Variational Imitation Learning with Diverse-quality Demonstrations

Voot Tangkaratt [1] **Bo Han** [2 1] **Mohammad Emtiyaz Khan** [1] **Masashi Sugiyama** [1 3]

Abstract

Learning from demonstrations can be challenging when the quality of demonstrations is diverse, and even more so when the quality is unknown and there is no additional information to estimate the quality. We propose a new method for imitation learning in such scenarios. We show that simple quality-estimation approaches might fail due to compounding error, and fix this issue by jointly estimating both the quality and reward using a variational approach. Our method is easy to implement within reinforcement-learning frameworks and also achieves state-of-the-art performance on continuous-control benchmarks. Our work enables scalable and data-efficient imitation learning under more realistic settings than before.

1. Introduction

Sequential decision making aims to learn a good policy that makes good decisions (Puterman, 1994). Imitation Learning (IL) is a specific case which learns such a policy from demonstrations (Schaal, 1999), and it performs well when high-quality demonstrations from experts are available (Ho & Ermon, 2016; Fu et al., 2018; Peng et al., 2019). However, in reality, the quality of demonstrations can be diverse, i.e., high- and low-quality demonstrations are mixed. This scenario typically happens when collecting demonstrations from experts is costly, e.g., in robotics where experts must have domain-specific knowledge (Mandlekar et al., 2018; Osa et al., 2018). Unfortunately, learning from diverse-quality demonstrations is challenging, because low-quality demonstrations often negatively affect learning performance, e.g., in robotics where they may cause damages to robots (Shiarlis et al., 2016). In this paper, we propose a new method to solve this learning problem under an assumption that diversity is caused by *noise-densities*.

Learning from diverse-quality demonstrations becomes less challenging when the quality of demonstrations is known. In such scenarios, we can use data-cleaning techniques to remove low-quality demonstrations (Han et al., 2011), or use *multi-modal* approaches to learn good policies that correspond to high-quality demonstrations (Li et al., 2017; Wang et al., 2017). In some scenarios, experts may not provide the quality directly. Instead, they provide additional information about the quality. With such information, learning is still relatively easy, since the quality can be estimated by their confidence scores (Wu et al., 2019), ranking scores (Brown et al., 2019), or a small number of high-quality demonstrations (Audiffren et al., 2015). However, these scenarios assume the availability of experts who provide the quality or additional information. Our goal in this paper is to go beyond these scenarios and perform IL under a more realistic setting where experts are not required.

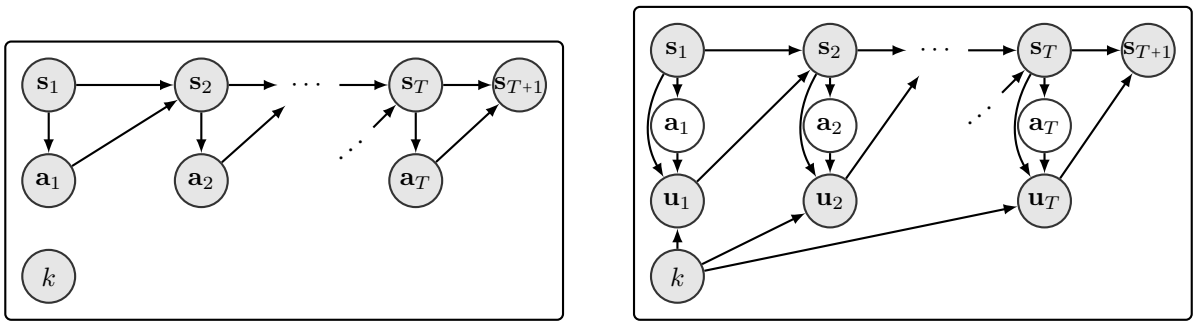
We propose a new method for IL with diverse-quality demonstrations by modeling the quality with a probabilistic graphic model under a noise-density assumption. We show that simple quality-estimation approaches might fail due to compounding error, and fix this issue by estimating the quality along with a reward function that represents an intention of experts' decision making. To handle large state-action spaces, we use a variational approach, which can be easily implemented within reinforcement-learning frameworks (Sutton & Barto, 1998) and is scalable to large state-action spaces by using neural networks. We also propose importance sampling to improve the data-efficiency of our method. The final method is called Variational IL with Diverse-quality demonstrations (VILD). Experiments on continuous-control tasks demonstrate that VILD is robust against diverse-quality demonstrations and achieves state-of-the-art performance.

2. IL with Diverse-quality Demonstrations

Before delving into our main contribution, we first give backgrounds about RL and IL. Then, we formulate a new setting in IL called *IL with diverse-quality demonstrations* and discuss deficiencies of existing methods.

[1]RIKEN Center for Advanced Intelligence Project, Japan [2]Department of Computer Science, Hong Kong Baptist University, Hong Kong [3]Department of Complexity Science and Engineering, The University of Tokyo, Japan. Correspondence to: Voot Tangkaratt <voot.tangkaratt@riken.jp>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.



(a) Expert demonstrations (b) Diverse-quality demonstrations

Figure 1. Graphical models describe $p^\star(\boldsymbol{\tau}_{\rm sa}, k)$ and $p_{\rm d}(\boldsymbol{\tau}_{\rm su}, k)$ of expert demonstrations and diverse-quality demonstrations, respectively. Shaded and unshaded nodes indicate observed and unobserved random variables, respectively. $\mathbf{s}_t \in \mathcal{S}$ is a state with transition densities $p_{\rm s}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$, $\mathbf{a}_t \in \mathcal{A}$ is an action with density $\pi^\star(\mathbf{a}_t|\mathbf{s}_t)$, $\mathbf{u}_t \in \mathcal{A}$ is a noisy action with density $p_{\rm n}(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k)$, and $k \in \{1, \dots, K\}$ is an identification number with distribution $\nu(k)$. Actions $\mathbf{a}_t$ are unobserved in Figure 1(b) because they are not executed in the MDP.

2.1. Reinforcement Learning

Reinforcement Learning (RL) (Sutton & Barto, 1998) aims to learn an optimal policy of a Markov decision process (MDP) (Puterman, 1994). We consider a finite-horizon continuous MDP defined by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, p_{\rm s}(\mathbf{s}'|\mathbf{s}, \mathbf{a}), \mu(\mathbf{s}_1), r(\mathbf{s}, \mathbf{a}))$ with a state $\mathbf{s}_t \in \mathcal{S} \subseteq \mathbb{R}^{d_{\mathbf{s}}}$, an action $\mathbf{a}_t \in \mathcal{A} \subseteq \mathbb{R}^{d_{\mathbf{a}}}$, a transition probability density $p_{\rm s}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)$, an initial state density $\mu(\mathbf{s}_1)$, and a reward function $r : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$, where the subscript $t \in \{1, \dots, T\}$ denotes the time step. We denote $\boldsymbol{\tau}_{\rm sa} = (\mathbf{s}_{1:T+1}, \mathbf{a}_{1:T})$ a (finite-horizon) trajectory of $\mathbf{s}_t$ and $\mathbf{a}_t$. A decision making of an agent is determined by a policy $\pi(\mathbf{a}_t|\mathbf{s}_t)$, which is a conditional probability density of action given state. RL seeks for an optimal policy $\pi^\star(\mathbf{a}_t|\mathbf{s}_t)$ which maximizes the expected cumulative reward: $\mathbb{E}_{p_\pi}[\Sigma_{t=1}^T r(\mathbf{s}_t, \mathbf{a}_t)]$, where $p_\pi(\boldsymbol{\tau}_{\rm sa}) = \mu(\mathbf{s}_1)\Pi_{t=1}^T p_{\rm s}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)\pi(\mathbf{a}_t|\mathbf{s}_t)$ is a trajectory probability density induced by π. A major limitation of RL is that it relies on the reward function which may be unavailable in practice (Schaal, 1999).

