A APPENDIX

A.1 INFORMATION BASED MEASURE DERIVATION

Given that the choice of p_z for sampling z_is is $p(z) \sim Uniform(1, n_z)$ a catergorical uniform distribution where the cardiality of $|Z| = n_z$ we can write the mutual infromation based measure as follows.

$$\begin{aligned}
I(s;z) &= I(z_i;s_j) \\
I(z;s) &= \textstyle\sum_{s_j,z_i} p(s_j,z_i) \log \frac{p(s_j,z_i)}{p(z_i)p(s_j)} \\
I(z;s) &= \textstyle\sum_{s_j,z_i} p(s_j,z_i) \log \frac{p(z_j|s_i)}{p(z_i)} \\
I(z;s) &= \textstyle\sum_{s_j,z_i} p(s_j|z_i).p(z_i) \log \frac{p(z_i|s_j)}{p(z_i)} \\
I(z;s) &= \textstyle\sum_{s_j,z_i} p(s_j|z_i).n_z^{-1} \log \frac{p(z_i|s_j)}{n_z^{-1}} \\
I(z;s) &= \textstyle\sum_{z_i} n_z^{-1}. \sum_{s_j,z_i} d_{\pi_i}(s) \log \frac{d_{\pi_i}(s)}{\sum_k d_{\pi_k(s)}.n_z^{-1}} \\
I(z;s) &= \textstyle\sum_{z_i} n_z^{-1}.KL(d_{\pi_{z_i}}(.)||d_{\pi_{z_{all}}}(.)) \\
I(z;s) &= E_{z_i\sim p(z)}[KL(d_{\pi_{z_i}}(.)||d_{\pi_{z_{all}}}(.))]
\end{aligned} \tag{13}$$

Here we can write $p(s_j|z_i) = d_{\pi_i}(s_j)$ as it is the steady state probability induced by the policy corresponding to a certain $z_i \in Z$. Furthermore we replace If we replace $p(z_i|s_j)$ in 13 with the following.

$$\begin{aligned}
p(z_i|s_j) &= \tfrac{p(s_j|z_i).p(z_i)}{\sum_k p(s_j|k).p(k)} \\
p(z_i|s_j) &= \tfrac{d_{\pi_i}.p(z_i)}{\sum_k d_{\pi_k}(s_j).p(k)} \\
p(z_i|s_j) &= \tfrac{d_{\pi_i}(s_j).n_z^{-1}}{\sum_k d_{\pi_k(s_j)}.n_z^{-1}} \\
p(z_i|s_j) &= \tfrac{d_{\pi_i}(s_j)}{\sum_k d_{\pi_k(s_j)}}
\end{aligned} \tag{14}$$

Since KL divergence is a convex function we can see that the information measure for the ith policy is convex on d_{π_i} induced by that policy π_i.

A.2 INFORMATION BASED MEASURE'S DEVIATION FROM IDEAL MEASURE

$$\begin{aligned}
\delta^{\mathsf{info}} &= E_{z_i\sim p(z)}[KL(d_{\pi_{z_i}}(.)||d_{\pi_{z_i^{-1}}}(.))] - E_{z_i\sim p(z)}[KL(d_{\pi_{z_i}}(.)||d_{\pi_{z_i^{all}}}(.))] \\
\delta^{\mathsf{info}} &= E_{z_i\sim p(z)}[\textstyle\sum_{s_j} d_{\pi_i(s)} \log \frac{d_{\pi_{z_i}}(s)}{d_{\pi_{z_i^{-1}}}(s)})] - E_{z_i\sim p(z)}[\sum_{s_j} d_{\pi_i(s)} \log \frac{d_{\pi_{z_i}}(s)}{d_{\pi_{z_i^{all}}}(s)})] \\
\delta^{\mathsf{info}} &= E_{z_i\sim p(z)}[\textstyle\sum_{s_j} d_{\pi_i(s_j)} (\log \frac{d_{\pi_{z_i}}(s_j)}{\sum_{k\neq i} d_{\pi_k}(s).(n-1)^{-1})} - \log \frac{d_{\pi_{z_i}}(s_j)}{\sum_k d_{\pi_k}(s_j).n^{-1}})] \\
\delta^{\mathsf{info}} &= E_{z_i\sim p(z)}[\textstyle\sum_{s_j} d_{\pi_i}(s_j) \log(\frac{\sum_k d_{\pi_k}(s).n^{-1}}{\sum_{k\neq i} d_{\pi_k}(s).(n-1)^{-1})})]
\end{aligned} \tag{15}$$

