A POLICY INVARIANCE UNDER INTERMEDIATE REWARDS

Assume the original Markov Decision Process (MDP) without the intermediate rewards is defined as $M = (S, A, T, \gamma, R)$, where S and A are state and action spaces, T is the state transition probabilities, γ is the discount factor, and R is the rewards. When we introduce the intermediate rewards R_m, the MDP is modified to $M' = (S, A, T, \gamma, R')$, where $R' = R + R_m$. The following theory provides a sufficient and necessary condition for the modified MDP M' to achieve the same optimal policy as the original MDP M.

Theorem 1. *The modified MDP $M' = (S, A, T, \gamma, R + F)$ with any shaping reward function F is guaranteed to be consistent with the optimal policy of the original MDP $M = (S, A, T, \gamma, R)$ if the shaping function F have the following form*

$$F(s, a, s') = \gamma\Phi(s') - \Phi(s), \tag{A.1}$$

where $\Phi : S \to \mathbb{R}$ is a potential function evaluated on states. For infinite-state case (i.e., the state space is an infinite set) the potential function is additionally required to be bounded.

Proof. Please refer to Ng et al. (1999) for detailed proof. □

From the above theorem, we can see our intermediate rewards in equation 3 is a potential based shaping function and the potential function is $\Phi(s) = -H(y \mid s)$. For classification task where $y \in \mathcal{Y} = \{1, 2, \ldots, K\}$ is a discrete variable, the entropy is naturally bounded, i.e., $0 \leq H(y \mid s) \leq log|\mathcal{Y}|$, where $|\mathcal{Y}|$ is the cardinality of the label space. For regression task where $y \in \mathbb{R}$, the entropy is bounded by $0 \leq H(y \mid s) \leq H(y \mid \emptyset)$. The upper bound $H(y \mid \emptyset)$ is determined by the given surrogate model. Similarly, for the intermediate rewards in equation 10, the potential function is $\Phi(s) = \frac{\log p(x_u|s)}{|u|}$, and it is bounded for a given surrogate model.

B EXPERIMENTS

B.1 CLASSIFICATION

For classification task, we conduct experiments on MNIST and two UCI datasets. We downsample the MNIST images to 16×16 to reduce the total number of features. Features are normalized into the range $[0, 1]$.

The surrogate model for classification task is a conditional extension of ACFlow, where the arbitrary conditional distributions are conditioned on the target variable y. The ACFlow model for MNIST is similar to the model they used in their original work (Li et al., 2019), which contains a stack of conditional coupling transformations and a conditional Gaussian likelihood module. We additionally condition them on the one-hot encoding of the target y. For UCI dataset, we use an autoregressive likelihood module. To train the surrogate model, we randomly select two non-overlapping subsets u and o and optimize the arbitrary conditional log likelihood

$$\begin{aligned} \log p(y, x_u \mid x_o) &= \log p(x_u \mid x_o) + \log P(y \mid x_u, x_o) \\ &= \log p(x_u \mid x_o) + \log \frac{p(x_u, x_o \mid y)P(y)}{\sum_{y'} p(x_u, x_o \mid y')P(y')}. \end{aligned} \tag{B.2}$$

The agent is implemented as a PPO policy. Given the current state x_o and the auxiliary information from the surrogate model, we extract a set embedding using set transformer (Lee et al., 2019). The inputs are first transformed to sets by concatenating with the one-hot encoding of their indexes. The set embedding is beneficial to deal with arbitrary dimensionality of the inputs. The policy network then takes the set embedding as inputs and outputs the next action. The critic network takes the same set embedding as inputs and output an estimate of the state values. To help the agent extract meaningful representations from its inputs, we let the prediction model f_θ take the same set embedding as input. The policy network, the critic network and the prediction function are all implemented as fully connected layers.

We run the baseline model JAFA (Shim et al., 2018) using their public code. We cross-validate the optimal architecture by modifying the number of layers and the size of each layer for both the agent and the classifier.