2.2. Imitation Learning

Imitation Learning (IL) was proposed to address the above limitation of RL (Schaal, 1999; Ng & Russell, 2000). IL aims to learn the optimal policy from demonstrations that encode information about the optimal policy, without using the reward function r of the MDP. A common setting of most IL methods is the setting of *IL with expert demonstrations*. Namely, demonstrations are collected by $K \geqslant 1$ demonstrators who execute actions $\mathbf{a}_t$ drawn from $\pi^\star(\mathbf{a}_t|\mathbf{s}_t)$ for every states $\mathbf{s}_t$. A graphical model describing this data collection process is depicted in Figure 1(a), where a random variable $k \in \{1, \dots, K\}$ denotes each demonstrator's identification number and $\nu(k)$ denotes the probability of collecting a demonstration from the k-th demonstrator. Under this assumption, expert demonstrations $\{(\boldsymbol{\tau}_{\rm sa}, k)_n\}_{n=1}^N$ are regarded to be drawn independently from

$$p^\star(\boldsymbol{\tau}_{\rm sa}, k) = \nu(k)\mu(\mathbf{s}_1)\prod_{t=1}^T p_{\rm s}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)\pi^\star(\mathbf{a}_t|\mathbf{s}_t). \quad (1)$$

Note that k can be omitted since k and $\boldsymbol{\tau}_{\rm sa}$ are independent.

IL has shown to work well in benchmark tasks (Ho & Ermon, 2016; Fu et al., 2018; Peng et al., 2019), but it has been rarely used in practice (Silver et al., 2012; Schroecker et al., 2019). One of the main reasons is that most methods assume the availability of high-quality demonstrations collected from experts according to Eq. (1). In practice, high-quality demonstrations are often too costly, and even when we obtain them, the number of demonstrations is often too few to accurately learn the optimal policy (Osa et al., 2018).

2.3. Diverse-quality Demonstrations

To make IL more practical, we consider *IL with diverse-quality demonstrations*, where demonstrations are collected from demonstrators with different level of expertise. Such demonstrations can be obtained cheaply via crowdsourcing (Mandlekar et al., 2018), but learning the optimal policy from them is challenging, as will be discussed below.

In this paper, we consider the following *noise-density assumption* of diverse-quality demonstrations. Namely, we assume that at each time step t, demonstrators execute noisy action $\mathbf{u}_t \sim p_{\rm n}(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k)$ where $\mathbf{u}_t \in \mathcal{A}$, instead of action $\mathbf{a}_t \sim \pi^\star(\mathbf{a}_t|\mathbf{s}_t)$. A graphical model describing this process is depicted in Figure 1(b). Under this assumption, diverse-quality demonstrations $\{(\boldsymbol{\tau}_{\rm su}, k)_n\}_{n=1}^N$ are regarded to be drawn from a probability density

$$p_{\rm d}(\boldsymbol{\tau}_{\rm su}, k) = \nu(k)\mu(\mathbf{s}_1)\prod_{t=1}^T p_{\rm s}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{u}_t) \times \int_{\mathcal{A}} \pi^\star(\mathbf{a}_t|\mathbf{s}_t)p_{\rm n}(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k)\mathrm{d}\mathbf{a}_t, \quad (2)$$

where $\boldsymbol{\tau}_{\mathrm{su}} = (\mathbf{s}_{1:T+1}, \mathbf{u}_{1:T})$ is a trajectory of $\mathbf{s}_t$ and $\mathbf{u}_t$. Indeed, the noise density p_{n} determines the level of demonstrator' expertise as well as the quality of demonstrations. The goal of IL with diverse-quality demonstrations is to learn the optimal policy using dataset $\{(\boldsymbol{\tau}_{\mathrm{su}}, k)_n\}_{n=1}^N$.

We emphasize that identification numbers do not contain information about the quality and do not need to be provided by experts. When the number is not given, a simple strategy is to set $k = n$ and $K = N$, which corresponds to assuming that each demonstrator collects one demonstration.

2.4. The Deficiency of Existing Methods

Indeed, methods for high-quality demonstrations in Section 2.2 are unsuitable for diverse-quality demonstrations in Section 2.3 due to the differences between $p^\star$ and p_{d}. Specifically, by comparing $p^\star$ and p_{d}, we can see that these methods would learn a policy $\widehat{\pi}$ that averages over noise-densities, i.e., $\widehat{\pi}(\mathbf{u}_t|\mathbf{s}_t) \approx \Sigma_{k=1}^K \nu(k) \int_{\mathcal{A}} \pi^\star(\mathbf{a}_t|\mathbf{s}_t) p_{\mathrm{n}}(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k) \mathrm{d}\mathbf{a}_t$. This averaging policy clearly differs from the optimal policy.

Multi-modal IL methods (Li et al., 2017; Hausman et al., 2017; Wang et al., 2017) are also unsuitable for diverse-quality demonstrations. Specifically, these methods aim to learn a multi-modal policy where different modalities estimate different policies. These methods are suitable for diverse demonstrations which are collected by experts with different optimal policies, because different modalities simply estimate optimal policy of different experts. However, these methods become unsuitable with diverse-quality demonstrations, because some modalities estimate policy of amateurs. For this reason, it is crucial to choose good modalities that estimate experts' policies, but doing so typically requires knowing the quality of demonstrations.

In supervised-learning, a well-known approach for handling diverse-quality data is to estimate the quality of data (Angluin & Laird, 1988; Raykar et al., 2010). Based on this approach, the quality of demonstrations may be estimated by using a parameterized model $p_{\theta,\omega}$ to estimate p_{d} as follows:

$$p_{\theta,\omega}(\boldsymbol{\tau}_{\mathrm{su}}, k) = \nu(k)\mu(\mathbf{s}_1) \prod_{t=1}^{T} p_{\mathrm{s}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{u}_t) \times \int_{\mathcal{A}} \pi_\theta(\mathbf{a}_t|\mathbf{s}_t) p_\omega(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k) \mathrm{d}\mathbf{a}_t. \quad (3)$$

The parameters $\boldsymbol{\theta}$ and $\boldsymbol{\omega}$ can be learned by a regression method (see Appendix B.2), where π_θ estimates the optimal policy and p_ω estimates the noise-density. However, this approach suffers from the issue of compounding error (Ross & Bagnell, 2010) and tends to perform poorly at test time. Namely, regression methods assume that data distributions are identical during training and testing. However, data distributions in IL depend on policies (Puterman, 1994), which leads to discrepancies between data distributions during training and testing. Due to this, compounding error can occur during testing, where prediction errors increase in future time steps due to changing data distributions.

Our goal is to tackle IL with diverse-quality demonstrations under this realistic setting (i.e., experts are unavailable), while avoiding these deficiencies.

3. VILD: A Robust Method for Diverse-quality Demonstrations

This section presents VILD, namely a robust method for tackling the challenge from diverse-quality demonstrations. Specifically, we build a parameterized model that explicitly describes the noise-density and a reward function (Section 3.1), and estimate its parameters by a variational approach (Section 3.2), which can be implemented easily by RL (Section 3.3). We also improve data-efficiency by using importance sampling (Section 3.4). Mathematical derivations are provided in Appendix A.

3.1. Modeling Diverse-quality Demonstrations

Our key idea to overcome the challenge of diverse-quality demonstrations is to estimate the quality of demonstrations. To avoid the deficiency of the model $p_{\theta,\omega}$ in Eq. (3), we utilize inverse RL (IRL) (Ng & Russell, 2000), where we learn a reward function from diverse-quality demonstrations. IL problems can be solved by a combination of IRL and RL, where we learn a reward function by IRL and then learn a policy from the reward function by RL[1]. This combination avoids the issue of compounding error, since the policy is learned by RL which takes into account the dependency between data distribution and policy (Ho & Ermon, 2016).

