VIME: Variational Information Maximizing Exploration

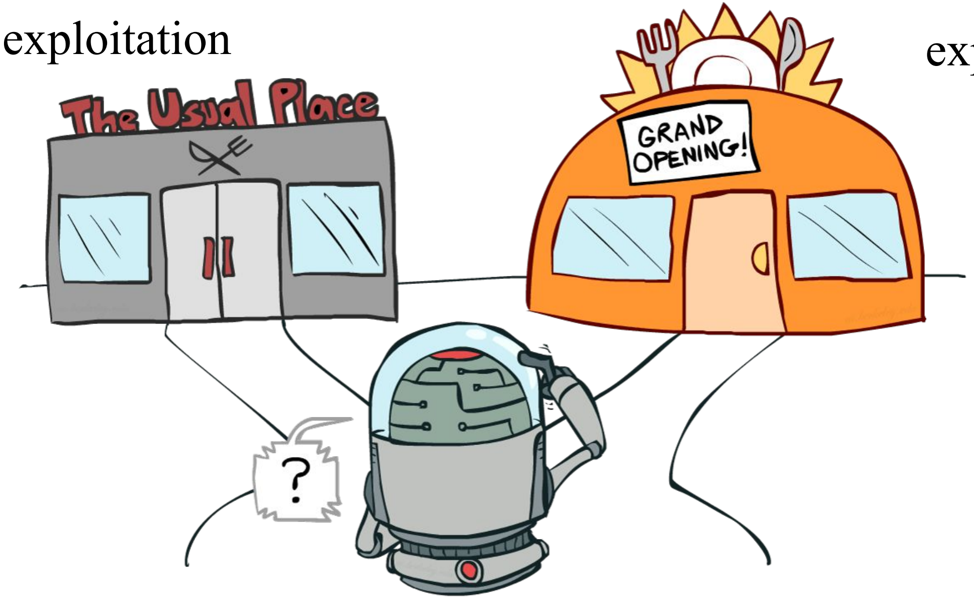
Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, Pieter Abbeel

Presenter: Daniel Flam-Shepherd

Exploration vs Exploitation

exploitation

exploration

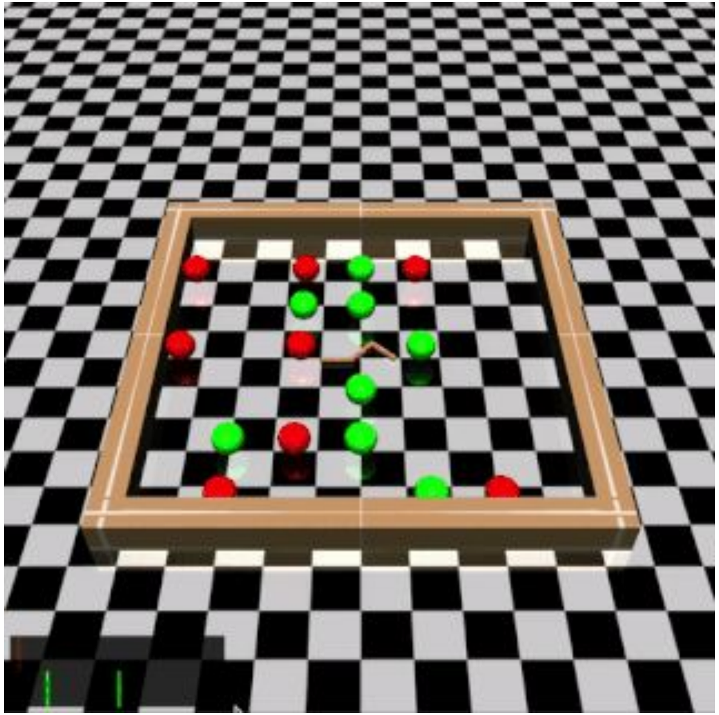


- An effective exploration strategy allows the agent to generate trajectories that are maximally informative about the environment

