Background

Reproducing kernel Hilbert space A Hilbert space is a complete vector space with the metric endowed by the inner product in the space. Define a kernel $k(\cdot,\cdot) : \mathcal{X} \times \mathcal{X} \rightarrow \mathcal{R}$. For any positive definite kernel $k(\cdot,\cdot)$, there exists a unique space of functions $f : \mathcal{X} \rightarrow \mathcal{R}$ called the reproducing kernel Hilbert space (RKHS), $\mathcal{H}_k$, which is a Hilbert space and satisfies $f(x) = \langle f, k(\cdot, x)\rangle_{\mathcal{H}_k}, \forall x \in \mathcal{X}, \forall f \in \mathcal{H}_k$. We further define a feature mapping that maps a point in the original space into a feature vector in the RKHS $\phi_k(x) : \mathcal{X} \rightarrow \mathcal{H}_k$ such that $\phi_k(x) = k(\cdot, x)$. The reproducing property of a RKHS indicates $\langle \phi_k(x), \phi_k(x')\rangle_{\mathcal{H}_{\|}} = \langle k(\cdot, x), k(\cdot, x')\rangle_{\mathcal{H}_k} = k(x, x')$. A kernel operation $y = k(x, x')$ takes two points in the original space (x, x') and output a real number y, which could be viewed as transforming the data into feature vectors $(\phi_k(x), \phi_k(x'))$, and then performing dot product in a RKHS. A large y indicates a short distance of the points (x, x') measured by dot product over the feature vectors in the RKHS.

Kernel MMD Kernel maximum mean discrepancy is a probability metric, which is firstly proposed in two-sample-test, distinguishing two distributions using finite samples (Gretton et al. 2012). Generative moment matching networks (GMMN) (Li, Swersky, and Zemel 2015) uses kernel maximum mean discrepancy to match a target distribution. It is proved that the $M_k^2(p, q) = 0$ iff $p = q$ for characteristic kernels (Sriperumbudur, Fukumizu, and Lanckriet 2011). Despite this appealing properties, using a fixed kernel often leads to poor performance (Dziugaite, Roy, and Ghahramani 2015; Bottou et al. 2018). Recent MMD GANs achieved better result estimating the MMD using a learnable kernel. (Li et al. 2017) measures MMD with $\max_\varphi M_\varphi^2(p, q)$, which considers a composition kernel k_φ combining Gaussian kernel k and a injected function f_φ. (Li et al. 2019) estimates MMD with $\max_{\varphi,\varphi} M_{\varphi,\varphi}^2(p, q)$, using an implicit kernel learning $k_{\psi,\varphi}$ combining learnable base kernel k_ψ and a injected function f_φ. In this paper, we propose a MMD with variational kernel learning, $\max_{E,\psi,\varphi} M_{E,\psi,\varphi}^2(p, q)$, which is sample efficient and robust to overfitting. Moreover, we established the connection between a MMD with kernel learning and an imitation learning objective. As far as our knowledge, we are the first to solve an imitation learning problem using MMD with kernel learning.

Proof of Theorems

Lemma 1. *Let $max_{E,\psi,\varphi} M_{E,\psi,\varphi}^2(p, q)$ be the MMD with variational kernel learning, and $max_{\psi,\varphi} M_{\psi,\varphi}^2(p, q)$ be the MMD with kernel learning. Given two distributions p, q, let p' and q' be the encoding distributions $p'(z) = E(z \mid x), x \sim p$, and $q'(z) = E(z \mid x), x \sim q$. Then $M_{E,\psi,\varphi}^2(p, q) = M_{\psi,\varphi}^2(p', q')$.*

Proof.

$$\begin{aligned} M_{E,\psi,\varphi}^2(p,q) &= \mathbb{E}_{x,x'\sim p}\left[\mathbb{E}_{z\sim E(z|x),z'\sim E(z'|x')}\left[k_{\psi,\varphi}(z,z')\right]\right] \\ &+\mathbb{E}_{y,y'\sim q}\left[\mathbb{E}_{z\sim E(z|y),z'\sim E(z'|y')}\left[k_{\psi,\varphi}(z,z')\right]\right] \\ &-2\mathbb{E}_{x\sim p,y\sim q}\left[\mathbb{E}_{z\sim E(z|x),z'\sim E(z'|y)}\left[k_{\psi,\varphi}(z,z')\right]\right] \\ &= \mathbb{E}_{x\sim p,z\sim E(z|x)}\mathbb{E}_{x'\sim p,z'\sim E(z'|x')}\left[k_{\psi,\varphi}(z,z')\right] \\ &+\mathbb{E}_{y\sim q,z\sim E(z|y)}\mathbb{E}_{y'\sim q,z'\sim E(z'|y')}\left[k_{\psi,\varphi}(z,z')\right] \\ &-2\mathbb{E}_{x\sim p,z\sim E(z|x)}\mathbb{E}_{p\sim q,z'\sim E(z'|y)}\left[k_{\psi,\varphi}(z,z')\right] \\ &= \mathbb{E}_{z,z'\sim p'}\left[k_{\psi,\varphi}(z,z')\right]+\mathbb{E}_{z,z'\sim q'}\left[k_{\psi,\varphi}(z,z')\right]-2\mathbb{E}_{z\sim p',z'\sim q'}\left[k_{\psi,\varphi}(z,z')\right] \\ &= M_{\psi,\varphi}^2(p',q') \end{aligned}$$