Specifically, our parameterized model for estimating p_{d} is based on a model of maximum entropy IRL (ME-IRL) (Ziebart et al., 2010), which learns a reward function from expert demonstrations by using a model $p_\phi(\boldsymbol{\tau}_{\mathrm{sa}}) \propto \mu(\mathbf{s}_1)\Pi_{t=1}^T p_{\mathrm{s}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) e^{r_\phi(\mathbf{s}_t, \mathbf{a}_t)}$, where $\boldsymbol{\phi}$ is the parameter. Based on this model, we propose to learn the reward function *and* noise-density by

$$p_{\phi,\omega}(\boldsymbol{\tau}_{\mathrm{su}}, k) = \frac{1}{Z_{\phi,\omega}} \nu(k)\mu(\mathbf{s}_1) \prod_{t=1}^{T} p_{\mathrm{s}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{u}_t) \times \int_{\mathcal{A}} e^{r_\phi(\mathbf{s}_t, \mathbf{a}_t)} p_\omega(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k) \mathrm{d}\mathbf{a}_t, \quad (4)$$

where $\boldsymbol{\phi}$ and $\boldsymbol{\omega}$ are parameters, and $Z_{\phi,\omega}$ is the normalization term ensuring that $p_{\phi,\omega}$ integrates to one. By comparing $p_{\phi,\omega}$ to p_{d}, the reward parameter $\boldsymbol{\phi}$ should be learned so that the cumulative reward is proportional to a joint probability density of actions given by the optimal policy, i.e., $e^{\Sigma_{t=1}^T r_\phi(\mathbf{s}_t, \mathbf{a}_t)} \propto \Pi_{t=1}^T \pi^\star(\mathbf{a}_t|\mathbf{s}_t)$. In other words, the cumu-

[1] IRL differs from RL; IRL learns a reward function from demonstrations, but RL learns a policy from a reward function.

lative reward is large for trajectories induced by the optimal policy. Therefore, the optimal policy can be learned by maximizing reward r_ϕ under transition probability p_{s}. Meanwhile, the model p_ω estimates the noise-density p_{n}, and the estimated level of demonstrators' expertise can be determined from p_ω.

To learn parameters of this model, we propose to minimize the KL divergence from the data distribution to the model: $\min_{\boldsymbol{\phi},\boldsymbol{\omega}} \mathrm{KL}(p_{\mathrm{d}}||p_{\phi,\omega})$. By ignoring constants and letting $l_{\phi,\omega}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{u}_t, k) = r_\phi(\mathbf{s}_t, \mathbf{a}_t) + \log p_\omega(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k)$, minimizating the KL divergence is equivalent to solving

$$\max_{\boldsymbol{\phi},\boldsymbol{\omega}} f(\boldsymbol{\phi},\boldsymbol{\omega}) - g(\boldsymbol{\phi},\boldsymbol{\omega}), \tag{5}$$

where

$$f(\boldsymbol{\phi},\boldsymbol{\omega}) = \mathbb{E}_{p_{\mathrm{d}}}\left[\sum_{t=1}^{T} \log\left(\int_{\mathcal{A}} e^{l_{\phi,\omega}(\mathbf{s}_t,\mathbf{a}_t,\mathbf{u}_t,k)} \mathrm{d}\mathbf{a}_t\right)\right], \tag{6}$$

$$g(\boldsymbol{\phi},\boldsymbol{\omega}) = \log Z_{\phi,\omega}. \tag{7}$$

Solving this maximization requires computing the integrals over both state space $\mathcal{S}$ (contained in g) and action space $\mathcal{A}$. Computing these integrals is feasible for small state-action spaces, but is infeasible for large state-action spaces. To scale up our model to large state-action spaces, we leverage a variational approach in the followings.

3.2. A Variational Approach for Parameter Estimation

The central idea of the variational approach is to lower-bound an integral by the Jensen inequality and a variational distribution (Jordan et al., 1999). The main benefit of the approach is that the integral can be computed via an expectation over an optimal variational distribution; This makes it easier to solve an optimization problem. However, finding the optimal variational distribution usually requires solving a sub-optimization problem.

Before we proceed, notice that the difference $f(\boldsymbol{\phi},\boldsymbol{\omega}) - g(\boldsymbol{\phi},\boldsymbol{\omega})$ is not a joint concave function of the integrals, and this prohibits using the Jensen inequality on this difference. However, we can separately lower-bound f and g by the Jensen inequality, since they are concave functions of their corresponding integrals. Specifically, a variational distribution $q_\psi(\mathbf{a}_t|\mathbf{s}_t, \mathbf{u}_t, k)$ with parameter $\boldsymbol{\psi}$ yields an inequality

$$\begin{aligned} f(\boldsymbol{\phi},\boldsymbol{\omega}) &\geqslant \mathbb{E}_{p_{\mathrm{d}}}\left[\sum_{t=1}^{T} \mathbb{E}_{q_\psi}\left[l_{\phi,\omega}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{u}_t, k)\right] + H_t(q_\psi)\right] \\ &= \mathcal{F}(\boldsymbol{\phi},\boldsymbol{\omega},\boldsymbol{\psi}), \end{aligned} \tag{8}$$

where we define $H_t(q_\psi) = -\mathbb{E}_{q_\psi}\left[\log q_\psi(\mathbf{a}_t|\mathbf{s}_t, \mathbf{u}_t, k)\right]$. It can be verified that $f(\boldsymbol{\phi},\boldsymbol{\omega}) = \max_{\boldsymbol{\psi}} \mathcal{F}(\boldsymbol{\phi},\boldsymbol{\omega},\boldsymbol{\psi})$. Meanwhile, a variational distribution $q_\theta(\mathbf{a}_t, \mathbf{u}_t|\mathbf{s}_t, k)$ with parameter $\boldsymbol{\theta}$ yields an inequality

$$\begin{aligned} g(\boldsymbol{\phi},\boldsymbol{\omega}) &\geqslant \mathbb{E}_{\bar{q}_\theta}\left[\sum_{t=1}^{T} l_{\phi,\omega}(\mathbf{s}_t, \mathbf{a}_t, \mathbf{u}_t, k)\right] + \mathcal{H}(\bar{q}_\theta) \\ &= \mathcal{G}(\boldsymbol{\phi},\boldsymbol{\omega},\boldsymbol{\theta}), \end{aligned} \tag{9}$$

where $\mathcal{H}(\bar{q}_\theta) = -\mathbb{E}_{\bar{q}_\theta}\left[\Sigma_{t=1}^{T} \log q_\theta(\mathbf{a}_t, \mathbf{u}_t|\mathbf{s}_t, k)\right]$, $\bar{q}_\theta(\boldsymbol{\tau}_{\mathrm{sau}}, k) = \nu(k)\mu(\mathbf{s}_1)\Pi_{t=1}^{T} p_{\mathrm{s}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{u}_t) q_\theta(\mathbf{a}_t, \mathbf{u}_t|\mathbf{s}_t, k)$, and $\boldsymbol{\tau}_{\mathrm{sau}} = (\mathbf{s}_{1:T+1}, \mathbf{a}_{1:T}, \mathbf{u}_{1:T})$. The lower-bound $\mathcal{G}$ resembles an objective function of maximum entropy RL (Ziebart et al., 2010). Based on the optimality results of maximum entropy RL, it can be verified that $g(\boldsymbol{\phi},\boldsymbol{\omega}) = \max_{\boldsymbol{\theta}} \mathcal{G}(\boldsymbol{\phi},\boldsymbol{\omega},\boldsymbol{\theta})$. Variational distributions $q_\psi^\star$ and $q_\theta^\star$ that maximize the lower-bounds ($\mathcal{F}$ and $\mathcal{G}$, respectively) are called optimal variational distributions.

By using the variational approach, Eq. (5) can be written as

$$\max_{\boldsymbol{\phi},\boldsymbol{\omega},\boldsymbol{\psi}} \min_{\boldsymbol{\theta}} \mathcal{F}(\boldsymbol{\phi},\boldsymbol{\omega},\boldsymbol{\psi}) - \mathcal{G}(\boldsymbol{\phi},\boldsymbol{\omega},\boldsymbol{\theta}). \tag{10}$$

It is feasible to solve Eq. (10) for large state-action spaces, since $\mathcal{F}$ and $\mathcal{G}$ are defined as expectations and can be optimized straightforwardly. In practice, we represent the variational distributions by parameterized functions (e.g., neural networks), and solve the optimization by stochastic gradient methods where expectations are approximated using mini-batch samples (Ranganath et al., 2014).