Image stolen from eric moura

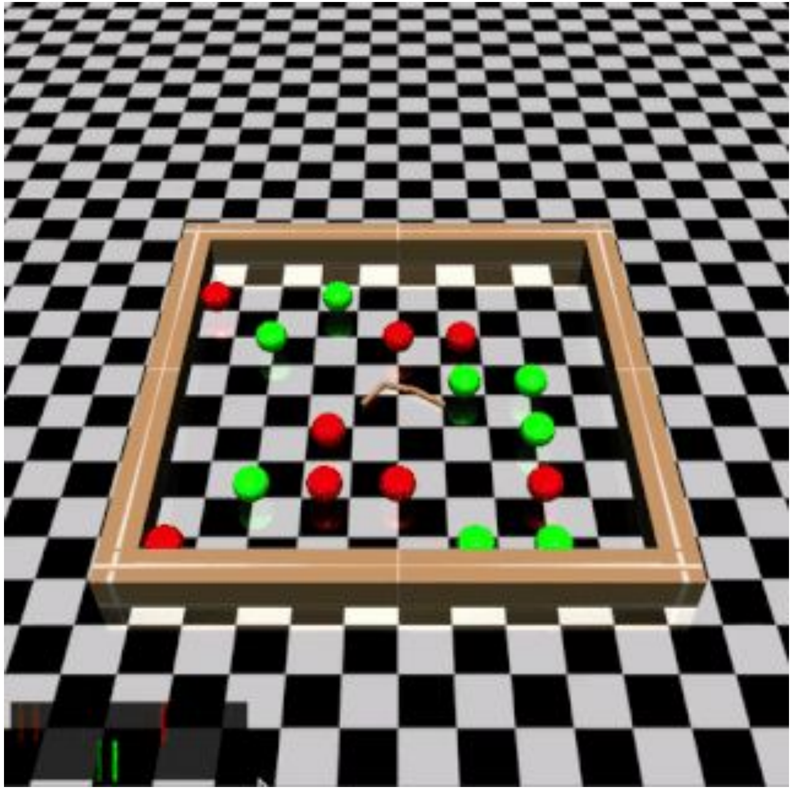
Why do we need Efficient Exploration

- Efficient exploration is an unsolved challenge in Reinforcement Learning
- In naive strategies, agents randomly stumble around until they enter rewarding situation
- Only works in toy tasks -- how do we scale to high dimensional action spaces



VIME is the solution

- A practical approach to exploration using uncertainty from a Bayesian Neural Network
- Improves a range of policy search methods and works in settings with sparse rewards



How is exploration typically handled ?

- For small tasks, we can turn to Bayesian RL and PAC-MDP methods
- The problem ? -- assumes discrete spaces.
- Otherwise -- use heuristic exploration strategies including :
  - acting randomly using epsilon-greedy or Boltzmann exploration, or utilizing Gaussian noise on the controls in policy gradient methods.
- The problem ? Can be highly inefficient.
- A few other methods have been proposed but nothing fully addresses exploration in continuous control.

Contributions

- This paper proposes a curiosity-driven exploration strategy -- agents are encouraged to take actions that result in states they deem surprising
- The authors propose a practical implementation, measuring information gain using variational inference. Specifically, the agent’s current understanding of the environment dynamics is represented by a BNN.
- VIME scales naturally to continuous state and action spaces and achieves significantly better performance than naïve exploration strategies.