□

Theorem 2. *Assume 1) the function $f_\varphi(x)$ is bounded and has a common Lipschitz constant $\sup_\varphi \| f_\varphi(x) \|_L \leq L_\varphi < \infty$; 2) the variance of function $h_\psi(\nu)$ is bounded $\mathbb{E}_\nu \left[\| h_\psi(\nu) \|^2\right] < \infty$; 3) and the kernel is bounded $sup_x k_{\psi,\varphi}(x, x) < \infty$. Let p' and q' be the encoding distributions, i.e., $p'(z) = E(z \mid x), x \sim p$, and $q'(z) = E(z \mid x), x \sim q$, then $max_{E,\psi,\varphi} M_{E,\psi,\varphi}^2(p, q)$ is continuous in the weak topology for the encodings z:*

$$\max_{E,\psi,\varphi} M_{E,\psi,\varphi}^2(p,q) \rightarrow 0 \Longleftrightarrow p' \overset{D}{\rightarrow} q'$$

Proof. To outline our proof, we firstly prove $p' \overset{D}{\rightarrow} q' \Longrightarrow \max_{E,\psi,\varphi} M_{E,\psi,\varphi}^2(p, q) \rightarrow 0$, and then prove the other direction, $max_{E,\psi,\varphi} M_{E,\psi,\varphi}^2(p, q) \rightarrow 0 \Longrightarrow p(z) \overset{D}{\rightarrow} p(z')$. In our first proof, we firstly prove $p' \overset{D}{\rightarrow} q' \Longrightarrow \max_{\psi,\varphi} M_{\psi,\varphi}^2(p', q') \rightarrow 0$, and then use Lemma 1 to prove $\max_{\psi,\varphi} M_{\psi,\varphi}^2(p', q') \rightarrow 0 \Longrightarrow \max_{E,\psi,\varphi} M_{E,\psi,\varphi}^2(p, q) \rightarrow 0$.

First we prove $p' \overset{D}{\rightarrow} q' \Longrightarrow \max_{E,\psi,\varphi} M_{E,\psi,\varphi}^2(p, q) \rightarrow 0$.

Let $\max_{\psi,\varphi} M_{\psi,\varphi}^2(p', q')$ be the MMD with kernel learning on the encodings $z \sim p', z' \sim q'$, we want to prove $p' \overset{D}{\rightarrow} q' \Longrightarrow \max_{\psi,\varphi} M_{\psi,\varphi}^2(p', q') \rightarrow 0$. Following the sketch in (Arbel et al. 2018), the only thing left to show is that $\| k_{\psi,\varphi}(x, \cdot) - k_{\psi,\varphi}(x', \cdot) \|_{\mathcal{H}_k}$ is Lipschitz. Since $\mathbb{E}_\nu \left[\| h_\psi(\nu) \|^2\right] < \infty$, and $f_\varphi(x)$ is bounded, then

$k_{\psi,\varphi}(x,x') \quad = \mathbb{E}_\nu \left[e^{ih_\psi(\nu)^T(f_\varphi(x)-f_\varphi(x'))} \right]$ is Lipschitz in $f_\varphi(x)$ (Li et al. 2019). Let the Lipschitz constant of $k_{\psi,\varphi}(x,x')$ be L_k, we have

$$\| k_{\psi,\varphi}(x,\cdot) - k_{\psi,\varphi}(x',\cdot) \|_{\mathcal{H}_k} \leq L_k \| f_\varphi(x) - f_\varphi(x') \| \leq L_k L_\varphi \| x - x' \|$$

As such,

$$p' \xrightarrow{D} q' \Longrightarrow \max_{\psi,\varphi} M^2_{\psi,\varphi}(p',q') \to 0$$

As proved by Lemma 1, we have $M^2_{E,\psi,\varphi}(p,q) = M^2_{\psi,\varphi}(p',q')$. As such,

$$\max_{\psi,\varphi} M^2_{\psi,\varphi}(p',q') \to 0 \Longrightarrow \max_{E,\psi,\varphi} M^2_{E,\psi,\varphi}(p,q) \to 0$$

Next we proof the other direction, $max_{E,\psi,\varphi} M^2_{E,\psi,\varphi}(p,q) \to 0 \Longrightarrow p; \xrightarrow{D} q'$

$$\begin{aligned} & \max_{E,\psi,\varphi} M^2_{E,\psi,\varphi}(p,q) \to 0 \\ \Rightarrow \quad & \max_{\psi,\varphi} M^2_{\psi,\varphi}(p',q') \to 0 \\ \Rightarrow \quad & M^2_{\psi',\varphi'}(p',q') \to 0 \\ \Rightarrow \quad & p' \xrightarrow{D} q' \end{aligned}$$

Where the 3rd step is due to that since the max operation is over all (ψ,φ), without loss of generality, we assume there exists parameterizations (ψ',φ') that recovers a Gaussian kernel k, the MMD with which is denoted as $M^2_{\psi',\varphi'}$. The 4th step is due to that MMD with any Gaussian kernel is weak (Gretton et al. 2012).

