the agent's as 1. However, these approximated energies are not derived from statistical modeling and lack theoretical correctness. It is worth noting that combining the idea of EBGAN [46] and GAIL [22] we can also design an adversarial style energy-based algorithm for imitation learning, however it is less interesting and also does not have statistical support about the learned energy.

6.2 Energy-Based Modeling

Our EBIL relies highly on EBMs, which have played an important role in a wide range of tasks including image modeling, trajectory modeling and continual online learning [9]. Thanks to the appealing features, EBMs have been introduced into many RL literature, for instance, parameterized as value function [36], employed in the actor-critic framework [21], applied to MaxEnt RL [19] and model-based RL regularization [3]. However, EBMs are always difficult to train due to the partition function [11]. Nevertheless, recent works have tackled the problem of training large-scale EBMs on high-dimensional data, such as DEEN [38] which is applied in our implementation. Except for DEEN, there still leave plenty of choices for efficiently training EBMs [9, 18, 30].

7 EXPERIMENTS

In this section we seek to empirically evaluate EBIL to figure out the effectiveness of our solution compared to previous works, especially the alternative optimization methods. We first conduct qualitative and quantitative experiments on a simple one-dimensional environment, where we illustrate the recovered reward in the whole state-action space and show the training stability of EBIL. Then, we test EBIL against baselines on benchmark environments using sub-optimal experts and release competitive performance.

7.1 Synthetic Task

In the synthetic task, we want to evaluate the qualitative performance of different IL methods by displaying the heat map of the learned reward signals and sampled trajectories. We want EBIL to be capable of guiding the agent to recover the expert policy and correspondingly generate the high-quality trajectories. Therefore, we evaluate EBIL on a synthetic environment where the agent tries to move in a one-dimensional space. Specifically, the state
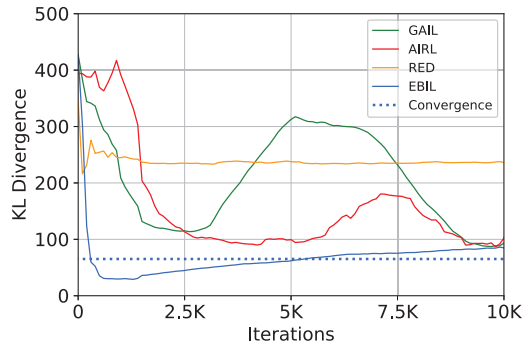
Figure 2: The KL divergence between the agent trajectories and the expert during the learning procedure, which indicates that EBIL is much more stable than the other methods. The blue dash line represents the converged result of EBIL.

space is $[-0.5, 10.5]$ and the action space is $[-1, 1]$. The environment initializes the state at 0, and we set the expert policy as static rule policies $\pi_E = \mathcal{N}(0.25, 0.06)$ when the state $s \in [-0.5, 5)$, and $\pi_E = \mathcal{N}(0.75, 0.06)$ when $s \in [5, 10.5]$. The sampled expert demonstration contains 40 trajectories with up to 30 timesteps in each one. For all methods, we choose Soft Actor Critic (SAC) [20] as the learning algorithm and we continue training each algorithm until convergence.

In this illustrative experiment, we compare EBIL aginst GAIL [22], AIRL [13] and RED [45], where GAIL and AIRL are two representative works of adversarial imitation learning which take an alternative updating on the reward and the policy. RED resembles EBIL which also take a two-stage training by first estimate the reward through the prediction error of a trained network with a randomized one. We plot the KL divergence between the agent's and the expert's trajectories during the training procedure in Fig. 2 and visualize the final estimated rewards with corresponding induced trajectories in Fig. 3.