3.3. Choices of Density Models in Practice

In practice, we need to specify density models in our optimization (Eq. (10)). For continuous-control tasks, we use

$$p_\omega(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k) = \mathcal{N}(\mathbf{u}_t|\mathbf{a}_t, \mathbf{C}_\omega(k)), \tag{11}$$

$$q_\theta(\mathbf{a}_t, \mathbf{u}_t|\mathbf{s}_t, k) = q_\theta(\mathbf{a}_t|\mathbf{s}_t)\mathcal{N}(\mathbf{u}_t|\mathbf{a}_t, \boldsymbol{\Sigma}_k), \tag{12}$$

where $\mathcal{N}(\mathbf{a}|\mathbf{b}, \mathbf{C})$ denotes a Gaussian distribution with mean $\mathbf{b}$ and covariance $\mathbf{C}$, and $\boldsymbol{\Sigma}_k$ is a hyper-parameter. We use the Gaussian distribution for p_ω to incorporate a prior assumption that noise-density p_{n} tends to Gaussian. Indeed, covariance $\mathbf{C}_\omega(k)$ gives an estimated expertise of the k-th demonstrator: the covariance is small for high-expertise demonstrators and vice-versa for low-expertise demonstrators[2]. Meanwhile, the choice for $q_\theta(\mathbf{a}_t, \mathbf{u}_t|\mathbf{s}_t, k)$ in Eq. (12) enables using RL to optimize $\boldsymbol{\theta}$, as will be described below. The choices for $q_\psi(\mathbf{a}_t|\mathbf{s}_t, \mathbf{u}_t, k)$ and $q_\theta(\mathbf{a}_t|\mathbf{s}_t)$ are flexible; We use Gaussians which are common for distributions over continuous action (Duan et al., 2016), but other choices such as the beta distributions can be used (Chou et al., 2017).

With the above density models, Eq. (10) is equivalent to

$$\max_{\boldsymbol{\phi},\boldsymbol{\omega},\boldsymbol{\psi}} \min_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\phi},\boldsymbol{\omega},\boldsymbol{\psi},\boldsymbol{\theta}), \tag{13}$$

[2]Different choices of p_ω incorporate different prior assumptions. For example, a Laplace distribution may be used to model demonstrations with outliers (see Appendix A.4) (Murphy, 2013).

where

$$\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\omega}, \boldsymbol{\psi}, \boldsymbol{\theta}) = \mathbb{E}_{p_{\mathrm{d}}}\left[\sum_{t=1}^{T} \mathbb{E}_{q_\psi}\left[r_\phi(\mathbf{s}_t, \mathbf{a}_t) - \frac{1}{2}\|\mathbf{u}_t - \mathbf{a}_t\|^2_{\mathbf{C}_\omega^{-1}(k)}\right] + H_t(q_\psi)\right] - \mathbb{E}_{\widetilde{q}_\theta}\left[\sum_{t=1}^{T} r_\phi(\mathbf{s}_t, \mathbf{a}_t)\right] - \mathcal{H}(\widetilde{q}_\theta) + \frac{T}{2}\mathbb{E}_\nu\left[\mathrm{Tr}(\mathbf{C}_\omega^{-1}(k)\boldsymbol{\Sigma}_k)\right].$$

Here, $\widetilde{q}_\theta(\boldsymbol{\tau}_{\mathrm{sa}}) = \mu(\mathbf{s}_1)\Pi_{t=1}^{T}\widetilde{p}_{\mathrm{s}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t)q_\theta(\mathbf{a}_t|\mathbf{s}_t)$ and $\widetilde{p}_{\mathrm{s}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) = \mathbb{E}_\nu\left[\int_{\mathcal{A}} p_{\mathrm{s}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{u}_t)\mathcal{N}(\mathbf{u}_t|\mathbf{a}_t, \boldsymbol{\Sigma}_k)\mathrm{d}\mathbf{u}_t\right]$. Recall that the optimal policy may be learned by maximizing reward r_ϕ under transition probability p_{s}. As can be seen, minimizing $\mathcal{L}$ w.r.t. $\boldsymbol{\theta}$ is equivalent to solving maximum entropy RL with reward r_ϕ and transition probability $\widetilde{p}_{\mathrm{s}}$. The discrepancy between p_{s} and $\widetilde{p}_{\mathrm{s}}$ is determined by $\boldsymbol{\Sigma}_k$: smaller value of $\boldsymbol{\Sigma}_k$ yields less discrepancy. Therefore, by choosing a reasonably small value of $\boldsymbol{\Sigma}_k$, we can optimize $\boldsymbol{\theta}$ by RL to obtain $q_\theta(\mathbf{a}_t|\mathbf{s}_t)$ which estimates the optimal policy. This is advantageous because we can use state-of-the-art RL methods without significant modifications to implementations.

To sum up, VILD solves Eq. (13) to learn policy $q_\theta(\mathbf{a}_t|\mathbf{s}_t)$, where $\boldsymbol{\theta}$ is optimized by RL with reward r_ϕ, while $\boldsymbol{\phi}$, $\boldsymbol{\omega}$, and $\boldsymbol{\psi}$ are optimized by stochastic gradient methods such as Adam (Kingma & Ba, 2015). Algorithm 1 shows the pseudo-code of VILD. We use a diagonal matrix for $\mathbf{C}_\omega(k)$ and also include a regularization term $L(\boldsymbol{\omega}) = T\mathbb{E}_\nu[\log|\mathbf{C}_\omega^{-1}(k)|]/2$ to penalize overly large values of $\mathbf{C}_\omega(k)$. Note that $\mathcal{L}$ already includes a penalty $\mathbb{E}_\nu[\mathrm{Tr}(\mathbf{C}_\omega^{-1}(k)\boldsymbol{\Sigma}_k)]$, but its strength is too small because $\boldsymbol{\Sigma}_k$ is chosen to be small. Similarly to prior works (Ho & Ermon, 2016), we implement VILD using feed-forward neural networks with two hidden-layers and use Monte-Carlo estimation to approximate expectations. We also pre-train the Gaussian mean of q_ψ to obtain reasonable initial predictions; We perform least-squares regression for 1000 gradient steps with target value $\mathbf{u}_t$. More implementation details are given in Appendix C[3].

3.4. Importance Sampling for Reward Learning

To improve the convergence rate of VILD when optimizing $\boldsymbol{\phi}$, we use importance sampling (IS). Specifically, the gradient $\nabla_\phi\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\omega}, \boldsymbol{\psi}, \boldsymbol{\theta})$ indicates that $\boldsymbol{\phi}$ needs to maximize the expected cumulative reward achieved by p_{d} and q_ψ, and at the same time minimize the expected cumulative reward achieved by q_θ. However, low-quality demonstrations drawn from p_{d} often yield low reward values which are not informative for maximization. For this reason, stochastic gradients estimated by these demonstrations tend to be uninformative, which leads to slow convergence and poor data-efficiency.