□

Theorem 3. *Solving the IL problem described by* (7) *is equivalent to solving a regularized IL problem with a cost function defined over the stochastic encoding $z \sim \mathbb{E}(z \mid s,a)$, with the cost function class being*

$$\begin{aligned} \mathcal{C}_{\mathcal{H}_{k_{\psi,\varphi}}} = \Bigg\{ & c(s,a) = \mathbb{E}_{z\sim E(z|s,a)}\left[c(z)\right] \\ & = \left\langle c, \mathbb{E}_{z\sim E(z|s,a)} \phi_{k_{\psi,\varphi}}(z) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \mid \forall \psi, \varphi \Bigg\}, \end{aligned}$$

and the regularizations being $I(X,Z) \leq I_c, \mathbb{E}_\nu[\| h_\psi(\nu) \|^2] < \infty, \| E \|_L \leq 1, \| f_\varphi \|_L \leq 1$.

Proof. Given the cost function class

$$\mathcal{C}_{\mathcal{H}_{k_{\psi,\varphi}}} = \left\{ c(s,a) = \mathbb{E}_{z\sim E(z|s,a)}\left[c(z)\right] = \left\langle c, \mathbb{E}_{z\sim E(z|s,a)} \phi_{k_{\psi,\varphi}}(z) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \mid \forall \psi, \varphi \right\}$$

the imitation learning objective becomes

$$\begin{aligned} \text{IL}(\pi_E) \quad & = \min_\pi \max_{c,E,\psi,\varphi} \left\langle c, \mathbb{E}_{s,a\sim\rho_\pi, z\sim\mathbb{E}(z|s,a)} \phi_{k_{\psi,\varphi}}(z) - \mathbb{E}_{s,a\sim\rho_{\pi_E}, z\sim\mathbb{E}(z|s,a)} \phi_{k_{\psi,\varphi}}(z) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\ & - \tfrac{1}{2} \| c \|^2_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \end{aligned}$$

Adding the regularizations $I(X,Z) \leq I_c, \mathbb{E}_\nu \left[\| h_\psi(\nu) \|^2\right] < \infty, \| E \|_L \leq 1, \| f_\varphi \|_L \leq 1$, it becomes

$$\begin{aligned} \text{IL}(\pi_E) \quad & = \min_\pi \max_{c,E,\psi,\varphi:\|E\|_L\leq 1, \|f_\varphi\|_L \leq 1} \\ & \left\langle c, \mathbb{E}_{s,a\sim\rho_\pi, z\sim\mathbb{E}(z|s,a)} \phi_{k_{\psi,\varphi}}(z) - \mathbb{E}_{s,a\sim\rho_{\pi_E}, z\sim\mathbb{E}(z|s,a)} \phi_{k_{\psi,\varphi}}(z) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\ & - \beta I(X,Z) - \lambda_h \mathbb{E}_\nu \left[\| h_\psi(\nu) \|^2\right] - \tfrac{1}{2} \| c \|^2_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \end{aligned}$$

Taking the derivative w.r.t c and set to 0, the optimal cost function c^* takes the form

$$\begin{aligned} & \mathbb{E}_{s,a\sim\rho_\pi, z\sim\mathbb{E}(z|s,a)} \phi_{k_{\psi,\varphi}}(z) - \mathbb{E}_{s,a\sim\rho_{\pi_E}, z\sim\mathbb{E}(z|s,a)} \phi_{k_{\psi,\varphi}}(z) - c^* = 0 \\ \Rightarrow \quad & c^* = \mathbb{E}_{s,a\sim\rho_\pi, z\sim\mathbb{E}(z|s,a)} \phi_{k_{\psi,\varphi}}(z) - \mathbb{E}_{s,a\sim\rho_{\pi_E}, z\sim\mathbb{E}(z|s,a)} \phi_{k_{\psi,\varphi}}(z) \end{aligned}$$