As illustrated in Fig. 3(b) and Fig. 2, the pre-estimated reward of EBIL successfully captures the density of the expert trajectories, and led by the interpretative reward the induced policy is able to quickly converge to the expert policy. By contrast, GAIL requires a noisy adversarial process to correct the policy. As a result, although GAIL achieves compatible final performance against EBIL (Fig. 3(c)), it suffers a slow, unstable training as shown in Fig. 2 and assigns meaningless reward in the state-action space, as revealed in Fig. 3(c). In addition, we are surprised to find that AIRL recovers an 'inverse'-type reward signals but still learns a good policy, as suggested in Fig. 3(d). We analyze such a problem in Appendix C.2, where we conclude that AIRL actually does not recover the energy of experts but the energy with an entropy term. Removing such an entropy term makes the result more reasonable. Finally, we notice that constructing the reward as the prediction errors does not help to recover a meaningful signal, as shown in Fig. 3(e), where RED suffers from the diverged reward and fails to imitate the expert accurately.

On the contrary, EBIL benefits from meaningful rewards which is tend to learn a deterministic policy which shows the most ood. Thus, the recovered policy has less variance but the mean is very close to the expert, which is a good property. In fact, an optimal

policy is usually a deterministic policy due to Bellman equation and a stochastic policy is always used to increase the exploration ability in a learning procedure. the exploration ability is much necesarry for methods as EBIL and RED. Although it is convincing to take the recovered energy as a meaningful reward signal for imitation learning, we must point out that as illustrated in Fig. 3(b), the recovered reward maybe be sparse in the space where the expert scarcely comes by. Thus, to gain better performance, the agents will be required to equip with an efficient RL algorithm with good exploration ability.

7.2 Imitation on Continous Control Benchmarks

Sub-optimal Demonstrations. We conduct evaluation experiments on the standard continuous control benchmarking MuJoCo environments. For each task, we train experts with OpenAI baselines version [8] of Proximal Policy Optimization (PPO) [40] and get sub-optimal demonstrations. We employ Trust Region Policy Optimization (TRPO) [39] as the learning algorithm in the implementation for all evaluated methods, and we do not apply BC initialization for all tasks. We consider 4 demonstrated trajectories by the sub-optimal expert, as [22, 45] do. We compare EBIL with several baseline methods and report the *converged* result in Tab. 1, which are evaluated throughout 50 test episodes.

As shown in Tab. 1, with sub-optimal demonstrations, EBIL can achieve the best or comparable performance among all environments, indicating that the recovered energy is able to become a good reward signal for imitating the experts. However, learning from such a sub-optimal expert seems challenging for GAIL and AIRL, which are less robust to reach a better performance. We notice that on some environment as InvertedDoublePendulum, GAIL and AIRL can work well during the training, but does not converge until the end. We think the problem is due to the training instability of GAN, which always need an early stop to get a better performance. On the contrary, EBIL provide steady reward signals instead of a alternative trained one that stabilizes the training. The performance of RED and GMMIL also indicates that the recovered reward by random distillation or the maximum mean discrepancy do not provide a stable guidance in such an sub-optimal setting. However, benefit from such a steady reward function, RED and GMMIL can perform better than GAIL and AIRL on some tasks.

Optimal Demonstrations. We also want to know how EBIL performs on optimal demonstrations compared with previous methods, especially adversarial inverse reinforcement learning methods such as GAIL, AIRL, and FAIRL. Therefore we evaluate EBIL on optimal demonstrations from [14], where the expert is trained by SAC. As before, the demonstration for each task contains 4 trajectories and each trajectory is subsampled by a factor of 20. Similar to [14], we finetune each model and checkpoint the model at its *best* validation loss and report the best resulting checkpoints on 50 test episodes. As shown in Tab. 2, we find that those adversarial algorithms can always achieve better performances on high-dimensional tasks as Hopper and Walker2d, where EBIL remains a gap between these methods. We think this problem is due to the accuracy of the energy model trained by DEEN, which only takes a simple MLP without further regularization operations. We must admit that benefiting from

