EMPOWERMENT UNDER UNCERTAINTY: APPROXIMATE BAYESIAN EXPLORATION IN MODEL-BASED RL

Oliver Obst*
Centre for Research in Mathematics and Data Science
Western Sydney University
Australia
o.obst@westernsydney.edu.au

Fabian C. Weigend
John A. Paulson School of Engineering and Applied Sciences
Harvard University
United States of America
fweigend@seas.harvard.edu

Lisa Graf
Neurorobotics Lab, Department of Computer Science
University of Freiburg
Germany
lgraf@cs.uni-freiburg.de

Exploration in reinforcement learning (RL) remains a fundamental challenge, particularly in environments with sparse rewards or uncertain dynamics. Intrinsic motivation methods such as curiosity (Pathak et al., 2017) and empowerment (Klyubin et al., 2005) offer principled exploration objectives by encouraging agents to acquire knowledge or gain influence over future states. Empowerment, formalised as the mutual information between actions and successor states, quantifies an agent's potential to control its future. However, existing formulations typically assume access to a known or deterministic transition model, limiting applicability in real-world scenarios where dynamics must be learned and remain uncertain (Klyubin et al., 2005; Choi et al., 2021).

We propose Bayesian Approximate Empowerment for Reinforcement Learning Exploration (BAERLE), a model-based method that enables empowerment-based exploration under transition uncertainty. BAERLE estimates the distribution over empowering actions using Stein Variational Gradient Descent (SVGD) (Liu & Wang, 2016), a particle-based variational inference technique. Unlike prior approaches that rely on ensemble variance or prediction error, BAERLE directly infers a particle approximation to the empowerment-maximising action distribution by propagating SVGD particles through a learned differentiable dynamics model. This naturally captures epistemic uncertainty through particle diversity and provides a flexible mechanism for both exploration and robustness.

In our approach, particles refer to samples of actions maintained in parallel to approximate the optimal action distribution that maximises empowerment. Unlike variational inference methods that use particles to represent uncertainty over latent variables or model parameters, each particle is an action in an agent's continuous action space.

Empowerment at a given state s is defined as the mutual information between actions A and next states S':

$$I(A; S' \mid s) = H(S' \mid s) - H(S' \mid A, s),$$

which quantifies the influence of actions on future outcomes. Since computing this exactly is intractable in most settings, we approximate the optimal action distribution using a variational objective:

$$I(A; S') = \max_{q(a)} \mathbb{E}_{a\sim q}\, \mathbb{E}_{s'\sim p_\phi(\cdot|s,a)} \left[\log p_\phi(s' \mid s, a) - \log p_\phi(s')\right],$$

*Corresponding author

where $p_\phi(s' \mid s, a)$ is the learned transition model and $p_\phi(s')$ is the marginal next-state distribution under $q(a)$.

We approximate the target distribution

$$\pi(a) \propto \exp\left(\mathbb{E}_{s'}\left[\log p_\phi(s' \mid s, a)\right] - \log p_\phi(s')\right)$$

by transporting a set of N particles $\{a_i\}_{i=1}^N$ via SVGD. Each particle is updated iteratively using the rule:

$$a_i \leftarrow a_i + \eta\, \phi(a_i),$$

where the SVGD direction $\phi(a)$ is given by

$$\phi(a) = \frac{1}{N} \sum_{j=1}^{N} \left[K(a_j, a)\, g(a_j) + \nabla_{a_j} K(a_j, a)\right],$$

with $K(\cdot, \cdot)$ an RBF kernel and

$$g(a) = \nabla_a \left[\mathbb{E}_{s'} \log p_\phi(s' \mid s, a) - \log p_\phi(s')\right].$$

This particle-based optimisation captures both the expected empowerment and its uncertainty through the diversity of actions sampled. Once particles converge, empowerment is estimated using a log-partition approximation over the particle log-scores. We interpret the final particle set as an approximate posterior over empowering actions.

