SCIENCE CHINA
Information Sciences

• LETTER •

September 2026, Vol. 69, Iss. 9, 199201:1–199201:2
https://doi.org/10.1007/s11432-025-4846-1

Cumulative predictive information-regularized reinforcement learning for energy-efficient robot skill learning

Bang YOU[1], Di GUO[2], Fuchun SUN[3] & Huaping LIU[3*]

[1]*School of Information Engineering, Wuhan University of Technology, Wuhan 430070, China*
[2]*School of Artificial Intelligence, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[3]*Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China*

Received 19 January 2025/Revised 9 January 2026/Accepted 3 March 2026/Published online 13 August 2026

Citation You B, Guo D, Sun F C, et al. Cumulative predictive information-regularized reinforcement learning for energy-efficient robot skill learning. Sci China Inf Sci, 2026, 69(9): 199201, https://doi.org/10.1007/s11432-025-4846-1

Deep reinforcement learning (RL) has demonstrated remarkable advancements in handling challenging robot tasks in embodied intelligence [1]. Typically, RL algorithms can automatically discover a policy, or a controller, that maps sensory observations into actions without the effort of modeling dynamics. However, they have no guarantee of achieving low energy consumption control, since policies are usually learned solely by maximizing expected cumulative rewards. This limitation prevents the practical applications of RL algorithms in untethered robotic systems, such as drones or legged robots, where batteries powering their operation provide only a limited energy supply, and enabling energy-efficient control is critical.

In this study, we aim to develop an RL framework in which energy-efficient controllers emerge automatically. A predictable controller, which solves tasks by generating action sequences that are easily predicted and consistent, can reduce energy consumption by avoiding unnecessary variations in motor movements, since unnecessary motor variations can produce high-frequency torque fluctuations and cause high energy consumption. Consider, for example, the forward walking of a legged robot. Moving the robot body in a periodic and alternating pattern is more energy-efficient compared to erratic locomotion gaits. Previous studies have already observed the benefits of periodic or predictable locomotion gaits in improving energy efficiency [2]. However, learning a predictable policy is challenging in RL settings.

A technique to learn predictable policies is to limit the complexity of the policy. For example, robust predictable control (RPC) [3] learns a predictable policy by minimizing the mutual information between states and their representations. However, limiting the information in states used for decision-making may degrade performance. Another solution is to improve the predictability of actions directly. A recent work, LZSAC [4], proposes to use compression algorithms to measure the predictability of actions and bias the agent to select actions that are easily predicted. However, directly predicting actions may not be effective due to the stochastic nature of policies.

The principle of mutual information maximization [5] in information theory is effective in capturing the underlying statistical dependencies in data, and has been well applied in the machine learning domain. Recent applications include learning representations, dataset distillation, and multi-agent collaboration. However, the use of mutual information maximization for inducing energy-efficient RL policies remains largely unexplored.

In this study, we propose to learn predictable RL policies by maximizing rewards and cumulative predictive information along the trajectories of an agent, which can improve energy efficiency by avoiding unnecessary motor movements. By additionally maximizing the predictive information, the agent is biased to learn state representations that are easily predicted, indirectly improving the predictability of actions. To effectively estimate the predictive information, we construct two parametric lower bounds on the predictive information, namely a lower bound based on a generative dynamic model, and a lower bound based on a discriminative model. We then use these lower bounds to construct an information-regularized reward function, which guides the agent to choose behaviors that result in transitions that the agent can predict well. The agent will receive rewards for choosing behaviors that are easily modeled by its learned models. Furthermore, we propose a practical algorithm that uses a single objective to jointly optimize the policy, the dynamic model, and the discriminative model to be self-consistent, resulting in a policy that generates predictable behaviors. We refer to the agent as the cumulative predictive information-regularized reinforcement learning agent or CPRL agent. We implement our algorithm based on the soft actor-critic (SAC) algorithm for learning a continuous policy and evaluate our algorithm on a set of challenging MuJoCo locomotion tasks. Empirical evaluations demonstrate that our algorithm outperforms leading baselines in terms of learning speed and final performance. We also observe that the action sequences produced by the CPRL agent are easier to predict, while achieving higher energy efficiency than previous approaches.

Method. The proposed method aims to build RL agents that use energy-efficient policies to handle control tasks. Predictable behaviors usually contain periodic and consistent movement patterns

* Corresponding author (email: hpliu@tsinghua.edu.cn)

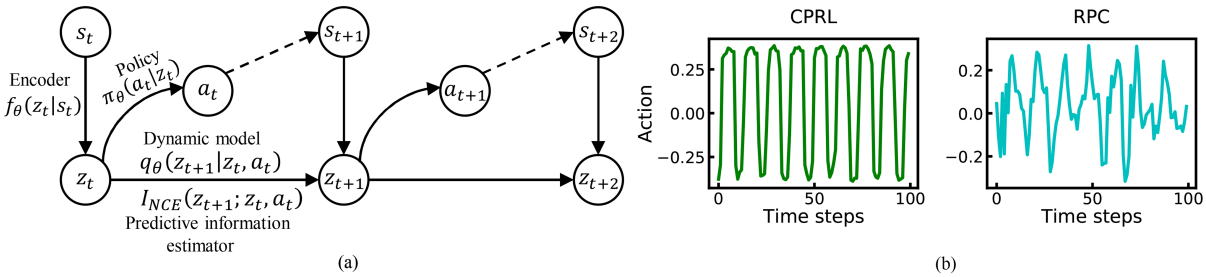

Figure 1 (Color online) (a) Our method uses one unified objective to jointly optimize the policy, the Kullback-Leibler divergence-based (KL-based) lower bound based on the encoder and the dynamic model, and the InfoNCE lower bound; (b) action sequences produced by CPRL show more periodic patterns than RPC.

