Proceedings of Machine Learning Research vol 120:1–11, 2020

Learning to Plan via Deep Optimistic Value Exploration

Tim Seyde* TSEYDE@CSAIL.MIT.EDU
Wilko Schwarting* WILKOS@CSAIL.MIT.EDU
MIT Computer Science and Artificial Intelligence Laboratory, USA

Sertac Karaman SERTAC@MIT.EDU
MIT Laboratory for Information and Decision Systems, USA

Daniela Rus RUS@CSAIL.MIT.EDU
MIT Computer Science and Artificial Intelligence Laboratory, USA

Abstract

Deep exploration requires coordinated long-term planning. We present a model-based reinforcement learning algorithm that guides policy learning through a value function that exhibits optimism in the face of uncertainty. We capture uncertainty over values by combining predictions from an ensemble of models and formulate an upper confidence bound (UCB) objective to recover optimistic estimates. Training the policy on ensemble rollouts with the learned value function as the terminal cost allows for projecting long-term interactions into a limited planning horizon, thus enabling deep optimistic exploration. We do not assume a priori knowledge of either the dynamics or reward function. We demonstrate that our approach can accommodate both dense and sparse reward signals, while improving sample complexity on a variety of benchmarking tasks.

Keywords: Reinforcement Learning, Deep Exploration, Model-Based, Value Function, UCB

1. Introduction

Reinforcement learning (RL) provides a framework for intelligent agents to acquire complex behaviors autonomously. Selecting an interaction strategy that ensures efficiency of the learning process remains a challenge. This concern is prevalent in domains with continuous state and action spaces and exacerbated by problem dimensionality. Robotics applications typically feature continuous control over high-dimensional state spaces. Enabling autonomous robots to learn temporally extended behaviors through interaction, therefore, requires focused, information-dense sampling strategies.

Model-based reinforcement learning (MBRL) informs decision-making in the real world by estimating the performance of candidate actions on an environment model. The control policy is optimized by solving a finite-horizon planning problem involving the model dynamics and objective function. Some formulations leverage nominal models to recover value estimates (Lowrey et al. (2019); Seyde et al. (2019)), while others employ learned models without considering behavior beyond the preview horizon (Kurutach et al. (2018); Chua et al. (2018); Clavera et al. (2018)). True autonomy arises at the intersection of these approaches, where the agent is capable of continuously refining its belief about environment dynamics and task objective, while efficiently planning towards goals that may extend far into the future.

* Equal contribution

In this work, we learn coordinated long-term planning in scenarios where both the environment dynamics and task objective are not known a priori. Our algorithm guides policy learning through a value function that exhibits optimism in the face of uncertainty. We capture uncertainty over values by combining predictions from a model ensemble (Lakshminarayanan et al. (2017); Pearce et al. (2018)) and formulate an upper confidence bound (UCB) objective (Auer et al. (2002); Krause and Ong (2011)) to recover optimistic performance estimates. Training the policy on ensemble rollouts with the learned value function as the terminal cost allows for projecting long-term interactions into a finite planning horizon, thus enabling deep optimistic exploration with minimal prior information. The contributions of this paper are:

- A policy optimization not requiring a priori knowledge about the dynamics or objective, that can accommodate sparse reward signals, and is applicable to high-dimensional control tasks
- A framework for efficient deep exploration that leverages an uncertainty-aware value function
- Improved sample complexity over state-of-the-art RL algorithms on a set of benchmark tasks

2. Related Work

Model-free reinforcement learning (MFRL) algorithms have solved a variety of challenging problems by forgoing sample complexity in favor of asymptotic performance (Silver et al. (2016); Fujimoto et al. (2018); Haarnoja et al. (2018); Hwangbo et al. (2019)). The efficiency of MBRL approaches has been demonstrated with parametric linear models (Levine and Abbeel (2014); Kumar et al. (2016)) and non-parametric Gaussian process models (Kuss and Rasmussen (2004); Deisenroth and Rasmussen (2011); Kamthe and Deisenroth (2018)) in low-dimensional settings. Moving to higher dimensions with hybridized states, modelling the dynamics with neural networks is a common practice for both state-space (Kurutach et al. (2018); Nagabandi et al. (2018); Chua et al. (2018); Clavera et al. (2018)) and latent space planning (Hafner et al. (2019)). Low sample density induces bias in these representations, which can be alleviated by ensembling (Kurutach et al. (2018); Chua et al. (2018); Clavera et al. (2018)). Finite horizon model rollouts are then used to either train a policy (Kurutach et al. (2018); Clavera et al. (2018)) or solve an MPC-type optimization (Nagabandi et al. (2018); Chua et al. (2018); Hafner et al. (2019)). These approaches limit predictions to the preview window and select actions either greedily or with added stochasticity, neglecting structured exploration. Here, we also mitigate model bias by considering ensemble rollouts but additionally leverage the associated uncertainty in achieving efficient long-term exploration.