To avoid pathological exploration of unpredictable but uncontrollable transitions, we apply a thresholding mechanism. If the expected empowerment at state s is below a small constant $\delta > 0$, we suppress the intrinsic reward. The reward is otherwise computed as:

$$r_{\text{int}} = \begin{cases} \alpha\, \mathbb{E}[I(A; S')] + \beta\, \text{Var}(I(A; S')) & \text{if } \mathbb{E}[I(A; S')] > \delta, \\ 0 & \text{otherwise,} \end{cases}$$

where α and β weight the contributions of control and exploration. This formulation explicitly combines empowerment with epistemic uncertainty, prioritising states where both control and learning potential are high.

We embed BAERLE into a model-based RL loop. The dynamics model f_ϕ is trained from collected transitions via maximum likelihood, and the policy is updated using a standard RL optimiser (e.g., PPO or SAC) with augmented rewards $r = r_{\text{ext}} + r_{\text{int}}$. SVGD updates are performed for a subset of states in each batch, allowing scalable computation of intrinsic rewards. All gradients are computed via automatic differentiation in PyTorch, and SVGD uses $N = 100$ particles with $T = 50$ iterations and a median-heuristic kernel bandwidth.

We embed BAERLE into a model-based RL loop. The dynamics model f_ϕ is trained from collected transitions via maximum likelihood, and the policy is updated using a standard RL optimiser (e.g., PPO or SAC) with augmented rewards $r = r_{\text{ext}} + r_{\text{int}}$. SVGD updates are performed for a subset of states in each batch, allowing scalable computation of intrinsic rewards. All gradients are computed via automatic differentiation in PyTorch, and SVGD uses $N = 100$ particles with $T = 50$ iterations and a median-heuristic kernel bandwidth.

Our implementation demonstrates that BAERLE is computationally tractable and compatible with modern deep RL pipelines. Early experiments on stochastic control tasks suggest that the method produces diverse and structured behaviour, but a full evaluation of exploration efficiency and downstream policy performance remains ongoing. We plan comparisons against curiosity-driven baselines, e.g., (Pathak et al., 2017; Houthooft et al., 2016).

In summary, BAERLE offers a scalable and principled approach to exploration under uncertainty, bridging empowerment with Bayesian inference. By explicitly considering both controllability and epistemic uncertainty, it provides a foundation for future work on intrinsically motivated agents that reason under uncertainty about their own influence. We present this as an initial investigation into the idea, with further analysis and benchmarking underway.

REFERENCES

Jongwook Choi, Archit Sharma, Honglak Lee, Sergey Levine, and Shixiang Shane Gu. Variational Empowerment as Representation Learning for Goal-Conditioned Reinforcement Learning. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 1953--1963. PMLR, July 2021. URL https://proceedings.mlr.press/v139/choi21b.html.

Rein Houthooft, Xi Chen, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. VIME: Variational Information Maximizing Exploration. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://proceedings.neurips.cc/paper_files/paper/2016/hash/abd815286ba1007abfbb8415b83ae2cf-Abstract.html.

A.S. Klyubin, D. Polani, and C.L. Nehaniv. Empowerment: A Universal Agent-Centric Measure of Control. In *2005 IEEE Congress on Evolutionary Computation*, volume 1, pp. 128--135, Edinburgh, Scotland, UK, 2005. IEEE. ISBN 978-0-7803-9363-9. doi: 10.1109/CEC.2005.1554676.

Qiang Liu and Dilin Wang. Stein Variational Gradient Descent: A General Purpose Bayesian Inference Algorithm. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016. URL https://papers.nips.cc/paper_files/paper/2016/hash/b3ba8f1bee1238a2f37603d90b58898d-Abstract.html.

Deepak Pathak, Pulkit Agrawal, Alexei A. Efros, and Trevor Darrell. Curiosity-Driven Exploration by Self-Supervised Prediction. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 488--489, Honolulu, HI, USA, July 2017. IEEE. ISBN 978-1-5386-0733-6. doi: 10.1109/CVPRW.2017.70.