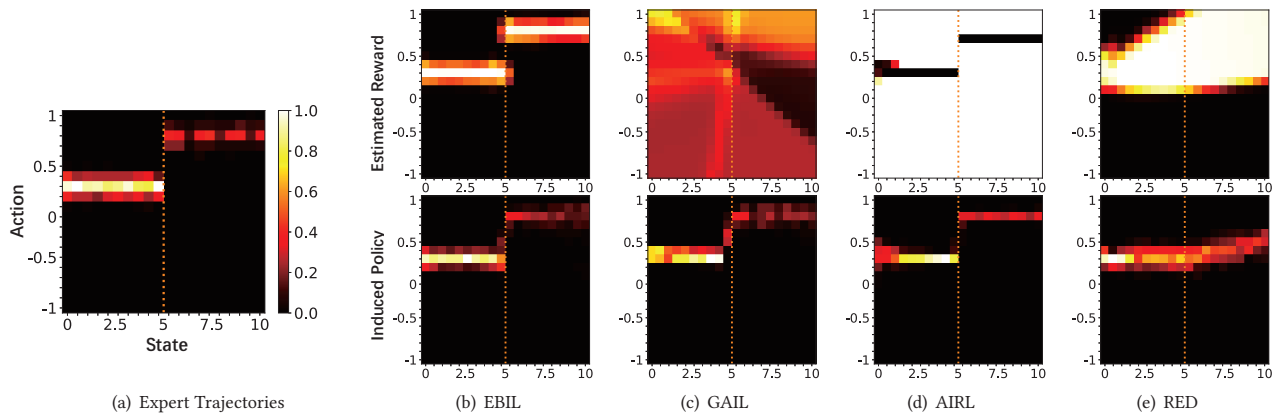(a) Expert Trajectories (b) EBIL (c) GAIL (d) AIRL (e) RED

Figure 3: Heat maps of the expert trajectories (leftmost), the (final) *estimated rewards* recovered by different methods (top) and the corresponding *induced policy* (bottom). The horizontal and the vertical axis denote the *state space* and the *action space* respectively. The red dotted line represents the position where the agent should change its policy. It is worth noting that EBIL and RED both learn fixed reward functions, while GAIL and AIRL iteratively update the reward signals. We do not compare BC since it learns the policy via supervised learning without recovering any reward signals.

Table 1: Comparison for different methods of the episodic true rewards with sub-optimal demonstrations on 5 continuous control benchmarks. The means and the standard deviations are evaluated over different random seeds.

Method	Humanoid	Hopper	Walker2d	Swimmer	InvertedDoublePendulum
Random	100.38 ± 28.25	14.21 ± 11.20	0.18 ± 4.35	0.89 ± 10.96	49.57 ± 16.88
BC	178.74 ± 55.88	28.04 ± 2.73	312.04 ± 83.83	5.93 ± 16.77	138.81 ± 39.99
GAIL	299.52 ± 81.52	$\mathbf{1673.32 \pm 57.13}$	329.01 ± 211.84	23.79 ± 21.84	327.42 ± 94.92
AIRL	286.63 ± 6.05	126.92 ± 62.39	215.79 ± 23.04	-13.44 ± 2.69	76.78 ± 19.63
GMMIL	416.83 ± 59.46	1000.87 ± 0.87	1585.91 ± 575.72	-0.73 ± 3.28	4244.63 ± 3228.14
RED	140.23 ± 19.10	641.08 ± 2.24	641.13 ± 2.75	-3.55 ± 5.05	6400.19 ± 4302.03
EBIL	$\mathbf{472.22 \pm 107.72}$	1040.99 ± 0.53	$\mathbf{2334.55 \pm 633.91}$	$\mathbf{58.09 \pm 2.03}$	$\mathbf{8988.37 \pm 1812.76}$
Expert (PPO)	1515.36 ± 683.59	1407.36 ± 176.91	2637.27 ± 1757.72	122.09 ± 2.60	6129.10 ± 3491.47

