Broader Impact

MISA framework provides a simple yet effective approach to policy learning from offline datasets. Although the results presented in this paper only consider simulated environments, given the generality of MISA, it could be potentially effective on learning real-robot policies in more complex environments. We should be cautious about the misuse of the method proposed. Depending on the specific application scenarios, it might be harmful to democratic privacy and safety.

A Proofs and Derivations

A.1 Proof for Theorem 4.1

We first show $\mathcal{I}_{\text{MISA}}$, $\mathcal{I}_{\text{MISA-DV}}$ and $\mathcal{I}_{\text{MISA-}f}$ are lower bounds for mutual information $I(S, A)$.

Let $\mu_{\theta,\phi}(a|s) \triangleq \frac{1}{\mathcal{Z}(s)}\pi_\theta(a|s)e^{T_\phi(s,a)}$, where $\mathcal{Z}(s) = \mathbb{E}_{\pi_\theta(a|s)}[e^{T_\phi(s,a)}]$, $\mathcal{I}_{\text{MISA}}$ can be written as:

$$\begin{aligned}
\mathcal{I}_{\text{MISA}} &\triangleq \mathbb{E}_{p(s,a)}\left[\log\frac{\pi_\theta(a|s)}{p(a)}\right] + \mathbb{E}_{p(s,a)}\left[T_\phi(s,a)\right] - \mathbb{E}_{p(s)}\log\mathbb{E}_{\pi_\theta(a|s)}\left[e^{T_\phi(s,a)}\right] \\
&= \mathbb{E}_{p(s,a)}\left[\log\frac{p(a|s)}{p(a)}\right] - \mathbb{E}_{p(s,a)}[\log p(a|s)] \\
&\quad + \mathbb{E}_{p(s,a)}[\log\pi_\theta(a|s)] + \mathbb{E}_{p(s,a)}\left[T_\phi(s,a)\right] - \mathbb{E}_{p(s)}[\log\mathcal{Z}(s)] \\
&= I(S,A) - \mathbb{E}_{p(s)}\left[D_{\text{KL}}(p(a|s)||\mu_{\theta,\phi}(a|s))\right] \leq I(S,A).
\end{aligned} \tag{18}$$

The above inequality holds as the KL divergence is always non-negative.

Similarly, let $\mu_{\theta,\phi}(s,a) \triangleq \frac{1}{\mathcal{Z}}p(s)\pi_\theta(a|s)e^{T_\phi(s,a)}$, where $\mathcal{Z}(s) = \mathbb{E}_{p(s)\pi_\theta(a|s)}[e^{T_\phi(s,a)}]$, $\mathcal{I}_{\text{MISA-DV}}$ can be written as:

$$\begin{aligned}
\mathcal{I}_{\text{MISA-DV}} &\triangleq \mathbb{E}_{p(s,a)}\left[\log\frac{\pi_\theta(a|s)}{p(a)}\right] + \mathbb{E}_{p(s,a)}\left[T_\phi(s,a)\right] - \log\mathbb{E}_{p(s)\pi_\theta(a|s)}\left[e^{T_\phi(s,a)}\right] \\
&= \mathbb{E}_{p(s,a)}\left[\log\frac{p(a|s)}{p(a)}\right] - \mathbb{E}_{p(s,a)}[\log p(a|s)] \\
&\quad + \mathbb{E}_{p(s,a)}[\log\pi_\theta(a|s)] + \mathbb{E}_{p(s,a)}\left[T_\phi(s,a)\right] - \log\mathcal{Z} \\
&= I(S,A) - D_{\text{KL}}(p(s,a)||\mu_{\theta,\phi}(s,a)) \leq I(S,A).
\end{aligned} \tag{19}$$

The above inequality holds as the KL divergence is always non-negative.