We adapt EDDI (Ma et al., 2018) to perform classification task by modifying the decoder to output Categorical distribution for y and Gaussian distribution for x. EDDI learns the distribution $p(y, x_o)$ by utilizing a VAE based model. The acquisition metric for EDDI is

$$\mathcal{U}_i = \mathbb{E}_{x_i \sim p(x_i|x_o)} D_{\mathrm{KL}}[p(z \mid x_i, x_o) \| p(z|x_o)] - \mathbb{E}_{y,x_i \sim p(y,x_i|x_o)} D_{\mathrm{KL}}[p(z \mid y, x_i, x_o) \| p(z \mid y, x_o)], \tag{B.3}$$

which is estimated using the proposal distribution. Then, a greedy policy that acquires the feature with maximum utility is employed. We similarly cross-validate the architecture for each dataset.

We also compare to a greedy policy using the surrogate model where the utility is calculated by equation 8. At each acquisition step, the one with maximum utility is selected.

B.2 Regression

For regression task, the target variable y is concatenated into the features x and the surrogate model learns the distribution $p(y, x_u \mid x_o)$ using the ACFlow. The agent is similarly implemented as the PPO policy with a set transformer based feature extractor. Baseline models include JAFA and EDDI, where the architecture is selected by cross validation. We also build a greedy policy using our surrogate model by estimating the utility following equation 6. For GSMRL and JAFA, the reward for a prediction $\hat{y}$ is calculated as the negative MSE $-\|\hat{y} - y\|_2^2$.

B.3 Time Series

Acquiring features for time series data requires the agent to integrate chronological constraints into the action space. For RL based approach, we manually set the probabilities of invalid action to zeros. For greedy approach, inspired by Thompson sampling (Thompson, 1933; Russo et al., 2017), we employ a prior distribution to encode our chronological constraint. Specifically, we set the prior as a Dirichlet distribution that is biased towards the selection of earlier time steps:

$$\pi(\rho) = \mathrm{Dir}\left[\alpha(T - (\max(o) + 1)), \ldots, \alpha(T - (T - 1))\right](\rho), \tag{B.4}$$

where α is a hyperparameter, T is the total time steps, $\max(o)$ represents the latest time step already acquired, and ρ is a distribution for acquisition over the remaining future time steps. However, we still desire that the acquired features are informative for target y. Hence, we update the prior to a posterior using time steps V that are drawn according to how informative they are:

$$p(V_n = t) \propto \exp(I(x_t; y \mid x_o)),\ t \in \{\max(o) + 1, \ldots, T - 1\},\ n \in \{1, \ldots, N\}, \tag{B.5}$$

where N is the number of samples. Due to conjugacy, the posterior is also a Dirichlet distribution

$$p(\rho \mid V) = \mathrm{Dir}\left[\alpha(T - (\max(o) + 1)) + \sum_{n=1}^{N} \mathbb{I}\{V_n = \max(o) + 1\}, \ldots\right](\rho). \tag{B.6}$$

Samples from posterior represent the probabilities of choosing each candidate, which now prefer both earlier time steps and informative features. We draw a sample from posterior and select the most likely time step at each acquisition step.

B.4 Unsupervised

To perform active feature acquisition on unsupervised tasks, a.k.a, active instance recognition, we modify the reward for prediction as the negative MSE of the unobserved features, i.e., $-\|\hat{x}_u - x_u\|_2^2$, where $\hat{x}_u$ is the imputed values of the unobserved features. The surrogate model is again an ACFlow model and the agent is similarly implemented as a PPO policy.

The JAFA is adapted to this task by changing the classifier to an auto-encoder like model, where the observed features x_o are encoded to predict the unobserved features x_u.

For EDDI, by plugging $y = x$ into equation B.3, we have the acquisition metric for this setting as

$$\mathcal{U}_i = \mathbb{E}_{x_i \sim p(x_i|x_o)} D_{\mathrm{KL}}[p(z \mid x_i, x_o) \| p(z|x_o)], \tag{B.7}$$

since the second KL term in equation B.3 equals to zero.

To build a greedy policy using our surrogate model, we estimate the utility using equation 9. Monte Carlo estimation is utilized to estimate the entropy.

Table 1: Hyperparameters for GSMRL and baselines.