Take in c^*, the imitation learning objective becomes

$$\begin{aligned}
\mathrm{IL}(\pi_E) \quad &= \min_\pi \max_{E,\psi,\varphi:\|E\|_L\le 1,\|f_\varphi\|_L\le 1} \\
&\left\langle c^*, \mathbb{E}_{s,a\sim\rho_\pi,z\sim\mathbb{E}(z|s,a)}\phi_{k_{\psi,\varphi}}(z) - \mathbb{E}_{s,a\sim\rho_{\pi_E},z\sim\mathbb{E}(z|s,a)}\phi_{k_{\psi,\varphi}}(z)\right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\
&-\beta I(X,Z) - \lambda_h \mathbb{E}_\nu\left[\| h_\psi(\nu) \|^2\right] - H(\pi) \\
&= \min_\pi \max_{E,\psi,\varphi:\|E\|_L\le 1,\|f_\varphi\|_L\le 1} \\
&\langle \mathbb{E}_{s,a\sim\rho_\pi,z\sim\mathbb{E}(z|s,a)}\phi_{k_{\psi,\varphi}}(z), \mathbb{E}_{s,a\sim\rho_\pi,z\sim\mathbb{E}(z|s,a)}\phi_{k_{\psi,\varphi}}(z)\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\
&+\left\langle \mathbb{E}_{s,a\sim\rho_{\pi_E},z\sim\mathbb{E}(z|s,a)}\phi_{k_{\psi,\varphi}}(z), \mathbb{E}_{s,a\sim\rho_{\pi_E},z\sim\mathbb{E}(z|s,a)}\phi_{k_{\psi,\varphi}}(z)\right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\
&-2\left\langle \mathbb{E}_{s,a\sim\rho_\pi,z\sim\mathbb{E}(z|s,a)}\phi_{k_{\psi,\varphi}}(z), \mathbb{E}_{s,a\sim\rho_{\pi_E},z\sim\mathbb{E}(z|s,a)}\phi_{k_{\psi,\varphi}}(z)\right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\
&-\beta I(X,Z) - \lambda_h \mathbb{E}_\nu\left[\| h_\psi(\nu) \|^2\right] - H(\pi) \\
&= \min_\pi \max_{E,\psi,\varphi:\|E\|_L\le 1,\|f_\varphi\|_L\le 1} \mathbb{E}_{s,a,s',a'\sim\rho_\pi}\left[\mathbb{E}_{z\sim E(z|s,a),z'\sim E(z'|s',a')}\left[k_{\psi,\varphi}(z,z')\right]\right] \\
&+\mathbb{E}_{s,a,s',a'\sim\rho_{\pi_E}}\left[\mathbb{E}_{z\sim E(z|s,a),z'\sim E(z'|s',a')}\left[k_{\psi,\varphi}(z,z')\right]\right] \\
&-2\mathbb{E}_{s,a\sim\rho_\pi,s',a'\sim\rho_{\pi_E}}\left[\mathbb{E}_{z\sim E(z|s,a),z'\sim E(z'|s',a')}\left[k_{\psi,\varphi}(z,z')\right]\right] \\
&-\beta I(X,Z) - \lambda_h \mathbb{E}_\nu\left[\| h_\psi(\nu) \|^2\right] - H(\pi) \\
&= \min_\pi \max_{E,\psi,\varphi:\|E\|_L\le 1,\|f_\varphi\|_L\le 1} M^2_{E,\psi,\varphi}(\rho_\pi, \rho_{\pi_E}) \\
&-\beta I(X,Z) - \lambda_h \mathbb{E}_\nu\left[\| h_\psi(\nu) \|^2\right] - H(\pi)
\end{aligned}$$

□

Theorem 4. *Let the policy π_θ be parameterized by θ, and $\epsilon \sim \mathcal{N}(0, I)$. The gradient of the policy optimization in VAKLIL has the form*

$$\begin{aligned}
&\nabla_\theta\left(\mathbb{E}_{x\sim\rho_{\pi_\theta}}\left[\hat{c}(x)\right] - H(\pi_\theta)\right) \\
=&\nabla_\theta \mathbb{E}_{x\sim\rho_{\pi_\theta},x'\sim\rho_{\pi_\theta},\epsilon}\left\langle \phi_{k_{\psi,\varphi}}(\mu_E(x') + \epsilon\Sigma_E^{1/2}(x')), \phi_{k_{\psi,\varphi}}(\mu_E(x))\right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\
&-\nabla_\theta \mathbb{E}_{x\sim\rho_{\pi_\theta},x''\sim\rho_{\pi_E},\epsilon}\left\langle \phi_{k_{\psi,\varphi}}(\mu_E(x'') + \epsilon\Sigma_E^{1/2}(x'')), \phi_{k_{\psi,\varphi}}(\mu_E(x))\right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\
&-\nabla_\theta H(\pi_\theta)\ .
\end{aligned}$$

Proof.