A.3 PROPOSED METHOD'S DEVIATION FROM IDEAL MEASURE

$$\begin{aligned}
\delta^{\mathsf{ours}} &= E_{z_i\sim p(z)}[KL(d_{\pi_{z_i}}(.)||d_{\pi_{z_i^{-1}}}(.))] - E_{z_i\sim p(z)}[KL(\mathsf{Uniform}(.)||d_{\pi_{z_i^{-1}}}(.))] \\
\delta^{\mathsf{info}} &= E_{z_i\sim p(z)}[\textstyle\sum_{s_j} d_{\pi_i(s)} \log \frac{d_{\pi_{z_i}}(s)}{d_{\pi_{z_i^{-1}}}(s)})] - E_{z_i\sim p(z)}[\sum_{s_j} d_{\pi_i(s)} \log \frac{n_s^{-1}}{d_{\pi_{z_i^{-1}}}(s)})] \\
\delta^{\mathsf{info}} &= E_{z_i\sim p(z)}[\textstyle\sum_{s_j} d_{\pi_i(s_j)} (\log \frac{d_{\pi_{z_i}}(s_j)}{\sum_{k\neq i} d_{\pi_k}(s).(n-1)^{-1})} - \log \frac{n_s^{-1}}{\sum_{k\neq i} d_{\pi_k}(s).(n-1)^{-1})})] \\
\delta^{\mathsf{info}} &= E_{z_i\sim p(z)}[\textstyle\sum_{s_j} d_{\pi_i(s_j)} (\log(d_{\pi_{z_i}}(s_j)) + \log(n_s))] \\
\delta^{\mathsf{info}} &= E_{z_i\sim p(z)}[\textstyle\sum_{s_j} d_{\pi_i(s_j)} \log(d_{\pi_{z_i}}(s_j))] + K \\
\delta^{\mathsf{info}} &= E_{z_i\sim p(z)}[-H(d_{\pi_{z_i}}(.))] + K
\end{aligned} \tag{16}$$

A.4 COMPARISON OF OUR DEVIATION VS INFORMATION BASED MEASURE'S DEVIATION

Since log is a monotonic function we can lower bound the information based measure's deviation as follows

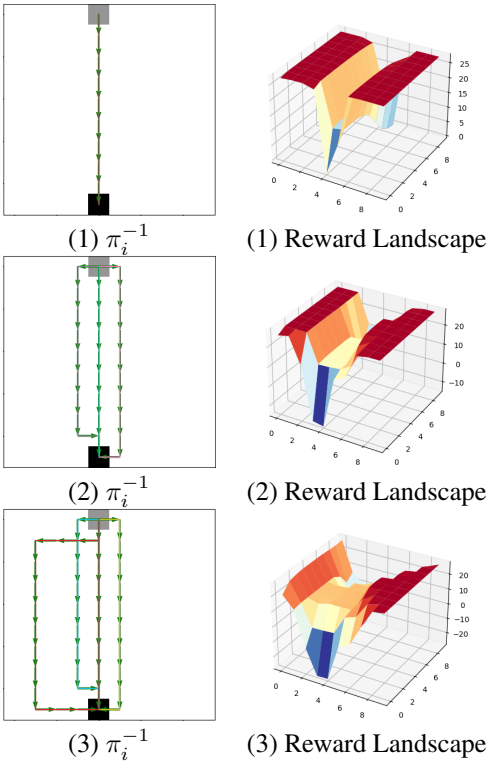
$$
\begin{aligned}
\delta^{\text{info}} = \quad & E_{z_i \sim p(z)}[\textstyle\sum_{s_j}(d_{\pi_i}(s_j)\log(\frac{\sum_{k\neq i} d_{\pi_k}(s).(n-1)_z^{-1})}{\sum_k d_{\pi_k}(s).n_z^{-1}}))] \\
\delta^{\text{info}} \leq \quad & E_{z_i \sim p(z)}[\textstyle\sum_{s_j}(d_{\pi_i}(s_j)\log(\frac{\sum_{k\neq i} d_{\pi_k}(s).(n-1)_z^{-1})}{d_{\pi_i}(s).n_z^{-1}}))[\\
\delta^{\text{info}} \leq \quad & E_{z_i \sim p(z)}[\textstyle\sum_{s_j}(d_{\pi_i}(s_j)(\log(\frac{\sum_{k\neq i} d_{\pi_k}(s).(n-1)^{-1}}{d_{\pi_i}(s)}))] + \log(n_z) \\
\delta^{\text{info}} \leq \quad & E_{z_i \sim p(z)}[-KL(d_{\pi_{z_i}}(.)||d_{\pi_{z_i^{-1}}}(.))] + \log(n_z)] \\
\delta^{\text{info}} \geq \quad & E_{z_i \sim p(z)}[KL(d_{\pi_{z_i}}(.)||d_{\pi_{z_i^{-1}}}(.))] - C \\
\delta^{\text{info}} \geq \quad & E_{z_i \sim p(z)}[H(d_{\pi_{z_i}}(.), d_{\pi_{z_i^{-1}}}(.)) - H(d_{\pi_{z_i}}(.))] - C
\end{aligned}
\tag{17}
$$