To avoid estimating such uninformative gradients, we use IS to estimate gradients using high-quality demonstrations which are sampled with high probability. Briefly, IS is a technique for estimating an expectation over a distribution by using samples from a different distribution (Robert & Casella, 2005). For VILD, we sample k from a distribution $\widetilde{\nu}(k) \propto \|\mathrm{vec}(\mathbf{C}_\omega^{-1}(k))\|_1$ which assigns high probabilities to k with high expertise (i.e., small $\mathbf{C}_\omega(k)$). With this distribution, the estimated gradients tend to be more informative for reward learning. To reduce a sampling bias, we use a truncated importance weight: $w(k) = \min(\nu(k)/\widetilde{\nu}(k), 1)$, which leads to an IS gradient:

$$\nabla_\phi^{\mathrm{IS}}\mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\omega}, \boldsymbol{\psi}, \boldsymbol{\theta}) = \mathbb{E}_{\widetilde{p}_{\mathrm{d}}}\left[w(k)\sum_{t=1}^{T}\mathbb{E}_{q_\psi}\left[\nabla_\phi r_\phi(\mathbf{s}_t, \mathbf{a}_t)\right]\right] - \mathbb{E}_{\widetilde{q}_\theta}\left[\sum_{t=1}^{T}\nabla_\phi r_\phi(\mathbf{s}_t, \mathbf{a}_t)\right], \quad (14)$$

where $\widetilde{p}_{\mathrm{d}}(\boldsymbol{\tau}_{\mathrm{su}}, k)$ is defined similarly to $p_{\mathrm{d}}(\boldsymbol{\tau}_{\mathrm{su}}, k)$ in Eq. (2) but with $\widetilde{\nu}(k)$ instead of $\nu(k)$. To obtain mini-batch samples from $\widetilde{p}_{\mathrm{d}}$, we sample k from $\widetilde{\nu}(k)$ and then uniformly sample demonstrations associated with k from dataset $\{(\boldsymbol{\tau}_{\mathrm{su}}, k)_n\}_{n=1}^N$. Computing $w(k)$ requires $\nu(k)$, which can be estimated accurately since k is a discrete random variable. For simplicity, we assume a uniform $\nu(k)$.

We note that the gradient in Eq. (14) is biased when $\nu(k)/\widetilde{\nu}(k) > 1$. Nonetheless, the biases may improve robustness against model misspecification, i.e., when the Gaussian model p_ω in Eq. (11) cannot exactly represent noise-density p_{n}. Specifically, the optimal solution of Eq. (13) may yield a poor policy when the model is misspecified. In such cases, informative biases can be introduced such that the solution has desirable properties[4]. For VILD, a desirable property is that the reward function yields relatively large values for high-expertise demonstrators. This is precisely the consequence of using $\widetilde{\nu}(k)$ for reward learning. Note that the usefulness of these biases still depend on the relative accuracy of estimated covariance.

3.5. Discussion

In this section, we discuss computational costs of VILD and a connection between VILD and maximum entropy IRL.

Computational costs. VILD does not incur large additional computational costs compared to prior methods. Specifically, additional costs of VILD include the cost of computing gradients w.r.t. $\boldsymbol{\omega}$ and $\boldsymbol{\psi}$ and the cost of sampling from q_ψ. For $\boldsymbol{\omega}$, the cost of computing gradients is very low because $\mathbf{C}_\omega(k)$ is a diagonal matrix. For $\boldsymbol{\psi}$, the cost of

[3] Source code: www.github.com/voot-t/vild_code

[4] For instance, an ℓ_2-regularization introduces biases to obtain a solution with a small ℓ_2-norm (Hastie et al., 2001).

Algorithm 1 VILD: Variational Imitation Learning with Diverse-quality demonstrations

Input: Diverse-quality demonstrations $\{(\boldsymbol{\tau}_{\mathrm{su}}, k)_n\}_{n=1}^N \sim p_{\mathrm{d}}(\boldsymbol{\tau}_{\mathrm{su}}, k)$ and a replay buffer $\mathcal{B} = \varnothing$.
Pre-train $q_\psi(\mathbf{a}_t|\mathbf{s}_t, \mathbf{u}_t, k)$ by least-squares regression. (see Appendix C)
while Not converge **do**
  while $|\mathcal{B}| < B$ with batch size B **do**
    Sample $\mathbf{a}_t \sim q_\theta(\mathbf{a}_t|\mathbf{s}_t)$ and $\boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{\epsilon}_t|\mathbf{0}, \boldsymbol{\Sigma}_k)$, observe $\mathbf{s}'_t \sim p(\mathbf{s}'_t|\mathbf{s}_t, \mathbf{a}_t + \boldsymbol{\epsilon}_t)$, and include $(\mathbf{s}_t, \mathbf{a}_t, \mathbf{s}'_t)$ into $\mathcal{B}$
  end while
  Update q_ψ by an estimate of $\nabla_{\boldsymbol{\psi}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\omega}, \boldsymbol{\psi}, \boldsymbol{\theta})$.
  Update p_ω by an estimate of $\nabla_{\boldsymbol{\omega}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\omega}, \boldsymbol{\psi}, \boldsymbol{\theta}) + \nabla_{\boldsymbol{\omega}} L(\boldsymbol{\omega})$.
  Update r_ϕ by an estimate of $\nabla^{\mathrm{IS}}_{\boldsymbol{\phi}} \mathcal{L}(\boldsymbol{\phi}, \boldsymbol{\omega}, \boldsymbol{\psi}, \boldsymbol{\theta})$.
  Update q_θ by an RL method (e.g., TRPO, SAC, or PPO) with reward function r_ϕ.
end while

computing gradients depends on the size of neural networks, and the cost of sampling depends on the number of samples drawn for Monte-Carlo estimation. In our experiments, we draw one sample from q_ψ to reduce the cost and use antithetic sampling to reduce estimation variances (Robert & Casella, 2005). Overall, additional costs of VILD are relatively low compared to the cost of collecting transition samples from MDP which is the main computational burden of many IL methods.

Relation to maximum entropy IRL. The model of VILD is based on the model of maximum entropy IRL (ME-IRL) (Ziebart et al., 2010) and VILD is closely related to ME-IRL. Specifically, VILD reduces to ME-IRL under an assumption that demonstrations are high-quality. This assumption is equivalent to letting q_ψ and p_ω be Dirac deltas: $q_\psi(\mathbf{a}_t|\mathbf{s}_t, \mathbf{u}_t, k) = \delta_{\mathbf{a}_t = \mathbf{u}_t}$ and $p_\omega(\mathbf{u}_t|\mathbf{a}_t, \mathbf{s}_t, k) = \delta_{\mathbf{u}_t = \mathbf{a}_t}$. In this case, the optimization in Eq. (10) is equivalent to

$$\max_{\boldsymbol{\phi}} \min_{\boldsymbol{\theta}} \mathbb{E}_{p_{\mathrm{d}}} \left[\sum_{t=1}^T r_\phi(\mathbf{s}_t, \mathbf{a}_t)\right] - \mathbb{E}_{q_\theta} \left[\sum_{t=1}^T r_\phi(\mathbf{s}_t, \mathbf{a}_t)\right] - \mathcal{H}(q_\theta), \qquad (15)$$

where $q_\theta(\boldsymbol{\tau}_{\mathrm{sa}}) = \mu(\mathbf{s}_1) \Pi_{t=1}^T p_{\mathrm{s}}(\mathbf{s}_{t+1}|\mathbf{s}_t, \mathbf{a}_t) q_\theta(\mathbf{a}_t|\mathbf{s}_t)$. Note that $\boldsymbol{\psi}$ and $\boldsymbol{\omega}$ do not appear in this objective because q_ψ and p_ω are Dirac deltas without parameters. This objective is equivalent to that of ME-IRL. In practice, when all demonstrations have high quality, we expect VILD to estimate small covariance for all demonstrations and this implies that $q_\psi(\mathbf{a}_t|\mathbf{s}_t, \mathbf{u}_t, k) \to \delta_{\mathbf{a}_t = \mathbf{u}_t}$ and $p_\omega(\mathbf{u}_t|\mathbf{a}_t, \mathbf{s}_t, k) \to \delta_{\mathbf{u}_t = \mathbf{a}_t}$. Based on this, we conjecture that VILD performs comparable to ME-IRL given high-quality demonstrations. Our experiment in Appendix D.4 supports this conjecture.

4. Experiments

We experimentally evaluate VILD (with IS and without IS) in continuous-control tasks. Performance is evaluated using a cumulative ground-truth reward along trajectories collected by policies (Ho & Ermon, 2016). We report the mean and standard error computed over 5 trials.