General Background -- Preliminaries

Assume a finite-horizon discounted Markov decision process

State set

Transition distribution

Initial state distribution

Horizon

$$(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \rho_0, \gamma, T)$$

Action Set

reward

Discount factor

With expected discounted return

$$\mu(\pi_\alpha) = \mathbb{E}_\tau [\sum_{t=0}^{T} \gamma^t r(s_t, a_t)]$$

For Trajectory $\tau = (s_0, a_0, \ldots),\ s_0 \sim \rho_0(s_0),\ a_t \sim \pi_\alpha(a_t|s_t),\ s_{t+1} \sim \mathcal{P}(s_{t+1}|s_t, a_t)$

2.2 curiosity driven exploration

- The agent engages in systematic exploration by seeking out state-action regions that are relatively unexplored.
- The agent models the environment dynamics via a model, parametrized by the random variable Θ with values $\theta \in \Theta$.

$$p(s_{t+1} | s_t, a_t; \theta)$$

- Assuming a prior $p(\theta)$, it maintains a distribution over dynamic models through a distribution over θ
- The history of the agent up until time step t is denoted as

$$\xi_t = \{s_1, a_1, \ldots, s_t\}$$

- The agent should take actions that maximize the reduction in uncertainty about dynamics

2.2 curiosity driven exploration

How can we achieve this? -- Maximizing the sum of reductions in entropy

$$\sum_t [H(\Theta|\xi_t, a_t) - H(\Theta|S_{t+1}, \xi_t, a_t)] = \sum_t I(S_{t+1}; \Theta|\xi_t, a_t)$$

The agent is encouraged to take actions that lead to states that are maximally informative about the dynamics model

$$I(S_{t+1}; \Theta|\xi_t, a_t) = \mathbb{E}_{s_{t+1}\sim\mathcal{P}(\cdot|\xi_t, a_t)}[D_{\mathrm{KL}}[p(\theta|\xi_t, a_t, s_{t+1})\,||p(\theta|\xi_t)]]$$

The trade-off between exploitation and exploration can now be realized explicitly as follows:

$$r'(s_t, a_T, s_{t+1}) = r(s_t, a_t) + \eta D_{\mathrm{KL}}[p(\theta|\xi_t, a_t, s_{t+1})\,||p(\theta|\xi_t)]$$

However, requires calculating the posterior which is generally intractable. $p(\theta|\xi_t, a_t, s_{t+1})$

Variational Bayes

We can derive the posterior distribution given a new state-action pair through Bayes' rule as

$$p(\theta|\xi_t, a_t, s_{t+1}) = \frac{p(\theta|\xi_t)p(s_{t+1}|\xi_t, a_t; \theta)}{p(s_{t+1}|\xi_t, a_t)} \qquad p(s_{t+1}|\xi_t, a_t) = \int_\Theta p(s_{t+1}|\xi_t, a_t; \theta)p(\theta|\xi_t)d\theta$$

This integral tends to be intractable when using highly expressive parametrized models. Use variational inference, approximate the true posterior with an approximation and minimize $D_{\mathrm{KL}}[q(\theta;\phi) \,||\, p(\theta|\mathcal{D})]$ which is equivalent to maximizing

$$L[q(\theta;\phi), \mathcal{D}] = \mathbb{E}_{\theta\sim q(\theta;\phi)}[p(\mathcal{D}|\theta)] - D_{\mathrm{KL}}[q(\theta;\phi) \,||\, p(\theta)]$$

Use this to compute an approximation to information gain using this lower bound.