the various improvements of GAN such as gradient penalty [16] that makes the training more stable, adversarial algorithms has advantages over the traditional statistical modeling of EBMs, especially in high-dimensional space, since they can continue to improve the learning of data distribution in the iterative training procedure, while score matching methods as DEEN only model the provided dataset from scratch. However, on an easier environment LunarLander where the EBM can provide many meaningful and dense rewards, EBIL outperforms the others while all the three adversarial algorithms have large variances.
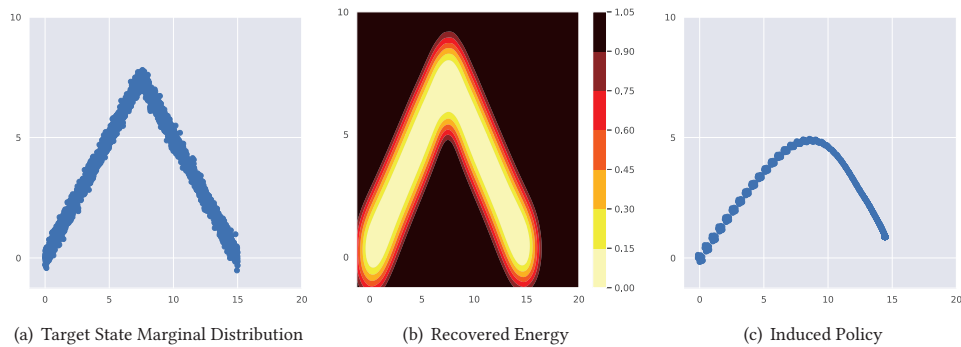
7.3 State Marginal Matching

In this section, we show that EBIL can also be effective in matching the state marginal distributions for a given dataset. Motivated by Ghasemipour et al. [14], we try to make the agent learn the desired policy without expert demonstrations. Specifically, unlike the traditional IL setting, the target state marginal distribution does not even have to be a realizable state-marginal distribution collected by expensive expert demonstrations, but easy-to-get heuristically-designed interpretable distributions. Therefore, we test our methods

Table 2: Comparison for different methods of the episodic true rewards with optimal demonstrations on 3 continuous control benchmarks. The means and the standard deviations are evaluated over different random seeds. We notice that EBIL can work better than adversarial inverse reinforcement learning algorithms on low-dimensional environments as LunarLander but remains performance gaps on harder tasks as Hopper and Walker2d.

Method	LunarLander	Hopper	Walker2d
Random	-232.81 ± 139.72	14.21 ± 11.20	0.18 ± 4.35
GAIL	-85.85 ± 59.22	3117.50 ± 2.96	4092.86 ± 7.46
AIRL	-66.07 ± 104.14	$\mathbf{3398.72 \pm 8.39}$	3987.23 ± 334.04
FAIRL	-116.90 ± 15.79	3353.78 ± 9.12	$\mathbf{4225.66 \pm 65.11}$
EBIL	$\mathbf{237.95 \pm 44.96}$	2401.93 ± 6.85	3026.60 ± 57.14
Expert	254.90 ± 24.11	3285.92 ± 2.14	4807.22 ± 166.35

(a) Target State Marginal Distribution (b) Recovered Energy (c) Induced Policy

Figure 4: (a) Heuristically-designed samples from target state marginal distribution. (b) The recovered energy by EBIL, representing the density of the demonstrations, where the darker color represent the lower energy and higher rewards. (c) The induced policy using the state-only reward constructed from the energy as shown in (b). The energy value is in a range of $[0, 1]$ **since we use *sigmoid* for the last layer of the DEEN network on this task.**
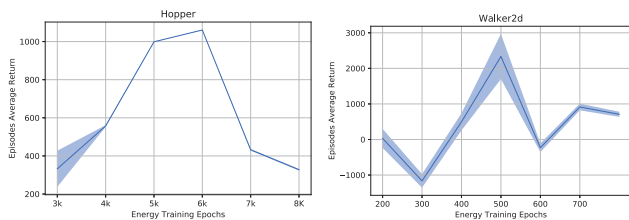
Figure 5: Ablation study on the average episode rewards evaluated on Hopper and Walker2 that are learned with energy models from different training epochs using sub-optimal demonstrations, where the solid line and the shade represent the mean and the standard deviation on different runs separately.

in a 2D point mass task similar to [14], where the agent can move around the ground. We first generate 4,000 heuristic samples as shown in Fig. 4(a), learn the energy from them and then train the policy guided by the energy.