Directed exploration strategies have been extensively studied for discrete action spaces. One line of research adds an information gain bonus to capture unexpected environment behavior (Stadie et al. (2015); Ostrovski et al. (2017); Pathak et al. (2017)). Interactions are then driven by uncertainty over the dynamics and not over long-term rewards. Osband et al. (2016a, 2017) consider the latter implicitly by extending their work on randomized value functions (Osband et al. (2016b)) in combination with DQN (Mnih et al. (2013)) to model a distribution over value functions and sampling from the posterior. Chen et al. (2017) build on this idea by combining the mean and variance of a value ensemble into a UCB objective for action selection. O'Donoghue et al. (2018) furthermore propose the Uncertainty Bellman equation to improve uncertainty propagation in deep exploration.

In the continuous domain, uncertainty-aware objectives based on predicted disagreement of model behavior (Still and Precup (2012); Houthooft et al. (2016); Henaff (2019)) or trajectory returns

(Depeweg et al. (2018)) have been well-studied for finite-horizon rollouts. Lowrey et al. (2019) consider infinite horizon planning by leveraging an ensemble of value functions in guiding an MPC into regions of uncertain returns for continuous control under a known nominal model. Here, we extend the idea of uncertainty driven value exploration to scenarios with unknown dynamics and objective functions by combining finite-horizon ensemble rollouts with an optimistic value function.

3. Preliminaries

We formulate the underlying optimization problem as a Markov decision process (MDP) defined by the tuple $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, f, r, \rho_0, \gamma\}$, where $\mathcal{S} \in \mathbb{R}^n$ denotes the state space, $\mathcal{A} \in \mathbb{R}^m$ the action space, $f \colon \mathcal{S} \times \mathcal{A} \to \mathcal{S}$ the transition function, $r \colon \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ the reward function, ρ_0 the initial state distribution, and $\gamma \in [0, 1)$ the discount factor. We define s_t and a_t to be the state and action at time t, respectively, and use the notation $r_t = r(s_t, a_t)$. Let $\pi_\theta \colon \mathcal{S} \to \mathcal{A}$ denote a deterministic policy parameterized by θ and define the discounted infinite horizon return $\eta(\theta) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$, where $s_0 \sim \rho_0$, $s_{t+1} = f(s_t, a_t)$, and $a_t = \pi_\theta(s_t)$. The objective is to find the optimal policy π_θ^* that maximizes the return $\eta(\theta)$, where we treat both the dynamics and reward function as unknown.

4. Model-Based Deep Reinforcement Learning

Model-based policy learning constructs an environment model to inform real-world interactions. In the online phase, the policy is used to interact with the environment and corresponding observations are appended to the memory $\mathcal{D}$. In the offline phase, the memory is queried to refine the model and to propagate information into the policy by training on simulated interactions.

4.1. Model Learning

The environment interaction at time t is represented as the tuple (s_t, a_t, s_{t+1}, r_t), corresponding to the discrete time transition $s_{t+1} = f(s_t, a_t)$. Here, we do not assume prior knowledge of the dynamics or the objective function and model them using function approximators $\hat{f}_\phi$ and $\hat{r}_\psi$, parameterized by ϕ and ψ, respectively. We apply standard supervised learning techniques in combination with episodic warm starts to optimize the generalized objective

$$\min_{\omega} \frac{1}{|\mathcal{D}|} \sum_{(s_t, a_t, s_{t+1}, r_t) \in \mathcal{D}} \left\| \tau - \hat{g}_\omega(s_t, a_t) \right\|_2^2, \tag{1}$$

where $\hat{g}_\omega = \{\hat{f}_\phi, \hat{r}_\psi\}$ denotes the function approximator and $\tau = \{s_{t+1} - s_t, r_t\}$ the target vector.