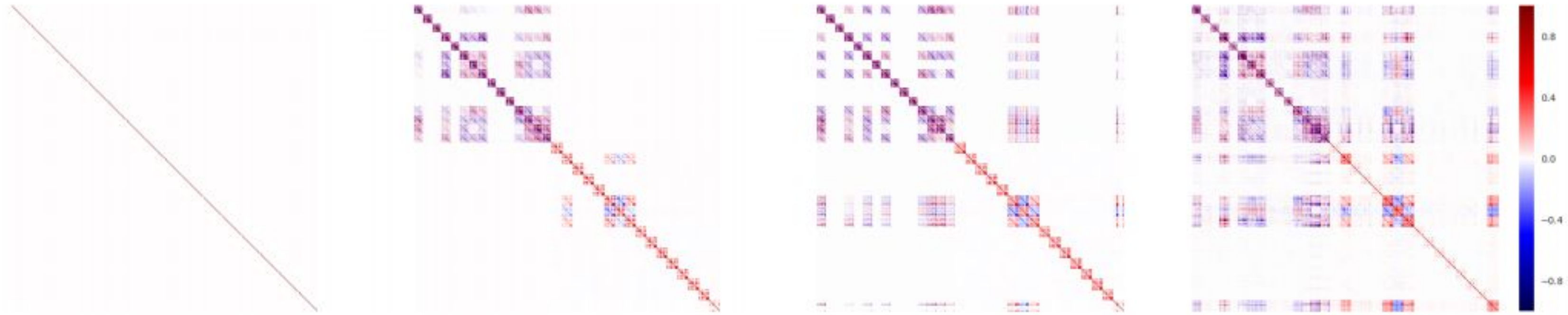
$$r'(s_t, a_T, s_{t+1}) = r(s_t, a_t) + \eta D_{KL}[q(\theta;\phi_{t+1}) \,||q(\theta;\phi_t)]$$

What do we use for $p(s_{t+1}|\xi_t, a_t; \theta)$?

A Bayesian Neural Network! They use a fully factorized Gaussian distribution of the weights

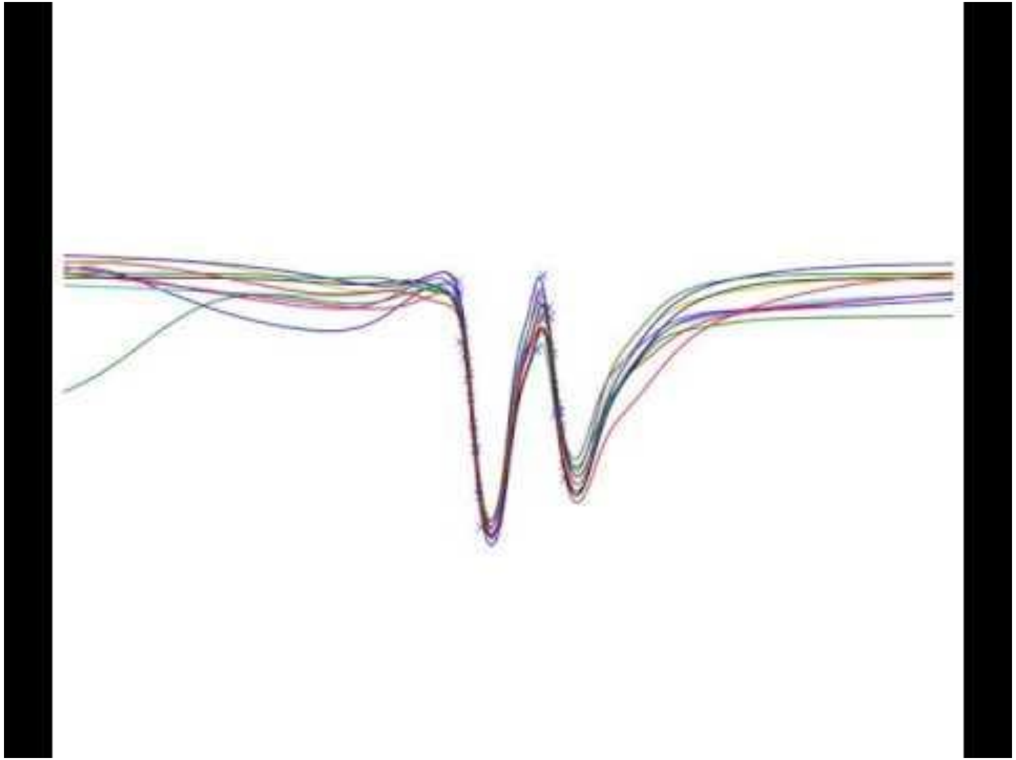
$$q(\theta;\phi) = \prod_{i=1}^{|\Theta|} \mathcal{N}(\theta_t|\mu_i, \sigma_i^2)$$

Ensure the sd is parameterized to be positive and optimize the lower bound using the local reparameterization trick



(a) Fully factorized (b) Matrix-variate (c) Block tridiagonal (d) Full covariance

Fitting some 1D toy data with a BNN!