As expected, Fig. 4(b) illustrates that DEEN successfully recovers the meaningful reward function, which describes the density of the target state marginal distribution. We then construct the reward function as $r(s) = h(-E(s))$ where h is a monotonic linear function, with which the agent is finally able to induce similar trajectories to the expert. It is worth noting that we shape the recovered reward with an encourage on the exploration on the x-axis, since we find that with the state-only reward, the agent tends to get stuck, i.e., learn to go forth and back at the very beginning near the starting point. Such a task reminds us of another possibility of finishing reinforcement learning tasks, where people do not have to carefully design the reward function but the samples on the desired trajectories. And by learning from the recovered energy of the provided samples with simple intuitive shaping that prevents the agent from cycling around, the agent can finally learn to complete the task.

7.4 Ablation Study

It is worth noting that in our experiments, we also find that a well-trained energy network may be hard for agents to learn the expert

policies on some environments. We regard it as the "sparse reward signals" problem, as discussed in Section 7.1. By contrast, sometimes a "half-trained" model may provide smoother rewards, and can help the agent to learn more efficiently. Actually, a similar phenomenon also occurs when training the discriminator in GAN. Therefore, to further understand the effect of the energy model from different training epochs, we conduct an ablation study on energy models trained from different epochs. The results are illustrated in Fig. 5, which verifies our intuition that a "half-trained" model can provide smoother rewards that solve the "sparse reward" problem, which is better for imitating the expert policy.

8 CONCLUSION

In this paper, we propose Energy-Based Imitation Learning (EBIL), which shows that it is feasible to compute a fixed reward function via directly estimating the expert's energy to help agents learn from the demonstrations, which breaks out of the traditional iterative IRL paradigm. We further discuss the relation of our method with Maximum Entropy Inverse Reinforcement Learning (MaxEnt IRL) and reveal that these methods are actually two sides of the same coin, where EBIL can be regarded as a simplified alternative of adversarial IRL methods. We conduct quantitative and qualitative evaluations in multiple tasks, showing that with recovering meaningful rewards, EBIL can finally lead to comparable performance against previous algorithms. For future work, we can try different energy estimation methods for expert demonstrations, and exploring more possibilities to utilize EBM to help the agent to learn in different reinforcement learning tasks as state-only learning. It is also intriguing to expand EBIL into multi-agent environments and construct interpretative reward signals for multi-agent learning.

REFERENCES

[1] Pieter Abbeel and Andrew Y Ng. 2004. Apprenticeship learning via inverse reinforcement learning. In *Proceedings of the twenty-first international conference on Machine learning*. 1.

[2] Michael Bloem and Nicholas Bambos. 2014. Infinite time horizon maximum causal entropy inverse reinforcement learning. In *53rd IEEE Conference on Decision and Control*. IEEE, 4911–4916.

[3] Rinu Boney, Juho Kannala, and Alexander Ilin. 2019. Regularizing Model-Based Planning with Energy-Based Models. *arXiv preprint arXiv:1910.05527* (2019).

[4] Avishek Joey Bose, Huan Ling, and Yanshuai Cao. 2018. Adversarial contrastive estimation. *arXiv preprint arXiv:1805.03642* (2018).

[5] Kiante Brantley, Wen Sun, and Mikael Henaff. 2020. Disagreement-Regularized Imitation Learning. In *International Conference on Learning Representations*. https://openreview.net/forum?id=rkgbYyHtwB

[6] Andrew Brock, Jeff Donahue, and Karen Simonyan. 2018. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096* (2018).