and can avoid unnecessary variations in robot motor movements, thereby decreasing the energy consumption required to achieve a control task. Hence, we expect to bias RL agents to produce predictable behaviors. To this end, the proposed optimization objective of the RL agent is to maximize the expected cumulative predictive information over the trajectories of the agent and the rewards. The agent is encouraged to retain predictive information cumulatively per episode. At a single time step, the predictive information measures the correlation between the next state representation and the current state representation paired with the current action. By introducing a Lagrange multiplier, we can convert the optimization objective with the information constraint into an unconstrained one.

Intuitively, by additionally maximizing predictive information, the agent is encouraged to learn representations that are easily predictable. A policy that relies on these predictable representations for decision-making is likely to produce temporally consistent behaviors. Furthermore, to obtain representations that are easily predictable, the agent tends to select predictable behaviors, which in turn result in states that are more predictable. The resulting predictable action sequence will improve energy efficiency by reducing unnecessary variations in movements.

To effectively estimate the predictive information, we construct a variational Kullback-Leibler divergence-based (KL-based) lower bound of the mutual information by introducing a parametric variational distribution, and an InfoNCE lower bound of the mutual information based on a discriminative model. We use both the InfoNCE lower bound and the KL-based lower bound to estimate the predictive information in order to avoid model collapse and capture environmental dynamics effectively. By plugging the lower bounds into the optimization objective, we can obtain the final objective that jointly optimizes the policy and other learnable models to maximize the rewards and the predictive information. The final objective is illustrated in Figure 1(a). Based on the final objective, we can obtain an information-regularized reward function that augments the environment rewards with the lower bounds on the predictive information. To achieve practical optimization, we develop the CPRL algorithm based on the information-regularized reward function and SAC. We provide a visualization of action sequences produced by our CPRL algorithm in Figure 1(b). Policies learned by CPRL generate a consistent and predictable action trajectory, featuring a high degree of periodicity. We refer to Appendixes B and C for the details of our method.

Experiments. We evaluate the performance of our method and the state-of-the-art baselines on six challenging continuous robot locomotion tasks. Experimental results show that our algorithm achieves faster learning speed and higher rewards than the baselines on the majority of tasks. Experimental details and results can be found in Appendixes D.1 and D.2.

Moreover, we conduct experiments to evaluate the energy efficiency of our algorithm based on the Cost of Transport. Experimental results demonstrate that our algorithm achieves higher energy efficiency than leading baselines. We also evaluate the predictability of our policies by both visualizing them and using a lossless compression algorithm. Empirical results show that our algorithm produces the most predictable action trajectories, featuring a high degree of periodicity, among all methods. We refer to Appendixes D.3 and D.4 for the details of these experiments.

Finally, we perform ablation studies to investigate the effect of individual components of our CPRL model. The results show that preserving predictive information is helpful in achieving good performance, while using the KL-based lower bound or the InfoNCE lower bound also improves performance. The ablation of hyperparameters demonstrates that our algorithm is robust to small changes in the key hyperparameters. Please refer to Appendix D.5 for the details of our ablation studies.

Conclusion and discussion. In this study, we presented CPRL, which learns energy-efficient policies by maximizing the cumulative predictive information in the trajectories of agents. Our approach uses one objective to jointly optimize the policy, the dynamic model, and the InfoNCE lower bound of predictive information to be self-consistent, resulting in a policy that generates predictable behaviors and state representations. Experimental results on a set of challenging locomotion tasks show that our algorithm achieves better performance and energy efficiency than leading approaches. Moreover, we observed that our CPRL agent generates periodic and predictable action sequences.

While achieving strong performance, CPRL shares a limitation common to methods for learning predictable policies: predictable policies can degrade the performance for some tasks, such as robot combat and playing poker, where complex and unpredictable behaviors are demanded.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 62025304, 62120106005) and Beijing Natural Science Foundation (Grant No. L253006).

Supporting information Appendixes A–D. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Liu H, Guo D, Cangelosi A. Embodied intelligence: a synergy of morphology, action, perception and learning. ACM Comput Surv, 2025, 57: 1–36
2 Kashiri N, Abate A, Abram S J, et al. An overview on principles for energy efficient robot locomotion. Front Robot AI, 2018, 5: 129
3 Eysenbach B, Salakhutdinov R R, Levine S. Robust predictable control. In: Proceedings of Advances in Neural Information Processing Systems, 2021
4 Saanum T, Elteto N, Dayan P, et al. Reinforcement learning with simple sequence priors. In: Proceedings of Advances in Neural Information Processing Systems, 2024
5 You B, Liu H, Peters J, et al. Maximum total correlation reinforcement learning. In: Proceedings of International Conference on Machine Learning, 2025