GSMRL	set transformer	$\{32, 64\} \times \{1, 2\}$
	set embedding size	$\{32, 64\}$
	policy network	$\{32, 64\} \times \{2, 3\}$
	critic network	$\{32, 64\} \times \{2, 3\}$
	prediction network	$\{64, 128\} \times \{2, 3\}$
	advantage λ	0.95
	discount factor γ	0.99
	PPO clip range	$[0.8, 1.2]$
	entropy coefficient	0.0
JAFA	set embedding size	$\{16, 32, 64, 128\}$
	Q network	$\{16, 32, 64, 128\} \times \{2, 3, 4, 5\}$
	prediction network	$\{16, 32, 64, 128\} \times \{2, 3, 4, 5\}$
EDDI	set embedding size	$\{10, 20, 50, 100\}$
	encoder	$\{32, 64, 128, 256\} \times \{3, 4, 5, 6\}$
	latent code	$\{10, 20, 50, 100\}$
	decoder	$\{32, 64, 128, 256\} \times \{3, 4, 5, 6\}$

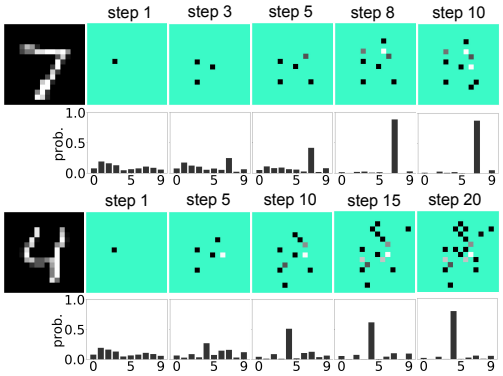


Figure D.1: Examples of the acquisition process for AFA task from GSMRL.

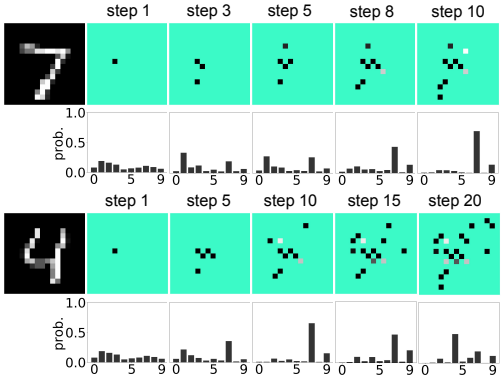


Figure D.2: Examples of the acquisition process for AFA task from GSM+Greedy.

C HYPERPARAMETERS

We search the hyperparameters for both our GSMRL and baselines using cross-validation. The range of the hyperparameters is listed in Table 1.

D ADDITIONAL RESULTS

Due to the space limit, we only show one example for the acquisition process in the main text. Figure D.1 and D.3 show some additional examples for AFA and AIR tasks respectively. In Fig. D.2 and D.4, we present several examples of the acquisition process from the greedy policy. Note that the predictions for both the greedy and the non-greedy policy are from the same pretrained ACFlow model, therefore the only difference is the acquired features. Comparing the greedy and the non-greedy policy suggests that the non-greedy policy eliminates the prediction uncertainty much faster than the greedy one.

In Fig. 4 and 10, we present the acquired features from our GSMRL for several testing examples. To better understand the overall distribution of the acquired features across all the testing instances, we plot the frequencies of each feature being acquired in Fig. D.5 and D.6 for both AFA and AIR on MNIST respectively. A higher value of the frequency means the corresponding feature is acquired for more testing instances. Specifically, the frequency for a feature equals to one means the corresponding feature is a common feature acquired for all testing instances. The frequency loosely represents the importance of each feature, which could help with model interpretation and reasoning about decision making. We will explore this direction in future works.

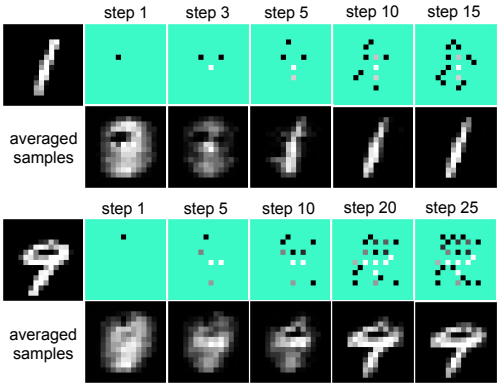


Figure D.3: Examples of the acquisition process for AIR task from GSMRL.