$$\begin{aligned}
&\nabla_\theta\left(\mathbb{E}_{x\sim\rho_{\pi_\theta}}\left[\hat{c}(x)\right] - H(\pi_\theta)\right) \\
&= \nabla_\theta \mathbb{E}_{x\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{x'\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{z'\sim E(z'|x')}k_{\psi,\varphi}\left(z', \mu_E(x)\right)\right]\right] \\
&-\nabla_\theta \mathbb{E}_{x\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{x''\sim\rho_{\pi_E}}\left[\mathbb{E}_{z''\sim E(z''|x'')}k_{\psi,\varphi}\left(z'', \mu_E(x)\right)\right]\right] - \nabla_\theta H(\pi_\theta) \\
&= \nabla_\theta \mathbb{E}_{x\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{x'\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{z'\sim E(z'|x')}\left\langle \phi_{k_{\psi,\varphi}}(z'), \phi_{k_{\psi,\varphi}}(\mu_E(x))\right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}}\right]\right] \\
&-\nabla_\theta \mathbb{E}_{x\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{x''\sim\rho_{\pi_E}}\left[\mathbb{E}_{z''\sim E(z''|x'')}\left\langle \phi_{k_{\psi,\varphi}}(z''), \phi_{k_{\psi,\varphi}}(\mu_E(x))\right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}}\right]\right] - \nabla_\theta H(\pi_\theta) \\
&= \nabla_\theta \mathbb{E}_{x\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{x'\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left\langle \phi_{k_{\psi,\varphi}}(\mu_E(x') + \epsilon\Sigma_E^{1/2}(x')), \phi_{k_{\psi,\varphi}}(\mu_E(x))\right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}}\right]\right] \\
&-\nabla_\theta \mathbb{E}_{x\sim\rho_{\pi_\theta}}\left[\mathbb{E}_{x''\sim\rho_{\pi_E}}\left[\mathbb{E}_{\epsilon\sim\mathcal{N}(0,I)}\left\langle \phi_{k_{\psi,\varphi}}(\mu_E(x'') + \epsilon\Sigma_E^{1/2}(x'')), \phi_{k_{\psi,\varphi}}(\mu_E(x))\right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}}\right]\right] - \nabla_\theta H(\pi_\theta)
\end{aligned}$$

□

Theorem 5. *Let $p_k(\omega) = p(h_\psi(\nu))$ be the spectral distribution.*

- *1) If $p_k(\omega)$ is fixed to be $(2\pi)^{-\frac{D}{2}}e^{-\frac{\|\omega\|^2}{2}}$, the kernel $k_{\psi,\varphi}(x,x')$ becomes a Gaussian kernel $k_\varphi(x,x') = e^{-\frac{(f_\varphi(x)-f_\varphi(x'))^2}{2}}$.*
- *2) If the kernel is estimated using random samples $\omega_i \sim p_k(\omega)$ and the complex exponential is replaced with cosine, then it becomes a random Fourier kernel $k_{\psi,\varphi}(x,x') = \kappa(x)\kappa(x')$, where $\kappa(x)=\sqrt{\frac{1}{D}}\left[cos\left(h_\psi(\nu_1)^T f_\varphi(x)+b_1\right),\ldots,cos\left(h_\psi(\nu_D)^T f_\varphi(x)+b_D\right)\right]'$.*

Proof. If $p_k(\omega) = p(h_\psi(\nu)) = (2\pi)^{-\frac{D}{2}} e^{-\frac{\|\omega\|^2}{2}}$, then

$$\begin{aligned} k_{\psi,\varphi}(x,x') &= \mathbb{E}_\nu \left[e^{i h_\psi(\nu)^T (f_\varphi(x) - f_\varphi(x'))} \right] \\ &= \mathbb{E}_\omega \left[e^{i\omega^T (f_\varphi(x) - f_\varphi(x'))} \right] \\ &= \int p(\omega) e^{i\omega^T (f_\varphi(x) - f_\varphi(x'))} d\omega \\ &= \int (2\pi)^{-\frac{D}{2}} e^{-\frac{\|\omega\|^2}{2}} e^{i\omega^T (f_\varphi(x) - f_\varphi(x'))} d\omega \\ &= e^{-\frac{(f_\varphi(x) - f_\varphi(x'))^2}{2}} \end{aligned}$$

If we evaluate the kernel using random samples $\omega_i \sim p_k(\omega)$, then

$$\begin{aligned} k_{\psi,\varphi}(x,x') &= \mathbb{E}_\nu \left[e^{i h_\psi(\nu)^T (f_\varphi(x) - f_\varphi(x'))} \right] \\ &= \frac{1}{D} \textstyle\sum_{j=1}^{D} e^{i h_\psi(\nu_j)^T (f_\varphi(x) - f_\varphi(x'))} \\ &= \frac{1}{D} \textstyle\sum_{j=1}^{D} e^{i h_\psi(\nu_j)^T f_\varphi(x) + b_j} e^{-i h_\psi(\nu_j)^T f_\varphi(x') - b_j} \\ &= \frac{1}{D} \textstyle\sum_{j=1}^{D} \cos\left(h_\psi(\nu_j)^T f_\varphi(x) + b_j \right) \cos\left(-h_\psi(\nu_j)^T f_\varphi(x') - b_j \right) \\ &= \frac{1}{D} \textstyle\sum_{j=1}^{D} \cos\left(h_\psi(\nu_j)^T f_\varphi(x) + b_j \right) \cos\left(h_\psi(\nu_j)^T f_\varphi(x') + b_j \right) \\ &= \kappa(x)\kappa(x') \end{aligned}$$

where $\kappa(x) = \sqrt{\frac{1}{D}} \left[\cos\left(h_\psi(\nu_1)^T f_\varphi(x) + b_1 \right), ..., \cos\left(h_\psi(\nu_D)^T f_\varphi(x) + b_D \right) \right]'$