[7] Yuri Burda, Harrison Edwards, Amos J. Storkey, and Oleg Klimov. 2018. Exploration by Random Network Distillation. *CoRR* abs/1810.12894 (2018). arXiv:1810.12894 http://arxiv.org/abs/1810.12894

[8] Prafulla Dhariwal, Christopher Hesse, Oleg Klimov, Alex Nichol, Matthias Plappert, Alec Radford, John Schulman, Szymon Sidor, Yuhuai Wu, and Peter Zhokhov. 2016. Openai baselines (2017). *URL https://github.com/openai/baselines* (2016).

[9] Yilun Du and Igor Mordatch. 2019. Implicit Generation and Generalization in Energy-Based Models. *CoRR* abs/1903.08689 (2019). arXiv:1903.08689 http://arxiv.org/abs/1903.08689

[10] Yilun Du and Igor Mordatch. 2019. Implicit Generation and Modeling with Energy Based Models. In *Advances in Neural Information Processing Systems*. 3603–3613.

[11] Chelsea Finn, Paul Christiano, Pieter Abbeel, and Sergey Levine. 2016. A connection between generative adversarial networks, inverse reinforcement learning, and energy-based models. *arXiv preprint arXiv:1611.03852* (2016).

[12] Chelsea Finn, Sergey Levine, and Pieter Abbeel. 2016. Guided cost learning: Deep inverse optimal control via policy optimization. In *International Conference on Machine Learning*. 49–58.

[13] Justin Fu, Katie Luo, and Sergey Levine. 2017. Learning robust rewards with adversarial inverse reinforcement learning. *arXiv preprint arXiv:1710.11248* (2017).

[14] Seyed Kamyar Seyed Ghasemipour, Richard Zemel, and Shixiang Gu. 2019. A Divergence Minimization Perspective on Imitation Learning Methods. *arXiv preprint arXiv:1911.02256* (2019).

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. 2672–2680.

[16] Ishaan Gulrajani, Faruk Ahmed, Martin Arjovsky, Vincent Dumoulin, and Aaron C Courville. 2017. Improved training of wasserstein gans. In *Advances in neural information processing systems*. 5767–5777.

[17] Michael Gutmann and Aapo Hyvärinen. 2010. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*. 297–304.

[18] Michael U Gutmann and Aapo Hyvärinen. 2012. Noise-contrastive estimation of unnormalized statistical models, with applications to natural image statistics. *Journal of Machine Learning Research* 13, Feb (2012), 307–361.

[19] Tuomas Haarnoja, Haoran Tang, Pieter Abbeel, and Sergey Levine. 2017. Reinforcement learning with deep energy-based policies. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org, 1352–1361.

[20] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. 2018. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. *arXiv preprint arXiv:1801.01290* (2018).

[21] Nicolas Heess, David Silver, and Yee Whye Teh. 2012. Actor-Critic Reinforcement Learning with Energy-Based Policies.. In *EWRL*. 43–58.

[22] Jonathan Ho and Stefano Ermon. 2016. Generative adversarial imitation learning. In *Advances in neural information processing systems*. 4565–4573.

[23] Ahmed Hussein, Mohamed Medhat Gaber, Eyad Elyan, and Chrisina Jayne. 2017. Imitation learning: A survey of learning methods. *ACM Computing Surveys (CSUR)* 50, 2 (2017), 21.

[24] Liyiming Ke, Matt Barnes, Wen Sun, Gilwoo Lee, Sanjiban Choudhury, and Siddhartha Srinivasa. 2019. Imitation Learning as f-Divergence Minimization.

[25] Kee-Eung Kim and Hyun Soo Park. 2018. Imitation learning via kernel mean embedding. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[26] Ilya Kostrikov, Kumar Krishna Agrawal, Debidatta Dwibedi, Sergey Levine, and Jonathan Tompson. 2018. Discriminator-actor-critic: Addressing sample inefficiency and reward bias in adversarial imitation learning. *arXiv preprint arXiv:1809.02925* (2018).

[27] Yann LeCun, Sumit Chopra, Raia Hadsell, M Ranzato, and F Huang. 2006. A tutorial on energy-based learning. *Predicting structured data* 1, 0 (2006).