Implementation : VIME

Algorithm 1: Variational Information Maximizing Exploration (VIME)

for each epoch n **do**
- **for** each timestep t in each trajectory generated during n **do**
  - Generate action $a_t \sim \pi_\alpha(s_t)$ and sample state $s_{t+1} \sim \mathcal{P}(\cdot|\xi_t, a_t)$, get $r(s_t, a_t)$.
  - Add triplet (s_t, a_t, s_{t+1}) to FIFO replay pool $\mathcal{R}$.
  - Compute $D_{\mathrm{KL}}[q(\theta; \phi'_{n+1}) \,\|\, q(\theta; \phi_{n+1})]$ by approximation $\nabla^\top H^{-1} \nabla$, following Eq. (16) for diagonal BNNs, or by optimizing Eq. (12) to obtain ϕ'_{n+1} for general BNNs.
  - Divide $D_{\mathrm{KL}}[q(\theta; \phi'_{n+1}) \,\|\, q(\theta; \phi_{n+1})]$ by median of previous KL divergences.
  - Construct $r'(s_t, a_t, s_{t+1}) \leftarrow r(s_t, a_t) + \eta D_{\mathrm{KL}}[q(\theta; \phi'_{n+1}) \,\|\, q(\theta; \phi_{n+1})]$, following Eq. (7).
- Minimize $D_{\mathrm{KL}}[q(\theta; \phi_n) \,\|\, p(\theta)] - \mathbb{E}_{\theta \sim q(\cdot; \phi_n)}[\log p(\mathcal{D}|\theta)]$ following Eq. (6), with $\mathcal{D}$ sampled randomly from $\mathcal{R}$, leading to updated posterior $q(\theta; \phi_{n+1})$.
- Use rewards $\{r'(s_t, a_t, s_{t+1})\}$ to update policy π_α using any standard RL method.

Implementation

The posterior distribution of the dynamics parameter can be computed through

$$\phi' = \arg\min_{\phi} \Big[\overbrace{\underbrace{D_{\mathrm{KL}}[q(\theta;\phi) \,\|\, q(\theta;\phi_{t-1})]}_{\ell_{\mathrm{KL}}(q(\theta;\phi))} - \mathbb{E}_{\theta\sim q(\cdot;\phi)} \left[\log p(s_t|\xi_t, a_t;\theta)\right]}^{\ell(q(\theta;\phi),s_t)} \Big], \tag{12}$$

The KL divergence term has very simple form

$$D_{\mathrm{KL}}[q(\theta;\phi) \,\|\, q(\theta;\phi')] = \frac{1}{2}\sum_{i=1}^{|\Theta|} \left(\left(\frac{\sigma_i}{\sigma_i'}\right)^2 + 2\log\sigma_i' - 2\log\sigma_i + \frac{(\mu_i'-\mu_i)^2}{\sigma_i'^2} \right) - \frac{|\Theta|}{2}. \tag{14}$$

Can also optimize efficiently using a single higher order step $\Delta\phi = H^{-1}(\ell)\nabla_\phi \ell(q(\theta;\phi), s_t)$

$$D_{\mathrm{KL}}\left[q(\theta;\phi+\lambda\Delta\phi) \,||\, q(\theta;\phi)\right] \approx \frac{1}{2}\lambda^2 \nabla_\phi \ell^\top H^{-1}(\ell_{\mathrm{KL}})\nabla_\phi \ell$$

Experiments

- The authors investigate

  (i) whether VIME improves learning when the reward is well shaped

  (ii) whether VIME can succeed in domains that have extremely sparse rewards,

  (iii) how η, as used in in Eq. (3), trades off exploration and exploitation behavior.