4.2. Policy Learning

Our goal is to find the optimal policy $\pi_\theta^*(s_t)$ that maximizes the return $\eta(\theta) = \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$. This objective is computationally intractable and we instead re-formulate it using model rollouts over a finite horizon T with a value function as the terminal return. The policy objective becomes

$$\eta(\theta, \phi, \psi, \nu, T) = \sum_{t=0}^{T-1} \gamma^t \hat{r}_\psi(s_t, a_t) + \gamma^T \hat{V}_\nu^\pi(s_T), \tag{2}$$

where $s_0 \in \mathcal{D}$, $a_t = \pi_\theta(s_t)$, $s_{t+1} = \hat{f}_\phi(s_t, a_t)$, and the true value function V^{π^*} is approximated under the current policy π_θ using a neural network parameterized by ν. The value function is trained synchronously with the policy using fitted value iteration on the Bellman backups described by equation (2), where we note that $V_\nu^\pi(s_0) = \eta(\theta, \phi, \psi, \nu, T)$ under the optimal policy $\pi_\theta = \pi^*$. For notational convenience, we use the abbreviation $\eta(\theta, \nu) \coloneqq \eta(\theta, \phi, \psi, \nu, T)$.

5. Deep Exploration through Model Uncertainty

The environment model trained in section 4.1 exhibits strong bias in regions of low sample density, while the deterministic policy of section 4.2 overfits to simulated data. Exploratory behavior then arises from exploiting model mismatches. To overcome these limitations, we introduce our method of learning a deterministic policy that is uncertainty-aware and intrinsically exhibits long-term explorative behavior. The policy training is guided by an optimistic value function that encodes the long-term potential of actions. Uncertainty estimates over environment behavior are recovered by leveraging an ensemble of models.

5.1. Model Learning with Uncertainty Estimation

In regions of low sample density, observed data constrains the network weights only weakly and the influence of random biases such as network initialization and the order of observed training samples become more prevalent. To reduce the effect of these biases, a model ensemble can be employed. We define an ensemble as a collection of M particles. Each particle is assigned a unique pairing of a dynamics and a reward function, $\{\{\hat{f}_{\phi_1}, \hat{r}_{\psi_1}\}, \ldots, \{\hat{f}_{\phi_M}, \hat{r}_{\psi_M}\}\}$, representing unique hypotheses on environment behavior. The particles are used in equation (2) to generate a set of predicted returns

$$\eta_i(\theta, \nu) = \sum_{t=0}^{T-1} \gamma^t \hat{r}_{\psi_i}(s_{i,t}, a_{i,t}) + \gamma^T \hat{V}_\nu^\pi(s_{i,T}), \tag{3}$$

where $s_{i,0} = s_0$, $a_{i,t} = \pi_\theta(s_{i,t})$, and $s_{i,t+1} = \hat{f}_i(s_{i,t}, a_{i,t})$. Particle distinctness is encouraged by varying the initial network weights and training batch order. Predicted trajectory returns with uncertainty estimates are then obtained by computing the ensemble mean μ_η and variance σ_η^2

$$\mu_\eta(\theta, \nu) = \frac{1}{M} \sum_{i=1}^{M} \eta_i(\theta, \nu), \quad \sigma_\eta^2(\theta, \nu) = \frac{1}{M} \sum_{i=1}^{M} \left(\eta_i(\theta, \nu) - \mu_\nu(\theta, \nu)\right)^2, \tag{4}$$

where μ_η provides a de-biased estimator of the return and σ_η^2 an estimator of prediction uncertainty.

5.2. Policy Learning with Directed Exploration

Replacing the policy objective in equation (2) by the predicted mean in equation (4) reduces model-specific bias in the policy. The resulting greedy strategy would be reliant on random exploration. Instead, we encourage active exploration by considering the potential improvement via the predicted variance in equation (4). We define the policy objective via the upper confidence bound (UCB)

$$\eta_{UCB}(\theta, \nu) = \mu_\eta(\theta, \nu) + \beta \sigma_\eta(\theta, \nu), \tag{5}$$