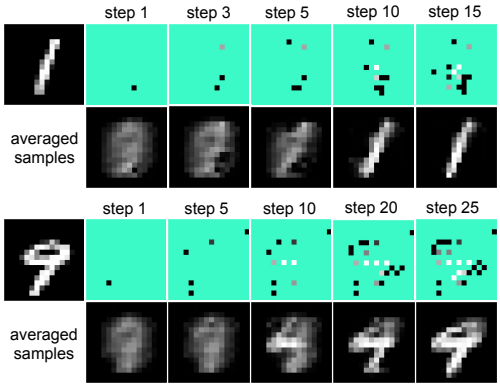


Figure D.4: Examples of the acquisition process for AIR task form GSM+Greedy.

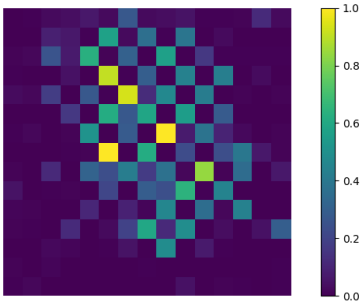


Figure D.5: Acquisition frequency for AFA.

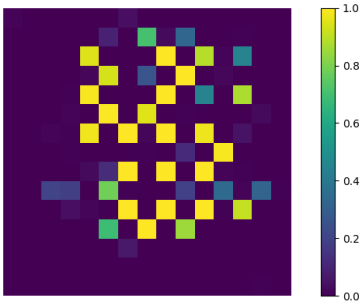
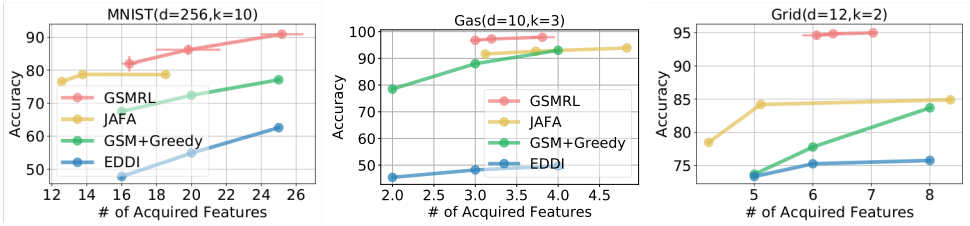
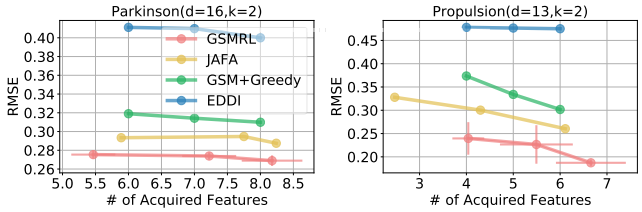


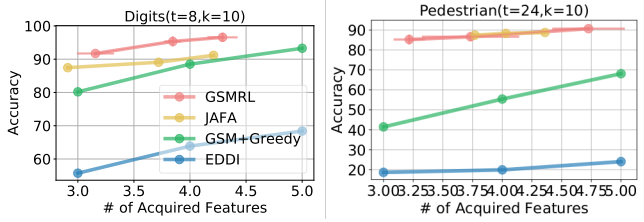
Figure D.6: Acquisition frequency for AIR.



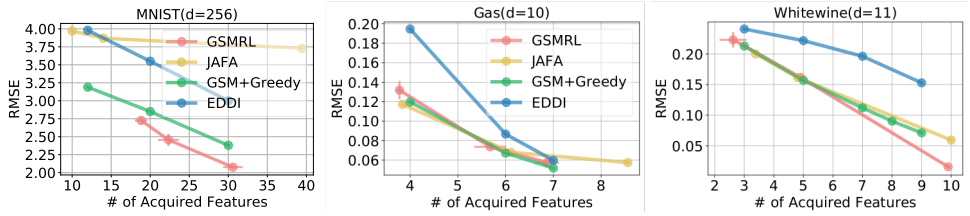
(a) Classification



(b) Regression



(c) Time Series



(d) Unsupervised

Figure D.7: Sensitivity analysis by running multiple times independently. Mean and standard deviation are reported for both the number of acquisitions and task performance.

In Fig. D.7, we analyse the sensitivity of our model to random initialization by running our model three times independently with different random seeds. We report the mean and standard deviation for both the number of acquisitions and the task performance. Baseline performance are presented for reference. We can see that our model is robust to random initialization and performs consistently better than baselines.