- The following Tasks are part of the experimental setup :

  1) CartPole ($S \subseteq R^4$, $A \subseteq R^1$), 2) CartPoleSwingup ($S \subseteq R^4$, $A \subseteq R^1$),
  3) DoublePendulum ($S \subseteq R^6$, $A \subseteq R^1$), 4) MountainCar ($S \subseteq R^3$, $A \subseteq R^1$),
  5) HalfCheetah ($S \subseteq R^{20}$, $A \subseteq R^6$), 6) Walker2D ($S \subseteq R^{20}$, $A \subseteq R^6$),
  7) SwimmerGather ($S \subseteq R^{33}$, $A \subseteq R^2$

Results : Sparse rewards

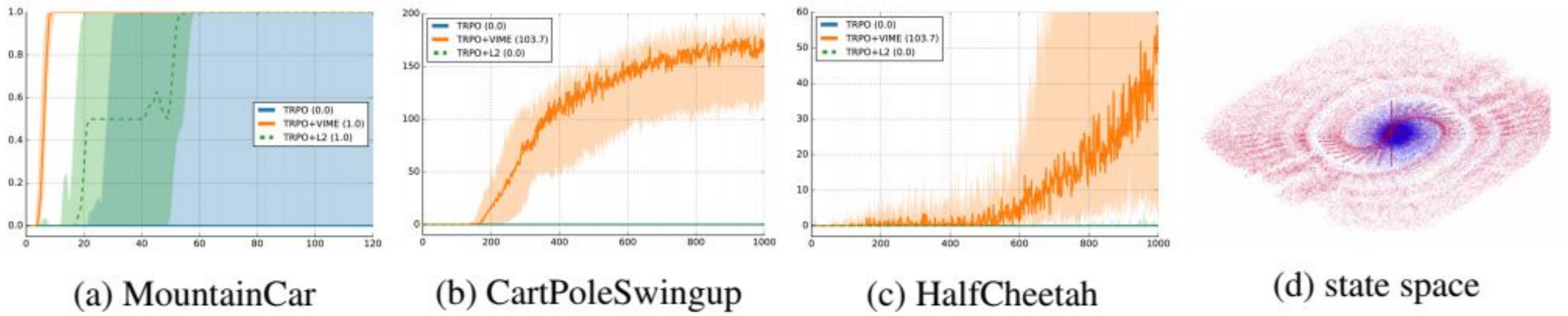


Figure 1: (a,b,c) TRPO+VIME versus TRPO on tasks with sparse rewards; (d) comparison of TRPO+VIME (red) and TRPO (blue) on MountainCar: visited states until convergence

Results : SwimmerGather

- They evaluate VIME on a difficult hierarchical task involving naturally sparse rewards
- VIME leads the agent to acquire coherent motion primitives without any reward guidance, achieving promising results on this challenging task.