[28] Fangchen Liu, Zhan Ling, Tongzhou Mu, and Hao Su. 2019. State Alignment-based Imitation Learning. *arXiv preprint arXiv:1911.10947* (2019).

[29] Andrew Y Ng, Stuart J Russell, et al. 2000. Algorithms for inverse reinforcement learning.. In *Icml*, Vol. 1. 2.

[30] Erik Nijkamp, Mitch Hill, Song-Chun Zhu, and Ying Nian Wu. 2019. Learning Non-Convergent Non-Persistent Short-Run MCMC Toward Energy-Based Model. arXiv:1904.09770 [stat.ML]

[31] Sebastian Nowozin, Botond Cseke, and Ryota Tomioka. 2016. f-gan: Training generative neural samplers using variational divergence minimization. In *Advances in neural information processing systems*. 271–279.

[32] Dean A Pomerleau. 1991. Efficient Training of Artificial Neural Networks for Autonomous Navigation. *Neural Computation* 3, 1 (1991), 88–97.

[33] Siddharth Reddy, Anca D. Dragan, and Sergey Levine. 2019. SQIL: Imitation Learning via Regularized Behavioral Cloning. *CoRR* abs/1905.11108 (2019). arXiv:1905.11108 http://arxiv.org/abs/1905.11108

[34] Stéphane Ross and Drew Bagnell. 2010. Efficient reductions for imitation learning. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*. 661–668.

[35] Stéphane Ross, Geoffrey Gordon, and Drew Bagnell. 2011. A reduction of imitation learning and structured prediction to no-regret online learning. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*. 627–635.

[36] Brian Sallans and Geoffrey E Hinton. 2004. Reinforcement learning with factored states and actions. *Journal of Machine Learning Research* 5, Aug (2004), 1063–1088.

[37] Saeed Saremi and Aapo Hyvarinen. 2019. Neural Empirical Bayes. *arXiv preprint arXiv:1903.02334* (2019).

[38] Saeed Saremi, Arash Mehrjou, Bernhard Schölkopf, and Aapo Hyvärinen. 2018. Deep energy estimator networks. *arXiv preprint arXiv:1805.08306* (2018).

[39] John Schulman, Sergey Levine, Pieter Abbeel, Michael Jordan, and Philipp Moritz. 2015. Trust region policy optimization. In *International conference on machine learning*. 1889–1897.

[40] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347* (2017).

[41] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. 2019. Sliced score matching: A scalable approach to density and score estimation. *arXiv preprint arXiv:1905.07088* (2019).

[42] Yang Song, Sahaj Garg, Jiaxin Shi, and Stefano Ermon. 2020. Sliced score matching: A scalable approach to density and score estimation. In *Uncertainty in Artificial Intelligence*. PMLR, 574–584.

[43] Wen Sun, Anirudh Vemula, Byron Boots, and J Andrew Bagnell. 2019. Provably efficient imitation learning from observation alone. *arXiv preprint arXiv:1905.10948* (2019).

[44] Pascal Vincent. 2011. A connection between score matching and denoising autoencoders. *Neural computation* 23, 7 (2011), 1661–1674.

[45] Ruohan Wang, Carlo Ciliberto, Pierluigi Vito Amadori, and Yiannis Demiris. 2019. Random Expert Distillation: Imitation Learning via Expert Policy Support Estimation. *CoRR* abs/1905.06750 (2019). arXiv:1905.06750 http://arxiv.org/abs/1905.06750

[46] Junbo Zhao, Michael Mathieu, and Yann LeCun. 2016. Energy-based generative adversarial network. *arXiv preprint arXiv:1609.03126* (2016).

[47] Brian D Ziebart. 2010. *Modeling purposeful adaptive behavior with the principle of maximum causal entropy*. Ph.D. Dissertation. figshare.

[48] Brian D Ziebart, Andrew L Maas, J Andrew Bagnell, and Anind K Dey. 2008. Maximum entropy inverse reinforcement learning.. In *Aaai*, Vol. 8. Chicago, IL, USA, 1433–1438.