where the scalar β quantifies the exploration-exploitation trade-off. The greedy policy is recovered at $\beta = 0$, while increasing β increases the optimism that uncertainty translates to potential for improvement. The objective flexibly scales exploratory behavior, while remaining more robust to outlier predictions from the ensemble than taking the maximum. The resulting policy is uncertainty-aware and behaves intrinsically explorative. However, exploration is restricted by the preview horizon as uncertainty will only propagate locally up until the terminal reward is queried. To address this limitation, we formulate our value function to encourage long-term exploration beyond the preview horizon by defining it over the infinite horizon UCB return. This modified value function implicitly encodes optimism in the face of uncertainty as it is biased towards the maximum return of the ensemble. It is approximated by the value network $\hat{V}_\nu^\pi$ and trained via fitted value iteration on $\eta_{UCB}(\theta, \nu)$. The policy and value function are trained concurrently on the objectives

$$\max_\theta \sum_{s_0 \in \mathcal{D}} \eta_{UCB}(\theta, \nu), \quad \min_\nu \sum_{s_0 \in \mathcal{D}} \left\| \eta_{UCB}(\theta, \nu) - \hat{V}_\nu^\pi(s_0) \right\|_2^2, \tag{6}$$

where $s_0 \in \mathcal{D}$. This process yields a purely deterministic policy capable of exploration through global uncertainty-awareness, while only requiring training on a finite preview horizon.

6. Deep Optimistic Value Exploration (DOVE)

The algorithm runs for K episodes, alternating between two phases: in the online phase, the policy is used to interact with the environment for N timesteps and the observed transitions are appended to memory $\mathcal{D}$. In the offline phase, the environment models are updated according to equation (1) using common supervised learning practices, while varying the batch order between ensemble members. The policy and value function are optimized iteratively on the objectives in equation (6). Each iteration consists of a policy optimization step under the current value function, followed by a value function optimization step under the updated policy. The corresponding initial conditions are generated by locally perturbing states from memory to ensure that information propagation is not limited to on-policy observations. All networks are trained with the Adam optimizer (Kingma and Ba (2014)). A schematic representation of the approach is provided in Algorithm 1.

Algorithm 1: Deep Optimistic Value Exploration (DOVE)

Initialize: $\mathcal{D} \leftarrow \emptyset$ and $\{\phi_i, \psi_i, \theta, \nu\} \leftarrow \mathcal{U}(a, b)$
for $i \leftarrow 1$ **to** K **do** `// episodes`
  for $t \leftarrow 1$ **to** N **do** `// timesteps`
    Execute $a_t = \pi_\theta(s_t)$ in environment, add transition to $\mathcal{D}$;
  end
  Train $\{\hat{f}_{\phi_i}, \hat{r}_{\psi_i}\}_{i=1}^M$ on transitions from $\mathcal{D}$ using supervised learning;
  for $b \leftarrow 1$ **to** B **do** `// batches`
    Sample observations from $\mathcal{D}$, perturb locally to generate initial conditions;
    Policy rollout on $\{\hat{f}_{\phi_i}, \hat{r}_{\psi_i}\}_{i=1}^M$, train π_θ to maximize η_{UCB};
    Policy rollout on $\{\hat{f}_{\phi_i}, \hat{r}_{\psi_i}\}_{i=1}^M$, train $\hat{V}_\nu^\pi$ to approximate η_{UCB};
  end
end

7. Experiments

In the following, we provide results for training agents with the DOVE algorithm in various settings. First, we show-case a task with sparse reward signals to illustrate how active exploration emerges when learning an optimistic value function that is uncertainty aware. Then, we demonstrate that DOVE improves performance over state-of-the-art on four higher dimensional benchmarking tasks. We employ an ensemble of size $M = 5$ and provide other relevant parameters in Appendix A [1].

7.1. Intuitive Example: Pendulum with Sparse Rewards

The simple pendulum has several function mappings with straight-forward graphic representations. We define a swing-up task with sparse reward feedback around the upright position. The agent does not have access to the nominal dynamics or reward function and sparsity avoids guidance towards the goal through the lack of smooth reward gradients. We remove random exploration by initializing the agent at rest in the downward configuration and the policy to not generate visible motions.