□

More Details of the Proposed Algorithm VAKLIL

Denote the state-action pair (s, a) as x. In kernel learning, we update E, ψ, φ with gradient ascent to maximize the objective $J = M^2_{E,\psi,\varphi}(\rho_\pi, \rho_{\pi_E}) - \beta I(X, Z) - \lambda_h \mathbb{E}_\nu \left[\| h_\psi(\nu) \|^2 \right]$, where the mutual information $I(X, Z)$ takes the form $I(X, Z) \leq \mathbb{E}_{x \sim \tilde{\rho}(x)} \left[D_{\text{KL}} \left[E(z \mid x) \parallel r(z) \right] \right]$, $\tilde{\rho}(x) = \frac{1}{2}(\rho_\pi + \rho_{\pi_E})$ is the mixed distribution, and $r(z) \sim \mathcal{N}(0, I)$ models a normal distribution. We apply the spectral normalization to the Encoder $E(z \mid x)$ and the function $f_\varphi(z)$ to satisfy the Lipschitz constraints $\| \ E \ \|_L \leq 1, \| \ f_\varphi \ \|_L \leq 1$. In policy optimization, given parameters E, ψ, φ, the optimal cost function c^* takes the form $c^* = \mathbb{E}_{x \sim \rho_\pi, z \sim \mathbb{E}(z|x)} \phi_{k_{\psi,\varphi}}(z) - \mathbb{E}_{x \sim \rho_{\pi_E}, z \sim \mathbb{E}(z|x)} \phi_{k_{\psi,\varphi}}(z)$. The policy is optimized according to arg $\min_\pi \mathbb{E}_{x \sim \rho_\pi} \left[\hat{c}(x) \right] - H(\pi)$, which we can solve with any maximum entropy reinforcement learning algorithms with cost $\hat{c}(x) = \mathbb{E}_{x' \sim \rho_\pi, z' \sim E(z'|x')} \left[k_{\psi,\varphi}(\mu_E(x), z') \right] - \mathbb{E}_{x' \sim \rho_{\pi_E},, z' \sim E(z'|x')} \left[k_{\psi,\varphi}(\mu_E(x), z') \right]$, where $\mu_E(x)$ is the mean of the encoder $E(z \mid x)$.

Derivation of Algorithms

Imitation Learning with a Broader Cost Function Class

Derivation of Eq. 4 and the optimal cost function c^*:

For the objective

$$\text{IL}(\pi_E) \quad = \min_\pi \max_{c,\psi,\varphi} \left\langle c, \mathbb{E}_{x \sim \rho_\pi} \phi_{k_{\psi,\varphi}}(x) - \mathbb{E}_{x \sim \rho_{\pi_E}} \phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} - \frac{1}{2} \| \ c \ \|^2_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi)$$

Take derivative w.r.t. c and set to 0,

$$\begin{aligned} & \mathbb{E}_{x \sim \rho_\pi} \phi_{k_{\psi,\varphi}}(x) - \mathbb{E}_{x \sim \rho_{\pi_E}} \phi_{k_{\psi,\varphi}}(x) - c^* = 0 \\ \Rightarrow \quad & c^* = \mathbb{E}_{x \sim \rho_\pi} \phi_{k_{\psi,\varphi}}(x) - \mathbb{E}_{x \sim \rho_{\pi_E}} \phi_{k_{\psi,\varphi}}(x) \end{aligned}$$

Given c^*, the above imitation learning objective is equivalent to

$$\begin{aligned} \text{IL}(\pi_E) \quad &= \min_\pi \max_{\psi,\varphi} \left\langle c^*, \mathbb{E}_{x \sim \rho_\pi} \phi_{k_{\psi,\varphi}}(x) - \mathbb{E}_{x \sim \rho_{\pi_E}} \phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \\ &= \min_\pi \max_{\psi,\varphi} \left\langle \mathbb{E}_{x \sim \rho_\pi} \phi_{k_{\psi,\varphi}}(x), \mathbb{E}_{x \sim \rho_\pi} \phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\ &+ \left\langle \mathbb{E}_{x \sim \rho_{\pi_E}} \phi_{k_{\psi,\varphi}}(x), \mathbb{E}_{x \sim \rho_{\pi_E}} \phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\ &-2 \left\langle \mathbb{E}_{x \sim \rho_\pi} \phi_{k_{\psi,\varphi}}(x), \mathbb{E}_{x \sim \rho_{\pi_E}} \phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \\ &= \min_\pi \max_{\psi,\varphi} \mathbb{E}_{x,x' \sim \rho_\pi} \left[k_{\psi,\varphi}(x,x') \right] + \mathbb{E}_{x,x' \sim \rho_{\pi_E}} \left[k_{\psi,\varphi}(x,x') \right] \\ &-2 \mathbb{E}_{x \sim \rho_\pi, x' \sim \rho_{\pi_E}} \left[k_{\psi,\varphi}(x,x') \right] - H(\pi) \\ &= \min_\pi \max_{\psi,\varphi} M^2_{\psi,\varphi}(\rho_\pi, \rho_{\pi_E}) - H(\pi) \end{aligned}$$

Derivation of optimal policy π^*:

In policy optimization, we fix the kernel parameters ψ, φ, and the optimal cost function c^* induced by kernel $k_{\psi,\varphi}$, then the objective becomes

$$\begin{aligned}
\pi^* &= \text{arg min}_\pi \left\langle c^*, \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(x) - \mathbb{E}_{x\sim\rho_{\pi_E}}\phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \\
&= \text{arg min}_\pi \left\langle c^*, \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \\
&= \text{arg min}_\pi \left\langle \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(x) - \mathbb{E}_{x\sim\rho_{\pi_E}}\phi_{k_{\psi,\varphi}}(x), \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \\
&= \text{arg min}_\pi \left\langle \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(x), \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\
&- \left\langle \mathbb{E}_{x\sim\rho_{\pi_E}}\phi_{k_{\psi,\varphi}}(x), \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(x) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \\
&= \text{arg min}_\pi \mathbb{E}_{x,x'\sim\rho_\pi}\left[k_{\psi,\varphi}(x,x')\right] - \mathbb{E}_{x\sim\rho_\pi,x'\sim\rho_{\pi_E}}\left[k_{\psi,\varphi}(x,x')\right] - H(\pi) \\
&= \text{arg min}_\pi \mathbb{E}_{x\sim\rho_\pi}\left[\mathbb{E}_{x'\sim\rho_\pi}\left[k_{\psi,\varphi}(x,x')\right] - \mathbb{E}_{x'\sim\rho_{\pi_E}}\left[k_{\psi,\varphi}(x,x')\right]\right] - H(\pi) \\
&= \text{arg min}_\pi \mathbb{E}_{x\sim\rho_\pi}\left[\hat{c}(x)\right] - H(\pi)
\end{aligned}$$

Sample Efficient and Robust Imitation Learning

The derivations of Eq. 7 and the optimal cost function c^* are given in the proof of Theorem 3.

Derivation of the mutual information $I(X,Z)$:

$$\begin{aligned}
I(X,Z) &= \int p(x,z)\log\tfrac{p(x,z)}{p(x)p(z)}dxdz \\
&= \int p(x)E(z \mid x)\log\tfrac{E(z|x)}{p(z)}dxdz \\
&\leq \int p(x)E(z \mid x)\log\tfrac{E(z|x)}{r(z)}dxdz \\
&= \mathbb{E}_{x\sim p(x)}\left[D_{\text{KL}}\left[E(z \mid x) \parallel r(z)\right]\right]
\end{aligned}$$

As such, a variational lower bound can be obtained by using an approximation $r(z)$ of the marginal. Practically, we choose prior $r(z) = \mathcal{N}(0, I)$, and use a mixed distribution $\frac{1}{2}(\rho_\pi + \rho_{\pi_E})$ for $p(x)$.

Derivation of the optimal policy π^*:

Given parameters E, ψ, φ, the optimal cost function c^* takes the form $c^* = \mathbb{E}_{x\sim\rho_\pi, z\sim\mathbb{E}(z|x)}\phi_{k_{\psi,\varphi}}(z) - \mathbb{E}_{x\sim\rho_{\pi_E}, z\sim\mathbb{E}(z|x)}\phi_{k_{\psi,\varphi}}(z)$. We optimize the policy using a simplified objective, where the mutual information constraint is excluded. Moreover, instead of evaluating the cost function as an expectation of encodings $c(x) = \mathbb{E}_{z\sim\mathbb{E}(z|x)}\left[c(z)\right]$, we evaluate it at the mean $c(x) = c(\mu_E(x))$. The reason is analogous to the dropout which is used in training but is disabled in evaluation. We incorporate a controllable level of uncertainty in kernel learnings to make the parameter updating more robust. In policy optimization, we fix the kernel parameters and want the evaluation of $c(x)$ being consistent, thus disabling the injected noise and evaluating cost as $c(x) = c(\mu_E(x))$.

$$\begin{aligned}
\pi^* &= \text{arg min}_\pi \left\langle c^*, \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(\mu_E(x)) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \\
&= \text{arg min}_\pi \left\langle \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(\mu_E(x)), \mathbb{E}_{x\sim\rho_\pi, z\sim\mathbb{E}(z|x)}\phi_{k_{\psi,\varphi}}(z) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} \\
&- \left\langle \mathbb{E}_{x\sim\rho_\pi}\phi_{k_{\psi,\varphi}}(\mu_E(x)), \mathbb{E}_{x\sim\rho_{\pi_E}, z\sim\mathbb{E}(z|x)}\phi_{k_{\psi,\varphi}}(z) \right\rangle_{\mathcal{H}_{k_{\psi,\varphi}}} - H(\pi) \\
&= \text{arg min}_\pi \mathbb{E}_{x,x'\sim\rho_\pi, z'\sim E(z'|x')}\left[k_{\psi,\varphi}(\mu_E(x), z')\right] \\
&-\mathbb{E}_{x\sim\rho_\pi, x'\sim\rho_{\pi_E},, z'\sim E(z'|x')}\left[k_{\psi,\varphi}(\mu_E(x), z')\right] - H(\pi) \\
&= \text{arg min}_\pi \mathbb{E}_{x\sim\rho_\pi}\left[\mathbb{E}_{x'\sim\rho_\pi, z'\sim E(z'|x')}\left[k_{\psi,\varphi}(\mu_E(x), z')\right] - \mathbb{E}_{x'\sim\rho_{\pi_E},, z'\sim E(z'|x')}\left[k_{\psi,\varphi}(\mu_E(x), z')\right]\right] \\
&-\mathbb{E}_{x\sim\rho_\pi, x'\sim\rho_{\pi_E},, z'\sim E(z'|x')}\left[k_{\psi,\varphi}(\mu_E(x), z')\right] - H(\pi) \\
&= \text{arg min}_\pi \mathbb{E}_{x\sim\rho_\pi}\left[\hat{c}(x)\right] - H(\pi)
\end{aligned}$$