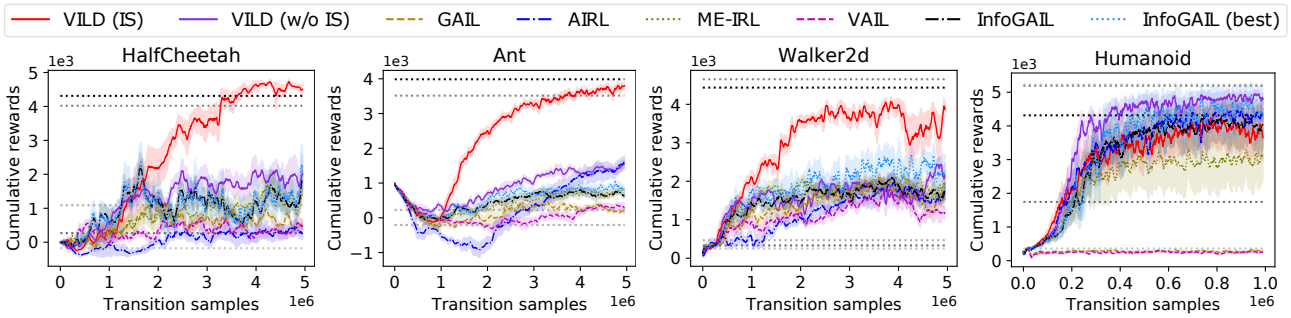
4.1. Comparison in Continuous-control Benchmarks

In this section, we evaluate VILD in continuous-control benchmark tasks (Brockman et al., 2016) (HalfCheetah, Ant, Walker2d, and Humanoid) under scenarios where the Gaussian model of p_ω is correct. Specifically, for each task, we generate two datasets using two types of Gaussian noise-density: $p_{\mathrm{n}}(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k) = \mathcal{N}(\mathbf{u}_t|\mathbf{a}_t, \boldsymbol{\sigma}_k^2)$ (time-action independent), and $p_{\mathrm{n}}(\mathbf{u}_t|\mathbf{s}_t, \mathbf{a}_t, k) = \mathcal{N}(\mathbf{u}_t|\mathbf{a}_t, \boldsymbol{\sigma}_k^2(\mathbf{a}, t))$ (time-action dependent). For each dataset, we use a pre-trained $\pi^\star$ and $K = 10$ demonstrators to generate approximately 10000 state-action pairs.

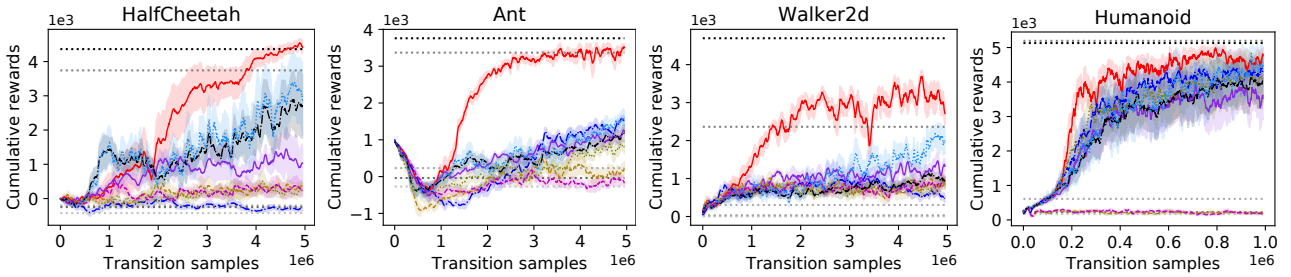
4.1.1. Comparison Against RL-based Methods

Firstly, we compare VILD against RL-based methods that use RL to optimize a policy. These methods include GAIL (Ho & Ermon, 2016), AIRL (Fu et al., 2018), VAIL (Peng et al., 2019), ME-IRL (Ziebart et al., 2010), and InfoGAIL (Li et al., 2017). These existing methods are well-known in IL, but they do not take diverse-quality into account. We use TRPO (Schulman et al., 2015) as an RL method, except on the Humanoid task where we use SAC (Haarnoja et al., 2018) since TRPO does not perform well. For InfoGAIL, a multi-modal IL method that learns a context-dependent policy, we report performance averaged over all contexts and performance with the best context (denoted by InfoGAIL (best)). Note that in InfoGAIL, modalities of a multi-modal policy are chosen based on values of context. The number of contexts is set to K.

Figure 2 shows the performance against the number of transition samples collected by RL. The results show that VILD with IS achieves state-of-the-art performance and outperforms the rest overall. VILD without IS also tends to outperform existing methods in terms of the final performance. However, it is outperformed by VILD with IS, except on the Humanoid task with time-action independent density (Figure 2(a)). This is perhaps because bias from IS may have a negative effect when the model choice is correct. Nonetheless, the overall good performance of VILD with IS



(a) Demonstrations are generated by time-action independent noise-density.



(b) Demonstrations are generated by time-action dependent noise-density.

Figure 2. Comparison on continuous-control benchmarks against RL-based methods that do not take diverse-quality into account. VILD-based methods perform overall better than the rest. Demonstrations are artificially generated. (VILD (IS) and VILD (w/o IS) denote VILD with and without IS, respectively. Horizontal dots denote performance of 5 demonstrators. Shaded area denotes standard errors.)

demonstrates that it is more robust against diverse-quality demonstrations compared to existing methods.

On the contrary, existing methods perform poorly, except on the Humanoid task, where all methods except GAIL and VAIL achieve statistically comparable performance according to t-test. This result implies that diverse-quality demonstrations in this task may not have strong negative-effects on the performance. This is perhaps because amateurs in this task perform relatively well compared to amateurs in other tasks (see Appendix D.1).

We also evaluate the accuracy of quality-estimation in VILD, where we compare estimated covariance $\mathbf{C}_\omega(k)$ against ground-truth covariance $\boldsymbol{\sigma}_k^2$ of noise-density. Figures of this comparison are given in Appendix D.1 due to space limitation. The results show that the estimation is quite accurate. Nonetheless, the estimation tends to be less accurate for low-expertise demonstrators. A reason for this phenomenon is that low-quality demonstrations are highly dissimilar, which makes quality-estimating more challenging.

4.1.2. Comparison Against SL-based Methods

Next, we compare VILD against supervised-learning (SL)-based methods, namely behavior cloning (BC) (Pomerleau, 1988), Co-teaching, and BC with diverse-quality demonstrations (BC-D). Specifically, BC performs regression without taking diverse-quality data into account. Co-teaching is a regression-extension of a recent classification method (Han et al., 2018) that is robust against diverse-quality data. BC-D takes diverse-quality data into account by performing regression based on the simple model $p_{\theta,\omega}$. We compare the performance of these methods against the performance of VILD with IS in the last 100 iteration of Figure 2.

Figure 3 shows the performance of these SL-based methods against the number of gradient steps. Performances for time-action dependent noise-density are similar and given in Appendix D.1. As seen, SL-based methods perform very poorly and their final performance is much worse compared to VILD with IS. In particular, for the Humanoid task which has the largest state-action space, SL-based methods could not improve upon the initial policy at all.