Consider the generalized KL-divergence [10, 8] between two un-normalized distributions $\tilde{p}(x)$ and $\tilde{q}(x)$ defined by

$$D_{\text{GKL}}(\tilde{p}(x)||\tilde{q}(x)) = \int \tilde{p}(x)\log\frac{\tilde{p}(x)}{\tilde{q}(x)} - \tilde{p}(x) + \tilde{q}(x)dx, \tag{20}$$

which is always non-negative and reduces to KL divergence when $\tilde{p}$ and $\tilde{q}$ are normalized. Let $\tilde{\mu}_{\theta,\phi}(a|s) \triangleq \pi_\theta(a|s)e^{T_\phi(s,a)-1}$ denote an un-normalized policy. We can rewrite $\mathcal{I}_{\text{MISA-}f}$ as

$$\begin{aligned}
\mathcal{I}_{\text{MISA-}f} &\triangleq \mathbb{E}_{p(s,a)}\left[\log\frac{\pi_\theta(a|s)}{p(a)}\right] + \mathbb{E}_{p(s,a)}\left[T_\phi(s,a)\right] - \mathbb{E}_{p(s)\pi_\theta(a|s)}\left[e^{T_\phi(s,a)-1}\right] \\
&= \mathbb{E}_{p(s,a)}\left[\log\frac{p(a|s)}{p(a)}\right] - \mathbb{E}_{p(s,a)}[\log p(a|s)] \\
&\quad + \mathbb{E}_{p(s,a)}[\log\pi_\theta(a|s)] + \mathbb{E}_{p(s,a)}\left[T_\phi(s,a)-1\right] + 1 - \mathbb{E}_{p(s)\pi_\theta(a|s)}\left[e^{T_\phi(s,a)-1}\right] \\
&= I(S,A) - \mathbb{E}_{p(s)}\left[D_{\text{GKL}}(p(a|s)||\tilde{\mu}_{\theta,\phi}(a|s))\right] \leq I(S,A).
\end{aligned} \tag{21}$$

So far, we have proven that $\mathcal{I}_{\text{MISA}}$, $\mathcal{I}_{\text{MISA-DV}}$ and $\mathcal{I}_{\text{MISA-}f}$ mutual information lower bounds. Then we are going to prove their relations by starting fromt he relation between $\mathcal{I}_{\text{MISA}}$ and $\mathcal{I}_{\text{MISA-DV}}$.

$$\begin{aligned}
\mathcal{I}_{\text{MISA}} - \mathcal{I}_{\text{MISA-DV}} &= D_{\text{KL}}(p(s,a)||\mu_{\theta,\phi}(s,a)) - \mathbb{E}_{p(s)}\left[D_{\text{KL}}(p(a|s)||\mu_{\theta,\phi}(a|s))\right] \\
&= \mathbb{E}_{p(s)}\mathbb{E}_{p(a|s)}\left[\log\frac{p(s,a)}{p(a|s)} - \log\frac{\mu_{\theta,\phi}(s,a)}{\mu_{\theta,\phi}(a|s)}\right] \\
&= \mathbb{E}_{p(s)}\mathbb{E}_{p(a|s)}\left[\log p(s) - \log\frac{1}{\mathcal{Z}}p(s)\mathcal{Z}(s)\right] \\
&= \mathbb{E}_{p(s)}\left[\log p(s) - \log\frac{1}{\mathcal{Z}}p(s)\mathcal{Z}(s)\right] \\
&= D_{\text{KL}}\left(p(s)||\frac{1}{\mathcal{Z}}p(s)\mathcal{Z}(s)\right) \geq 0,
\end{aligned} \tag{22}$$

where $\frac{1}{\mathcal{Z}}p(s)\mathcal{Z}(s)$ is a self-normalized distribution as $\mathcal{Z} = \mathbb{E}_{p(s)[\mathcal{Z}(s)]}$. Therefore, we have $\mathcal{I}_{\text{MISA}} \geq \mathcal{I}_{\text{MISA-DV}}$.