Figure 1 (A-C) depicts the learned representations of the reward function, value function, and policy. Both the value function and policy accurately capture the desired swing-up and stabilization behavior, solving the task. The role of exploring through model uncertainty in building these representations is apparent by examining how the first non-zero reward is obtained in episode 5. Figure 1 depicts the UCB value function and its uncertainty before and after the interaction (D-E, respectively). Before, the downward configuration is well explored and high uncertainty remains around the upright configuration (D). The optimistic agent plans to explore this high uncertainty region as it holds potential for improvement. It performs a swing-up and rotates with positive velocity, effectively cutting uncertainty in that region of the state space (E). The observed reward is immediately propagated into the value function and guides the agent in optimizing its swing-up and stabilization behavior. Based on the UCB trade-off, the agent is furthermore capable of ignoring uncertain areas of the state space that are non-conducive to the task. This is demonstrated by the remaining reward uncertainty at high velocity regimes around the upright position in Figure 1 (A). These states would not allow for immediate stabilization. The UCB formulation therefore induces high selectivity in planning long-term interactions. This remains valid even in sparse reward settings, where reward uncertainty can be leveraged in the absence of informative mean estimates. We observe similar exploratory behavior on the mountain car with sparse rewards as displayed in Appendix B [1].

7.2. Performance on Benchmarking Tasks

The previous section highlighted the algorithm's ability to explore efficiently even with only sparse reward signals. In this section, we compare performance to state-of-the-art MBRL and MFRL algorithms on four benchmarking tasks. The tasks vary in their respective timescales, episode lengths and motion objectives, therefore constituting a concise setting to demonstrate the versatility of the approach. The *Reacher* task requires generalization of a 2D reaching behavior to arbitrary goal locations. The *Hopper*, *HalfCheetah* and *SlimHumanoid* tasks require fast forward locomotion under non-smooth impact dynamics. All tasks penalize the usage of control inputs. We analyze the learning progress on five random seeds over the first 100 episodes to highlight sample-efficiency. It needs to be noted that, while our algorithm does not have access to the nominal dynamics and

1. Please refer to the extended version of this article at https://dspace.mit.edu/handle/1721.1/125161.

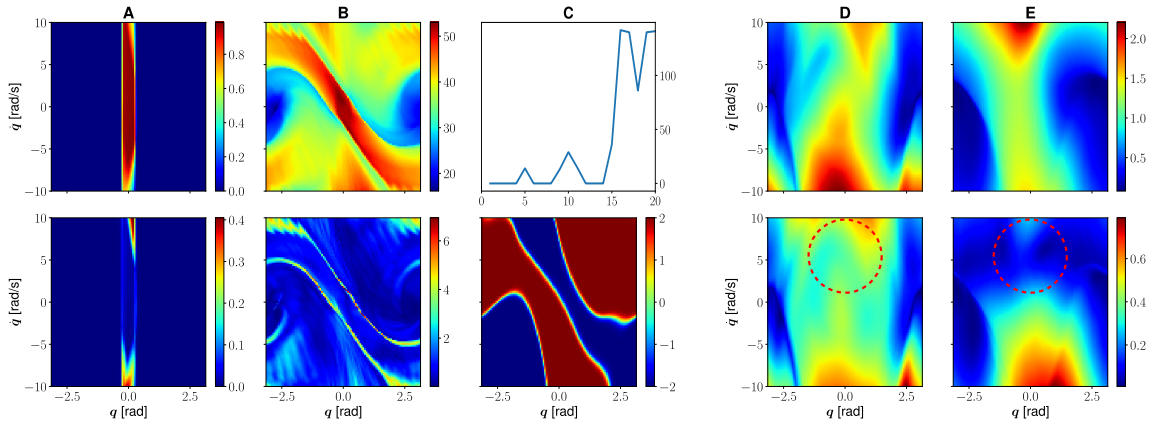

Figure 1: Pendulum swing-up from downward configuration with sparse rewards within $|\theta| \leq 15°$. Top to bottom: (A) learned reward function mean and standard deviation. (B) learned value function and standard deviation. (C) episodic rewards and learned policy. (D), (E) learned value function and standard deviation before and after episode 5, respectively. DOVE actively explores the high potential rewards of spinning with positive velocity and reduces associated uncertainty (red circle). Furthermore, uncertain regions not allowing for immediate stabilization at the top are ignored (A).