Notice that the performance of SL-based methods sharply degrades as training progresses. We conjecture that this degradation is due to compounding error caused by overfitting. Specifically, these methods may learn reasonably good policies early on (e.g., in Ant and Walker2d), but the policies overfit to training data as training progresses. During testing, these overfitted policies may make incorrect predictions which cause compounding error. In addition, diverse-quality demonstrations also makes the issue more severe, since neural networks tend to overfit to low-quality data (Arpit et al., 2017). Due to these reasons, BC performs poorly as it suffers from issues of compounding error and diverse-quality demonstrations. Meanwhile, Co-teaching

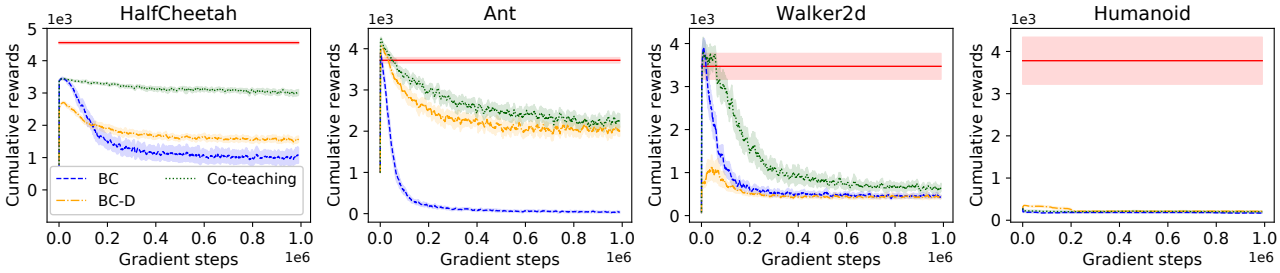


Figure 3. Comparison on continuous-control benchmarks against supervised-learning-based methods. BC-D and Co-teaching take diverse-quality into account, while BC does not. VILD with IS (red horizontal lines) clearly performs better than these methods. Demonstrations are artificially generated by time-action independent noise-density.

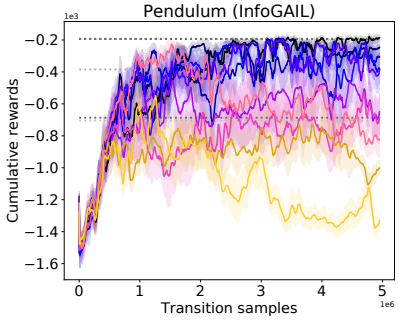


Figure 4. Performance of InfoGAIL on Pendulum with different values of context. Clearly, choosing a good value of context is crucial for InfoGAIL.

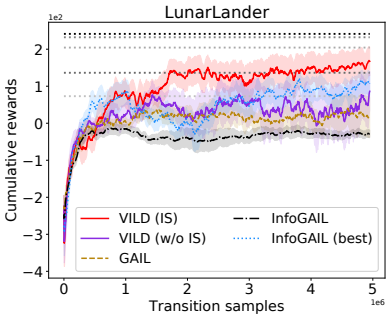


Figure 5. Comparison on LunarLander against GAIL and InfoGAIL. The model of VILD is incorrect, but VILD with IS still outperforms comparison methods.

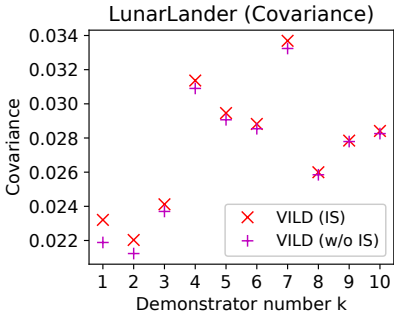


Figure 6. Results of quality-estimation by VILD in LunarLander. The estimated covariance $\mathbf{C}_\omega(k)$ yields relatively accurate quality for each demonstrator.

and BC-D are quite robust against diverse-quality demonstrations, but they still suffer from compounding error and perform worse than VILD with IS. Overall, these results indicate that SL methods for diverse-quality data are not suitable for diverse-quality demonstrations.

4.1.3. Investigating InfoGAIL in Pendulum Task

From Figure 2, we can see that InfoGAIL performs poorly when its performance is averaged over all contexts. Using the best context improves its performance, but the improvement is quite mild. We investigated this phenomenon and found that the learned multi-modal policy yields similar performance for all contexts (see Appendix D.1), which implies that InfoGAIL fails to learn a good multi-modal policy. This is perhaps because learning a multi-modal policy is challenging in large state-action spaces. To verify our claim that choosing good modalities is crucial for multi-modal IL (Section 2.4), we perform an experiment in a Pendulum task. This task has a much smaller state-action space and we expect InfoGAIL to learn a good multi-modal policy.

Figure 4 shows the performance of InfoGAIL for different values of context (denoted by different colors). As seen, the performance of InfoGAIL crucially depends on the value of context. Namely, a well-chosen context yields a policy with good performance, whereas a poorly-chosen context yields a policy with poor performance. Indeed, averaging these policies over all contexts yields a policy with average performance. This result supports our claim that choosing good modalities is crucial for multi-modal IL methods. However, recall that doing so is typically difficult when the quality of demonstrations is unknown. In our experiments, good modalities could be chosen based on performance, but this is not possible when a ground-truth reward function for performance evaluation is not available.

4.2. Robustness Against Incorrect Model Choices

Next, we evaluate the robustness of VILD against incorrect model choices. Specifically, we evaluate VILD when p_n is not Gaussian. We consider a LunarLander task, where an optimal policy is available for generating high-quality demonstrations (Brockman et al., 2016). To generate diverse-quality demonstrations, we perturb parameters of the optimal policy using half-Gaussian distributions with variance depending on k. We use $K = 10$ to generate a dataset with approximately 20000 state-action pairs. We compare VILD against GAIL and InfoGAIL; We expect other RL-based methods to perform similarly to GAIL, based on benchmark results. We use PPO (Schulman et al., 2017) as an RL method. We use a log-sigmoid reward function for VILD to make comparison against GAIL fair (see Appendix D.2).

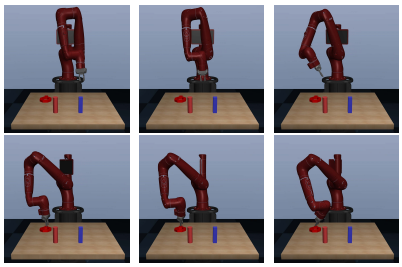


Figure 7. RobosuiteReacher task. Rewards are inverse proportional to distance between the end-effector and red object. Depicted trajectory is obtained by VILD with IS (left to right, top to bottom).

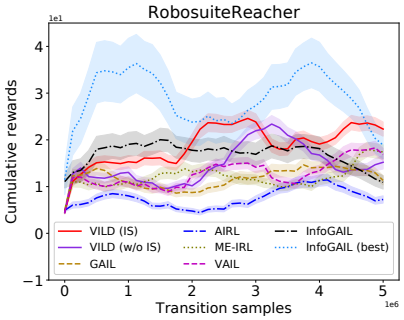


Figure 8. Comparison on RobosuiteReacher against RL-based methods using real-world demonstrations. VILD with IS performs overall better than methods that do not take diversity into account.

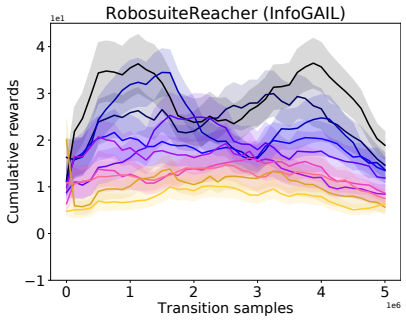


Figure 9. Performance of InfoGAIL on RobosuiteReacher with different values of context. Performance of InfoGAIL is highly unstable.

Figure 5 shows the performance. It can be seen that VILD with IS outperforms comparison methods and learns a good policy. This result indicates that VILD with IS is robust against incorrect model choices. On the other hand, VILD without IS does not perform as well as VILD with IS. The discrepancy between them is perhaps due to IS biases which enable VILD with IS to learn a better solution. Meanwhile, GAIL and InfoGAIL do not perform well. Using the best context can improve the performance of InfoGAIL, but its performance is still poor compared to VILD with IS.

Figure 6 shows results of quality-estimation by VILD. The results show that the quality-estimation is reasonably accurate under this scenario. Namely, the value of $\mathbf{C}_\omega(k)$ of high-expertise demonstrators (i.e., $k = 1, 2, 3$) is relatively smaller than that of low-expertise demonstrators (i.e., $k = 8, 9, 10$). Note that we cannot directly evaluate quality-estimation against the ground-truth because the noise-density is not a Gaussian distribution.