Similarly, the relation between $\mathcal{I}_{\text{MISA-DV}}$ and $\mathcal{I}_{\text{MISA-}f}$ is given by:

$$\begin{aligned}
\mathcal{I}_{\text{MISA-DV}} - \mathcal{I}_{\text{MISA-}f} &= \mathbb{E}_{p(s)}\left[D_{\text{GKL}}(p(a|s)||\tilde{\mu}_{\theta,\phi}(a|s))\right] - D_{\text{KL}}(p(s,a)||\mu_{\theta,\phi}(s,a)) \\
&= \mathbb{E}_{p(s)}\mathbb{E}_{p(a|s)}\left[\log\frac{p(a|s)}{p(s,a)} - \log\frac{\tilde{\mu}_{\theta,\phi}(a|s)}{\mu_{\theta,\phi}(s,a)}\right] - 1 + \mathbb{E}_{p(s)}\mathbb{E}_{\pi_\theta(a|s)}\left[e^{T_\phi(s,a)-1}\right] \\
&= \mathbb{E}_{p(s)}\mathbb{E}_{p(a|s)}\left[-\log p(s) - \log\frac{\tilde{\mu}_{\theta,\phi}(a|s)}{\mu_{\theta,\phi}(s,a)}\right] - 1 + \mathbb{E}_{p(s)}\mathbb{E}_{\pi_\theta(a|s)}\left[e^{T_\phi(s,a)-1}\right] \\
&= \mathbb{E}_{p(s)}\mathbb{E}_{p(a|s)}\left[\log\frac{\mu_{\theta,\phi}(s,a)}{p(s)\tilde{\mu}_{\theta,\phi}(a|s)}\right] - \mathbb{E}_{\mu_{\theta,\phi}(s,a)}[1] + \mathbb{E}_{p(s)}\mathbb{E}_{\pi_\theta(a|s)}\left[e^{T_\phi(s,a)-1}\right] \\
&= \mathbb{E}_{p(s,a)}\left[\log\frac{e}{\mathcal{Z}}\right] - \mathbb{E}_{\mu_{\theta,\phi}(s,a)}[1] + \mathbb{E}_{p(s)}\mathbb{E}_{\pi_\theta(a|s)}\left[e^{T_\phi(s,a)-1}\right] \\
&= \mathbb{E}_{\mu_{\theta,\phi}(s,a)}\left[\log\frac{e}{\mathcal{Z}}\right] - \mathbb{E}_{\mu_{\theta,\phi}(s,a)}[1] + \mathbb{E}_{p(s)}\mathbb{E}_{\pi_\theta(a|s)}\left[e^{T_\phi(s,a)-1}\right] \\
&= \mathbb{E}_{\mu_{\theta,\phi}(s,a)}\left[\log\frac{\mu_{\theta,\phi}(s,a)}{p(s)\tilde{\mu}_{\theta,\phi}(a|s)}\right] - \mathbb{E}_{\mu_{\theta,\phi}(s,a)}[1] + \mathbb{E}_{p(s)}\mathbb{E}_{\pi_\theta(a|s)}\left[e^{T_\phi(s,a)-1}\right] \\
&= D_{\text{GKL}}\left(\mu_{\theta,\phi}(s,a)||p(s)\tilde{\mu}_{\theta,\phi}(a|s)\right) \geq 0,
\end{aligned} \tag{23}$$

where $p(s)\tilde{\mu}_{\theta,\phi}(a|s)$ is an unnormalized joint distribution. Therefore, we have $I(S,A) \geq \mathcal{I}_{\text{MISA}} \geq \mathcal{I}_{\text{MISA-DV}} \geq \mathcal{I}_{\text{MISA-}f}$.

Algorithm 1 Mutual Information Regularized Offline RL

Input: Initialize Q network Q_ϕ, policy network π_θ, dataset $\mathcal{D}$, hyperparameters α_1 and α_2.
for $t \in \{1, \ldots, \texttt{MAX\_STEP}\}$ **do**
  Train the Q network by gradient descent with objective $J_Q(\phi)$ in Eqn. 12:
  $\phi := \phi - \eta_Q \nabla_\phi J_Q(\phi)$
  Improve policy network by gradient ascent with object $J_\pi(\theta)$ in Eqn. 13:
  $\theta := \theta + \eta_\pi \nabla_\theta \mathbb{E}_{s\sim\mathcal{D}, a\sim\pi_\theta(a|s)}[Q_\phi(s,a)] + \alpha_2 \nabla_\theta I_{\text{MISA}}$
end
Output: The well-trained π_θ.