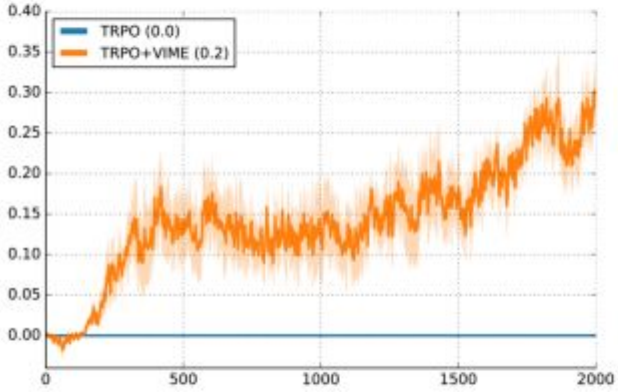
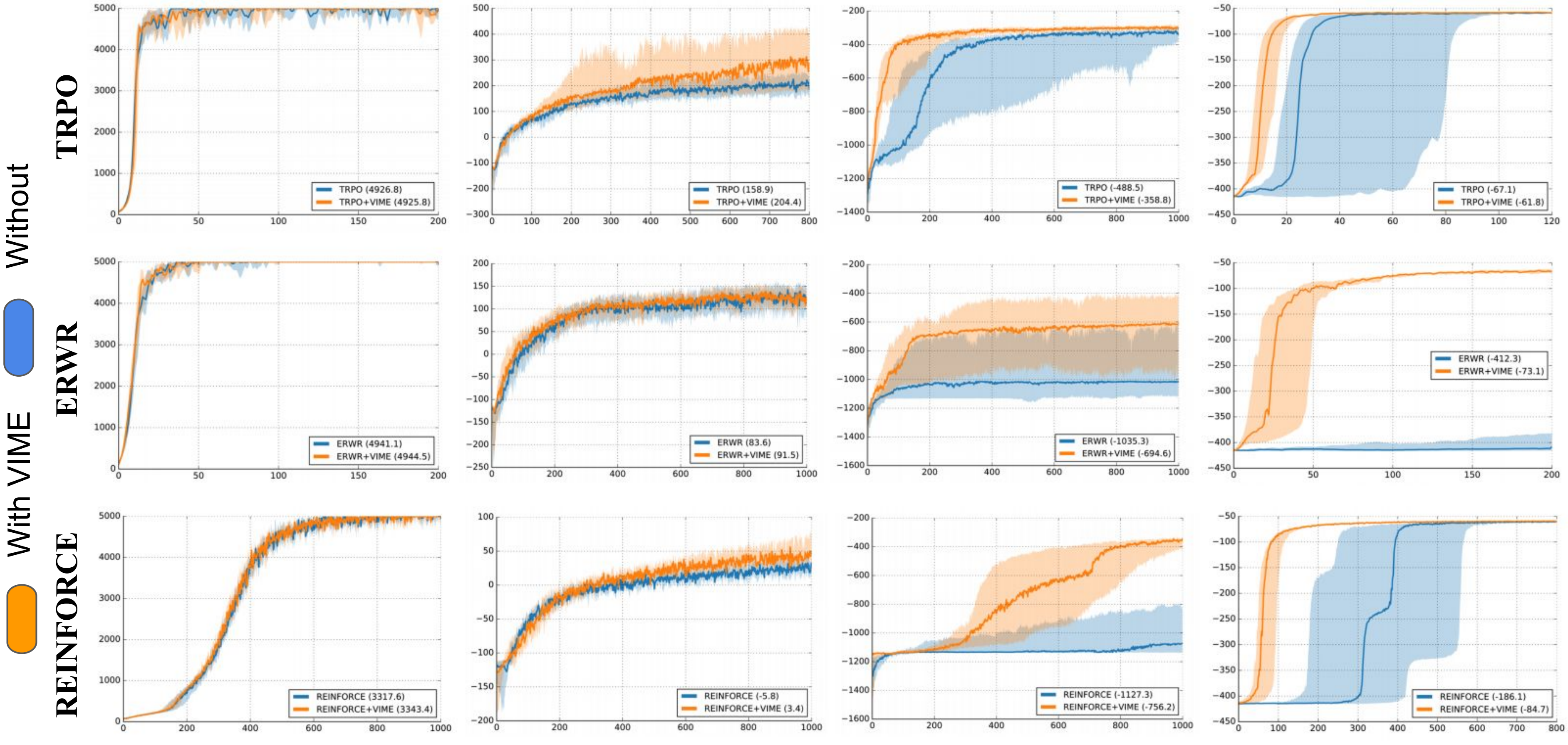


Figure 5: Performance of TRPO with and without VIME on the challenging hierarchical task SwimmerGather.