4.3. Robustness Against Real-world Demonstrations

Lastly, we evaluate the robustness of VILD against real-world demonstrations collected by crowdsourcing (Mandlekar et al., 2018). While the public datasets were collected for Assembly tasks in a Robosuite platform (Fan et al., 2018), we consider a Reacher task, where demonstrations in Assembly tasks are clipped when the robot's end-effector contacts the object. We use a Reacher dataset with approximately 5000 state-action pairs. We evaluate RL-based methods where we use TRPO as an RL method. For VILD, we use a log-sigmoid reward function which improves the performance.

Figure 7 shows the task and a trajectory obtained by VILD with IS, while Figure 8 shows the performance obtained by collecting 5 million transition samples for RL training. VILD with IS clearly outperforms comparison methods except InfoGAIL (best). Meanwhile, VILD without IS tends to outperform existing methods except VAIL, InfoGAIL, and InfoGAIL (best). Overall, the results demonstrate that, given 5 million transition samples, VILD with IS is more robust against real-world demonstrations compared to methods that do not take diversity into account and InfoGAIL that do not use the best context.

Note that the final performance of InfoGAIL (best) is comparable to that of VILD with IS, but InfoGAIL (best) learns faster. Nonetheless, InfoGAIL (best) is unstable as its performance fluctuates between good and poor. This instability can be observed for most values of context as shown in Figure 9. This is perhaps due to a large state-action space which makes learning a multi-modal policy challenging.

5. Conclusion

This paper explored a realistic setting in IL where demonstrations have diverse-quality. We showed the deficiency of existing methods, and proposed a robust method called VILD, which learns both the reward function and noise-density by using the variational approach. Empirical evaluations on continuous-control tasks demonstrated that our work enables scalable and data-efficient IL in this setting.

In this work, we considered the noise-density assumption where the quality is determined by noise. In future, we will consider different assumptions for determining the quality.

Acknowledgements

We thank the anonymous reviewers for their constructive feedback. BH was supported by the Early Career Scheme (ECS) through the Research Grants Council of Hong Kong under Grant No.22200720, HKBU Tier-1 Start-up Grant and HKBU CSD Start-up Grant. MS was supported by KAKENHI 17H00757.

References

Angluin, D. and Laird, P. Learning from noisy examples. *Machine Learning*, 1988.

Arpit, D., Jastrzebski, S. K., Ballas, N., Krueger, D., Bengio, E., Kanwal, M. S., Maharaj, T., Fischer, A., Courville, A. C., Bengio, Y., and Lacoste-Julien, S. A closer look at memorization in deep networks. In *International Conference on Machine Learning*, 2017.

Audiffren, J., Valko, M., Lazaric, A., and Ghavamzadeh, M. Maximum entropy semi-supervised inverse reinforcement learning. In *International Joint Conferences on Artificial Intelligence*, 2015.

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. OpenAI Gym. *CoRR*, abs/1606.01540, 2016.

Brown, D. S., Goo, W., Nagarajan, P., and Niekum, S. Extrapolating beyond suboptimal demonstrations via inverse reinforcement learning from observations. In *International Conference on Machine Learning*, 2019.

Chou, P.-W., Maturana, D., and Scherer, S. Improving stochastic policy gradients in continuous control with deep reinforcement learning using the beta distribution. In *International Conference on Machine Learning*, 2017.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pp. 1329–1338, 2016.

Fan, L., Zhu, Y., Zhu, J., Liu, Z., Zeng, O., Gupta, A., Creus-Costa, J., Savarese, S., and Fei-Fei, L. Surreal: Open-source reinforcement learning framework and robot manipulation benchmark. In *Conference on Robot Learning*, 2018.

Fu, J., Luo, K., and Levine, S. Learning robust rewards with adversarial inverse reinforcement learning. In *International Conference on Learning Representation*, 2018.

Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, 2018.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I. W., and Sugiyama, M. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, 2018.

Han, J., Kamber, M., and Pei, J. *Data Mining: Concepts and Techniques, 3rd edition*. Morgan Kaufmann, 2011.

Hastie, T., Tibshirani, R., and Friedman, J. *The Elements of Statistical Learning*. Springer New York Inc., 2001.

Hausman, K., Chebotar, Y., Schaal, S., Sukhatme, G. S., and Lim, J. J. Multi-modal imitation learning from unstructured demonstrations using generative adversarial nets. In *Advances in Neural Information Processing Systems*, 2017.

Ho, J. and Ermon, S. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, 2016.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. An introduction to variational methods for graphical models. *Machine Learning*, 1999.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Li, Y., Song, J., and Ermon, S. Infogail: Interpretable imitation learning from visual demonstrations. In *Advances in Neural Information Processing Systems*, 2017.

Mandlekar, A., Zhu, Y., Garg, A., Booher, J., Spero, M., Tung, A., Gao, J., Emmons, J., Gupta, A., Orbay, E., Savarese, S., and Fei-Fei, L. ROBOTURK: A crowdsourcing platform for robotic skill learning through imitation. In *Conference on Robot Learning*, 2018.

Murphy, K. P. *Machine learning : a probabilistic perspective*. 2013.

Ng, A. Y. and Russell, S. J. Algorithms for inverse reinforcement learning. In *International Conference on Machine Learning*, 2000.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., and Peters, J. An algorithmic perspective on imitation learning. *Foundations and Trends in Robotics*, 2018.

Peng, X. B., Kanazawa, A., Toyer, S., Abbeel, P., and Levine, S. Variational discriminator bottleneck: Improving imitation learning, inverse RL, and GANs by constraining information flow. In *International Conference on Learning Representations*, 2019.

Pomerleau, D. ALVINN: an autonomous land vehicle in a neural network. In *Advances in Neural Information Processing Systems*, 1988.

Puterman, M. L. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. 1994.

Ranganath, R., Gerrish, S., and Blei, D. M. Black box variational inference. In *International Conference on Artificial Intelligence and Statistics*, 2014.

Raykar, V. C., Yu, S., Zhao, L. H., Valadez, G. H., Florin, C., Bogoni, L., and Moy, L. Learning from crowds. *Journal of Machine Learning Research*, 2010.

Robert, C. P. and Casella, G. *Monte Carlo Statistical Methods*. 2005.

Ross, S. and Bagnell, D. Efficient reductions for imitation learning. In *International Conference on Artificial Intelligence and Statistics*, 2010.

Schaal, S. Is imitation learning the route to humanoid robots? *Trends in Cognitive Sciences*, 1999.

Schroecker, Y., Vecerik, M., and Scholz, J. Generative predecessor models for sample-efficient imitation learning. In *International Conference on Learning Representations*, 2019.

Schulman, J., Levine, S., Moritz, P., Jordan, M., and Abbeel, P. Trust region policy optimization. In *International Conference on Machine Learning*, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347, 2017.

Shiarlis, K., Messias, J. V., and Whiteson, S. Inverse reinforcement learning from failure. In *International Conference on Autonomous Agents & Multiagent Systems*, 2016.

Silver, D., Bagnell, J. A., and Stentz, A. Learning autonomous driving styles and maneuvers from expert demonstration. In *International Symposium on Experimental Robotics*, 2012.

Sutton, R. S. and Barto, A. G. *Reinforcement Learning - an Introduction*. MIT Press, 1998.

Wang, Z., Merel, J., Reed, S. E., de Freitas, N., Wayne, G., and Heess, N. Robust imitation of diverse behaviors. In *Advances in Neural Information Processing Systems*, 2017.

Wu, Y., Charoenphakdee, N., Bao, H., Tangkaratt, V., and Sugiyama, M. Imitation learning from imperfect demonstration. In *International Conference on Machine Learning*, 2019.

Ziebart, B. D., Bagnell, J. A., and Dey, A. K. Modeling interaction via the principle of maximum causal entropy. In *International Conference on Machine Learning*, 2010.