**Algorithm 1** Mini-batch PolicyWalk for Multiple Dynamics

**INPUT:** Environments $\{E_m\}_{m=1}^M$, Demonstrations $\{D_m\}_{m=1}^M$, Prior $P(R)$, Step size $\delta$, Mini-batch size $N$
**OUTPUT:** Sampled Rewards $\{R_i\}_{i=1}^t$

Pick a random vector $R \in \mathbb{R}^{|S|}/\delta$
$\{\pi_m\}_{m=1}^M \leftarrow \{\text{Policy iteration}(E_m, R)\}_{m=1}^M$
**for** $i = 1$ **do** $t$
  Pick a reward vector $\tilde{R}$ uniformly at random from the neighbors of $R \in \mathbb{R}^{|S|}/\delta$
  $u \leftarrow$ Sample from uniform distribution $U(0, 1)$
  $\tilde{\mathbb{N}} \leftarrow$ Sampled $N$ integers from $\{n \in \mathbb{N} | \ n \leq M\}$ without repetition
  Compute $Q^\pi(s, a, R, E) \quad \forall\{s, a, (E_n, \pi_n)\} \in \mathcal{S} \times \mathcal{A} \times \{(E_n, \pi_n)\}_{n\in\tilde{\mathbb{N}}}$
  **if** $\exists\{s, a, (E, \pi)\} \in \mathcal{S} \times \mathcal{A} \times \{(E_n, \pi_n)\}_{n\in\tilde{\mathbb{N}}}, Q^\pi(s, \pi(s), \tilde{R}, E) < Q^\pi(s, a, \tilde{R}, E)$ **then** ▷ If any sampled policy is not optimal
    $\{\tilde{\pi}_n\}_{n\in\tilde{\mathbb{N}}} \leftarrow \{\text{Policy iteration}(E_n, \tilde{R})\}_{n\in\tilde{\mathbb{N}}}$
    **if** $\frac{1}{M}\log\left(u\frac{P(R)}{P(\tilde{R})}\right) < \frac{1}{N}\sum_{n\in\tilde{\mathbb{N}}}\log P\left(D_n, E_n|\tilde{R}\right) - \log P\left(D_n, E_n|R\right)$ **then**
      $R \leftarrow \tilde{R}$
      $\{\pi_n\}_{n\in\tilde{\mathbb{N}}} \leftarrow \{\tilde{\pi}_n\}_{n\in\tilde{\mathbb{N}}}$
  **else if** $\frac{1}{M}\log\left(u\frac{P(R)}{P(\tilde{R})}\right) < \frac{1}{N}\sum_{n\in\tilde{\mathbb{N}}}\log P\left(D_n, E_n|\tilde{R}\right) - \log P\left(D_n, E_n|R\right)$ **then**
    $R \leftarrow \tilde{R}$
  $R_i \leftarrow R$

## 5 EXPERIMENTS

The experimental environment is a windy grid world, in which each state has a wind direction, and the agent transitions to a wind direction with a certain probability regardless of the agent's action. Hence, we can create environments with different dynamics by varying the wind direction of each state. In this experiment, the probability of forced transition to the wind direction is set to 30%. The number of wind directions is five (i.e., up, down, left, right, and no wind), and the wind direction of each state is independent. The reward is 1.0 in a upper right-hand corner state (4, 4), and zero elsewhere.

We evaluate the estimated reward with a score that is known as the expected value difference (EVD) [1] which is a measure of how sub-optimal the learned policy is under the expert true reward. To calculate EVD, we used 100 environments with different wind directions (dynamics) generated from uniform distributions, and each experiment was conducted 10 times.

In Figure 1, the number of demonstrated environments is fixed to eight, and the number of environments used in each MCMC step (i.e., mini-batch size) are varied across {1, 2, 4, 8}. As Figure 1 shows, EVD does not change significantly with the mini-batch size $N$. Figure 2 evaluates our method with a mini-batch size of $N = 1$ by varying the number $M$ of environments of expert's demonstrations. EVD decreases as the number of environments for reward estimation decreases.

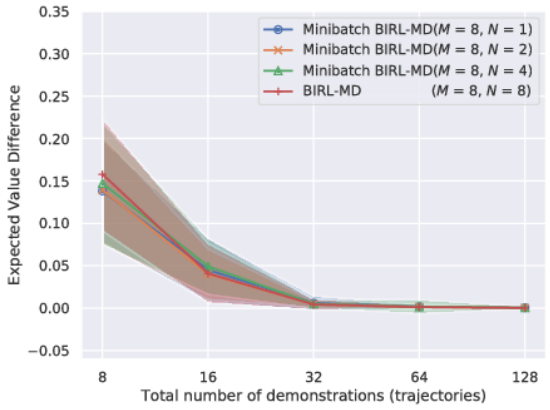


**Figure 1: Evaluation of mini-batch BIRL-MD for different mini-batch size $N$.**

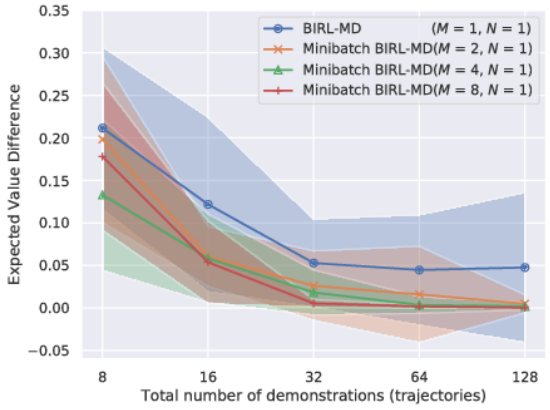


**Figure 2: Evaluation of Minibatch BIRL-MD for Different Number of Environments of Expert's Demonstrations $M$ with fixed mini-batch size $N$.**

## 6 CONCLUSIONS AND FUTURE WORK

This paper formulates a Bayesian inverse reinforcement learning problem for expert demonstrations (i.e., sequences of both sensor inputs to expert and expert's actions) under multiple environments with different dynamics. An advantage of our method is its scalability with respect to the number of environments in which expert demonstrations are generated. Figure 1 and Table **??** show that our method can approximate the posterior distribution with a small mini-batch size in a computational time that is comparable with that of BIRL.

## REFERENCES

[1] Sergey Levine, Zoran Popovic, and Vladlen Koltun. 2011. Nonlinear inverse reinforcement learning with gaussian processes. In *Advances in Neural Information Processing Systems*. 19–27.
[2] Deepak Ramachandran and Eyal Amir. 2007. Bayesian inverse reinforcement learning. *IJCAI International Joint Conference on Artificial Intelligence* (2007), 2586–2591.