A.2 Derivation of MISA Gradients

We detail how the unbiased gradient is derived in Sec.4.3.

$$\frac{\partial \mathcal{I}_{\text{MISA}}}{\partial \theta} = \mathbb{E}_{s,a\sim D}\left[\frac{\log \pi_\theta(a \mid s)}{\partial \theta}\right] - \mathbb{E}_{s\sim D}\left[\frac{\partial \log \mathbb{E}_{\pi_\theta(a|s)}[e^{Q_\phi(s,a)}]}{\partial \theta}\right]$$

$$= \mathbb{E}_{s,a\sim D}\left[\frac{\log \pi_\theta(a \mid s)}{\partial \theta}\right] - \mathbb{E}_{s\sim \mathcal{D}}\left[\mathbb{E}_{\pi_\theta(a|s)}\left[\frac{e^{Q_\phi(s,a)}}{\mathbb{E}_{\pi_\theta(a|s)}\left[e^{Q_\phi(s,a)}\right]}\frac{\log \pi_\theta(a \mid s)}{\partial \theta}\right]\right] \tag{24}$$

$$= \mathbb{E}_{s,a\sim D}\left[\frac{\log \pi_\theta(a \mid s)}{\partial \theta}\right] - \mathbb{E}_{s\sim D, a\sim p_{\theta,\phi}(a|s)}\left[\frac{\log \pi_\theta(a \mid s)}{\partial \theta}\right] \tag{25}$$

for Eqn. 24, we use the log-derivative trick.

B Implementation Details

We follow the network architectures of CQL [23] and IQL [22], where a neural network of 3 encoding layers of size 256 is used for antmaze-v0 environments, and 2 encoding layers for other tasks, followed by an output layer. We use ELU activation function [11] and SAC [17] as the base RL algorithm. Besides, we use a learning rate of 1×10^{-4} for both the policy network and Q-value network with a cosine learning rate scheduler. When approximating $\mathbb{E}_{\pi_\theta(a|s)}\left[e^{T_\psi(s,a)}\right]$, we use 50 Monte-Carlo samples. To sample from the non-parametric distribution $p_{\theta,\phi}(a \mid s) = \frac{\pi_\theta(a|s)e^{Q_\phi(s,a)}}{\mathbb{E}_{\pi_\theta(a|s)}\left[e^{Q_\phi(s,a)}\right]}$, we use Hamiltonian Monte Carlo algorithm. In addition, for unbiased gradient estimation with MCMC samples, we use a burn-in steps of 5. For all tasks, we average the mean returns over 10 evaluation trajectories and 5 random seeds. In particular, following [22], we evaluate the antmaze-v0 environments for 100 episodes instead. To stabilize the training of our agents in antmaze-v0 environments, we follow [23] and normalize the reward by $r' = (r - 0.5) * 4$. As MCMC sampling is slow, we trade-off its accuracy with efficiency by choosing moderately small iteration configurations. We set the MCMC burn-in steps to 5, number of leapfrog steps to 2, and MCMC step size to 1.

For practical implementations, we follow the CQL-Lagrange [23] implementation by constraining the Q-value update by a "budget" variable τ and rewrite Eqn. 12 as

$$\min_{Q}\max_{\gamma_1 \geq 0} \gamma_1 (\mathbb{E}_{s\sim\mathcal{D}}\left[\log \mathbb{E}_{\pi_\theta(a|s)}\left[e^{Q_\phi(s,a)}\right]\right] - \mathbb{E}_{s,a\sim\mathcal{D}}\left[Q_\phi(s,a)\right] - \tau) - J_Q^{\mathcal{B}}(\phi). \tag{26}$$

Eqn. 26 implies that if the expected value of Q-value difference is less than the threshold τ, γ_1 will adjust to close to 0; if the Q-value difference is higher than the threshold τ, γ_1 will be larger and penalize Q-values harder. We set $\tau = 10$ for antmaze-v0 environments and $\tau = 3$ for adroit-v0 and kitchen-v0 environments. For gym-locomotion-v2 tasks, we disable this function and direction optimize Eqn. 12, because these tasks have a relatively short horizon and dense reward, and further constraining the Q values is less necessary. Our code is implemented in JAX [7] with Flax [19]. All experiments are conducted on NVIDIA 3090 GPUs.