If we compared both δ^{info} and δ^{ours}

$$
\begin{aligned}
\delta^{\text{info}} - \delta^{\text{ours}} = \quad & H(d_{\pi_{z_i}}(.), d_{\pi_{z_i^{-1}}}(.)) - H(d_{\pi_{z_i}}(.)) + H(d_{\pi_{z_i}}(.)) + M \\
\delta^{\text{info}} - \delta^{\text{ours}} = \quad & H(d_{\pi_{z_i}}(.), d_{\pi_{z_i^{-1}}}(.)) + M
\end{aligned}
\tag{18}
$$

For a discrete case the cross entropy is a non negative quantity. Thus always $\delta_{\text{ours}} \leq \delta_{\text{info}}$ making our measure close to the ideal measure than the information based measure.

A.5 Visualization of the diversity measure landscape for the grid world problem



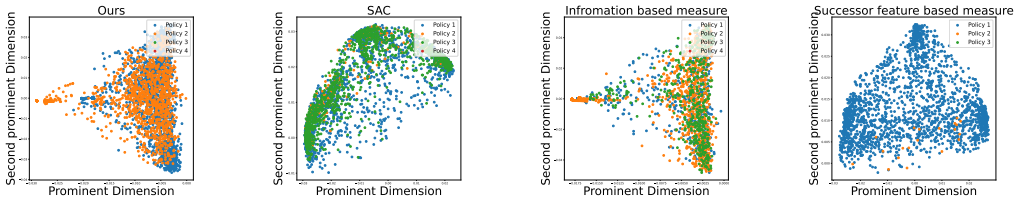
(1) π_i^{-1} (1) Reward Landscape

(2) π_i^{-1} (2) Reward Landscape

(3) π_i^{-1} (3) Reward Landscape

Figure 5: This figure illustrates learned diversity measure's landscape corresponding policy of others in different occasion in the gridworld setting. For the figures in the right the x, y axis denote the grid and the z axis denotes the value of the reward function. This figure illustrates on how visted states are penalized by a function independent of the policy π_i under the proposed diversity measure.

A.6 COVERAGE: CONTINUAL CONTROL TASK



(a) Ours (b) SAC (c) Information based (d) SF

Figure 6: This figure illustrates the state space coverage by the policies generated by each agent in the ant environment. Dimensionality reduction was performed through PCA on the visited states. Here the x and y axes represents the two primary dimensions resulting from PCA. Our proposed method demonstrated better coverage in the areas on the right hand side of the space a compared to the lack of coverage in case of SAC. All three of the polices showing better performance than the standard SAC on test task while having certain level of coverage in this area leads us to hypothesize that the diverse polices are generated by the visitation of these states. Our proposed measure with a higher coverage was able perform better in average the than other methods both enhancing our hypothesis and the ability of our measure to induce better diversity.