Results : environments with well shaped rewards

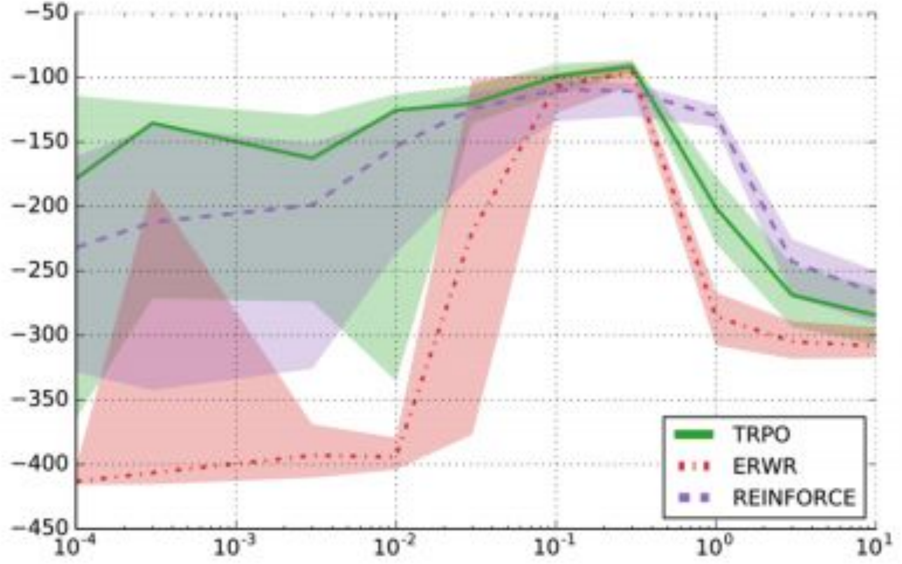


(a) CartPole (b) CartPoleSwingup (c) DoublePendulum (d) MountainCar

Results : Trading off exploration and exploitation

How does η trade off exploration and exploitation behavior?

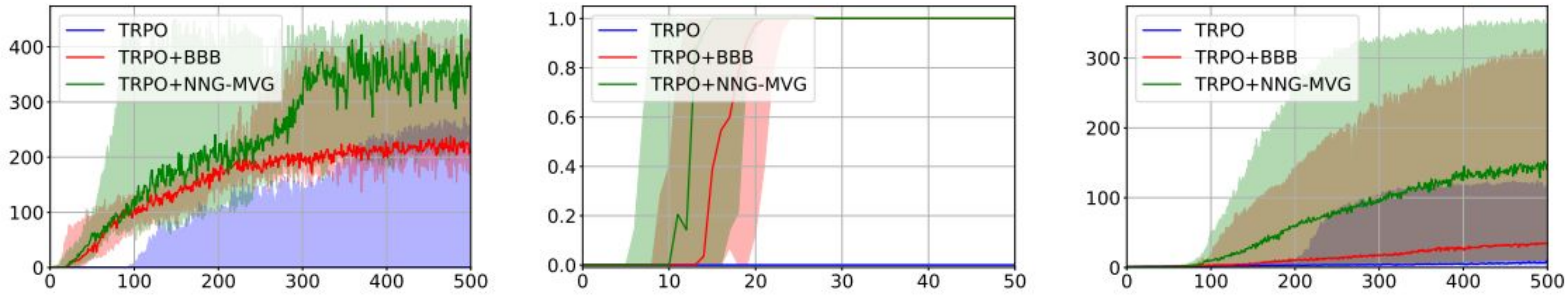
- Setting η too high clearly results in prioritizing exploration over getting additional external reward.
- Too low of an η value reduces the method to the baseline algorithm



$$r'(s_t, a_T, s_{t+1}) = r(s_t, a_t) + \eta D_{\mathrm{KL}}\left[p(\theta|\xi_t, a_t, s_{t+1})\,||p(\theta|\xi_t)\right]$$

Discussion and Limitations

- Empirical results show that VIME performs significantly better than heuristic exploration methods across various continuous control tasks and algorithms.
- Mean Field Variational Inference underestimates uncertainty → however you can replace the diagonal Gaussian posterior with one including covariance like [Noisy Natural Gradient as Variational Inference].



(a) CartPoleSwingup (b) MountainCar (c) DoublePendulum