where $\hat{c}(x) = \mathbb{E}_{x'\sim\rho_\pi, z'\sim E(z'|x')}\left[k_{\psi,\varphi}(\mu_E(x), z')\right] - \mathbb{E}_{x'\sim\rho_{\pi_E},, z'\sim E(z'|x')}\left[k_{\psi,\varphi}(\mu_E(x), z')\right]$

Experiment

Descriptions of Environments

We evaluate our algorithm over five OpenAI Gym environments, Ant (RoboschoolAnt-v1), HalfCheetah (RoboschoolHalfCheetah-v1), Humanoid (RoboschoolHumanoid-v1), HumanoidFlagrun (RoboschoolHumanoidFlagrun-v1), Walker2D (RoboschoolWalker2d-v1), and a complex transportation environment. The number of states and number of actions in each environment are described in Table 3.

Environment	HalfCheetah	Ant	Walker2D	Humanoid	HumanoidFlagrun	Transportation
states	26	28	22	44	44	101
actions	6	8	6	17	17	200

Table 3: Descriptions of environments

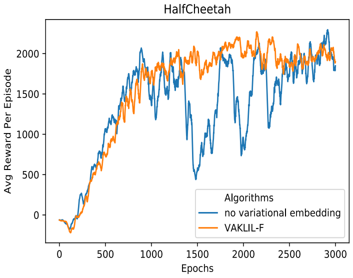

Figure 3: The comparison of VAKLIL-F with and without the variational embedding.

The transportation task involves controlling the movement of vehicles to maximize the utility function of the entire system. The environment is implemented based on MATSim (Ciari 2015), an open-source high-fidelity large-scale agent-based transportation simulator and framework. This transportation environment contains 100 vehicles, 23 roads, and 2 facilities. It has 101 discrete state variables which includes 100 variables to describe the location of the vehicles and 1 variable to indicate the current time, and 200 continuous action variables to control which roads will the vehicles move to, and when the vehicles will enter or leave facilities. One typical simulation run contains 1440 time steps — one day in minutes. The state transition is implemented following the fundamental diagram of traffic flow that simulates the movement of vehicles in different traffic conditions, The reward function emulates the Charypar-Nagel scoring function (Ciari 2015) which rewards performing the correct activities at the correct time, and penalizes traveling on roads and late arrival.

Experimental Setup

The policy networks are implemented as 2-hidden layer neural networks for the all the tasks. To be specific, for the five OpenAI Gym tasks, the dimension of the first hidden layer of the policy network is 64, and that of the second hidden layer is 32; for the transportation task, the dimension of all the hidden layers are 128. All neural networks implement the "relu" activation layer.

The discriminator networks are implemented according to the nature of each algorithm. The discriminator networks for the GAIL and AL are implemented as 2-hidden layer neural networks. The encoder and the discriminator for VAIL are both implemented as 1-hidden layer neural networks. The encoder E, the injected function f_φ and h_ψ in VAKLIL-F and VAKLIL-G are all implemented as 1-hidden layer neural networks.

We collect 10-20 expert trajectories for each environment. To be specific, we collect 12, 12, 12, 10, 10, 20 trajectories for HalfCheetah, Ant, Walker2D, Humanoid, HumanoidFlagrun, and Transportation, respectively. For each task and each algorithm, we repeat the experiment 3-6 times. The performance is reported with the mean and variance of these experimental runs at different evaluation metrics. In each experiment, we iteratively update the discriminator/kernels and the policy for 2000-4000 epochs. In each epoch, we firstly sample the state-action pairs from interacting with the environment, and then update the discriminator/kernel for 10 gradient descent steps with the same samples, finally update the policy with 5 TRPO steps with the same samples.

We run the experiments using the NVIDIA TESLA V100 TENSOR CORE 16GB GPU. The typical run time for one experiment with the OpenAI Gym environment is several hours, and that for the transportation environment is a day.

Additional Experiments

Our algorithm contains two novel components: (i) using a *learned* kernel for MMD vs a fixed one, and (ii) using a variational embedding. To demonstrate the contribution of these two components, we conduct a further experiment. In the following, we use "1" for using a *learned* kernel for MMD vs a fixed one, and "2" for using a variational embedding. Figure 3 showed "2" makes the learning more stable. If only "2" is considered, this reduces to VAIL (Peng et al. 2018), which is GAIL plus variational bottleneck, and has been compared in the paper.