reward function, the MBRL algorithms we compare to leverage nominal rewards in their planning. The resulting performance curves are provided in Figure 2, where baseline performance is taken from Wang et al. (2019). Across all tasks, DOVE performs better than or on par with the MBRL and the MFRL baselines. This holds despite DOVE having to learn the reward signals used for planning, whereas the MBRL baselines plan on nominal rewards. The performance gap widens with increasing task dimensionality, relatively doubling scores on *Hopper* and *HalfCheetah* and increasing scores tenfold on *SlimHumanoid* after 100 episodes, underlining DOVE's ability to effectively focus exploration only on regions of the state space that exhibit strong potential for improvement. We achieve this by guiding the policy learning through optimistic value estimates, thereby predicting long-term behavior and enabling targeted deep exploration. DOVE can then efficiently learn complex, temporally extended motion patterns while planning only over short time horizons of $T < 0.5$ s. The results in Figure 2 also show that DOVE improves performance over DVE, a baseline with model ensembles and a regular value function without UCB component. Our ablation study in Appendix C [1] confirms that performance is further reduced when only considering a single ensemble particle with a non-UCB value function. This highlights the importance of deep, directed exploration facilitated by model ensembling and an uncertainty-aware optimistic value function.

8. Discussion & Conclusion

We propose DOVE, an MBRL algorithm that enables sample-efficient deep exploration with a deterministic policy. Policy learning is guided by a value function that exhibits optimism in the face of uncertainty. The value function encodes an upper confidence bound over performance estimates from a model ensemble. Training the policy on finite horizon model rollouts with the optimistic value function as the terminal reward enables computationally tractable deep optimistic exploration.

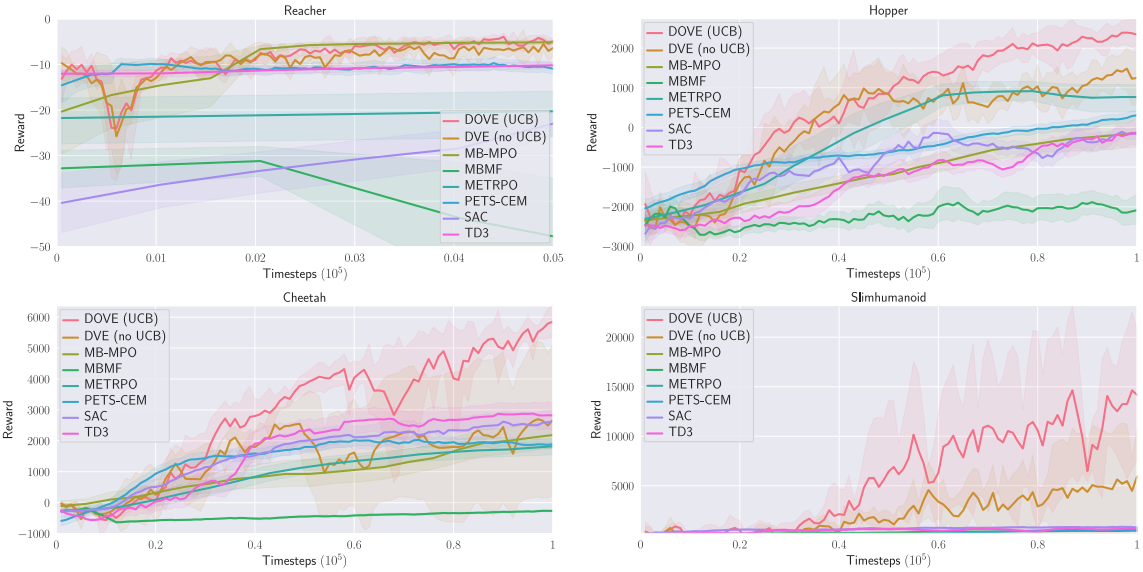

Figure 2: Performance on the *Reacher*, *Hopper*, *Cheetah*, and *SlimHumanoid* benchmarking tasks. DOVE is compared against state-of-the-art MBRL and MFRL algorithms and DVE, a variation of DOVE that only uses a non-UCB value function ($\beta = 0$). Performance is evaluated over 100 episodes, averaged over 5 random seeds and compared to baseline data from Wang et al. (2019). DOVE performs best on all tasks, while scaling gracefully with increased problem dimensionality.

The approach assumes no prior knowledge over the dynamics, reward function or policy, and learns all representations jointly. Experimental evaluation shows that DOVE efficiently solves tasks that only provide sparse reward signals and extends well to higher dimensional systems. Performance improvements over various state-of-the-art MBRL and MBFL algorithms have been demonstrated on the *Reacher*, *Hopper*, *Cheetah*, and *SlimHumanoid* benchmarking tasks. We observed that DOVE scaled much better with problem dimensionality than the other MBRL algorithms, highlighting the approach's ability to effectively focus exploration only on regions with strong potential for improvement. Furthermore, DOVE was able to do so by planning based on the learned reward functions and did not have access to the nominal rewards. In the future, we hope to build on these results by utilizing a model predictive controller in the online phase. Planning over a receding horizon is likely to further increase sample efficiency, as online re-planning will mitigate some effects of model mismatch. We are furthermore interested in extending our work to latent space planning from first-person perspective images, as they provide a concise representation of environment states that are difficult to measure directly.

Acknowledgments

This work was supported in part by the Office of Naval Research (ONR) Grant N00014-18-1-2830, Qualcomm and Toyota Research Institute (TRI). This article solely reflects the opinions and conclusions of its authors and not TRI, Toyota, or any other entity. We thank them for their support.

References

Peter Auer, Nicolo Cesa-Bianchi, and Paul Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

Richard Y Chen, Szymon Sidor, Pieter Abbeel, and John Schulman. UCB exploration via Q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

Kurtland Chua, Roberto Calandra, Rowan McAllister, and Sergey Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, pages 4754–4765, 2018.

Ignasi Clavera, Jonas Rothfuss, John Schulman, Yasuhiro Fujita, Tamim Asfour, and Pieter Abbeel. Model-based reinforcement learning via meta-policy optimization. In *Conference on Robot Learning*, pages 617–629, 2018.

Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In *International Conference on Machine Learning*, pages 465–472, 2011.

Stefan Depeweg, Jose-Miguel Hernandez-Lobato, Finale Doshi-Velez, and Steffen Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In *International Conference on Machine Learning*, pages 1184–1193, 2018.

Scott Fujimoto, Herke Hoof, and David Meger. Addressing Function Approximation Error in Actor-Critic Methods. In *International Conference on Machine Learning*, pages 1582–1591, 2018.

Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International Conference on Machine Learning*, pages 1856–1865, 2018.

Danijar Hafner, Timothy Lillicrap, Ian Fischer, Ruben Villegas, David Ha, Honglak Lee, and James Davidson. Learning Latent Dynamics for Planning from Pixels. In *International Conference on Machine Learning*, pages 2555–2565, 2019.

Mikael Henaff. Explicit Explore-Exploit Algorithms in Continuous State Spaces. In *Advances in Neural Information Processing Systems*, pages 9372–9382, 2019.

Rein Houthooft, Xi Chen, Yan Duan, John Schulman, Filip De Turck, and Pieter Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

Jemin Hwangbo, Joonho Lee, Alexey Dosovitskiy, Dario Bellicoso, Vassilios Tsounis, Vladlen Koltun, and Marco Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), 2019.

Sanket Kamthe and Marc Deisenroth. Data-Efficient Reinforcement Learning with Probabilistic Model Predictive Control. In *International Conference on Artificial Intelligence and Statistics*, pages 1701–1710, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Andreas Krause and Cheng S Ong. Contextual gaussian process bandit optimization. In *Advances in neural information processing systems*, pages 2447–2455, 2011.

Vikash Kumar, Emanuel Todorov, and Sergey Levine. Optimal control with learned local models: Application to dexterous manipulation. In *2016 IEEE International Conference on Robotics and Automation (ICRA)*, pages 378–383. IEEE, 2016.

Thanard Kurutach, Ignasi Clavera, Yan Duan, Aviv Tamar, and Pieter Abbeel. Model-ensemble trust-region policy optimization. In *International Conference on Learning Representations*, 2018.

Malte Kuss and Carl E Rasmussen. Gaussian processes in reinforcement learning. In *Advances in neural information processing systems*, pages 751–758, 2004.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

Sergey Levine and Pieter Abbeel. Learning neural network policies with guided policy search under unknown dynamics. In *Advances in Neural Information Processing Systems*, pages 1071–1079, 2014.

Kendall Lowrey, Aravind Rajeswaran, Sham Kakade, Emanuel Todorov, and Igor Mordatch. Plan Online, Learn Offline: Efficient Learning and Exploration via Model-Based Control. In *International Conference on Learning Representations*, 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.

Anusha Nagabandi, Gregory Kahn, Ronald S Fearing, and Sergey Levine. Neural network dynamics for model-based deep reinforcement learning with model-free fine-tuning. In *2018 IEEE International Conference on Robotics and Automation (ICRA)*, pages 7559–7566. IEEE, 2018.

Ian Osband, Charles Blundell, Alexander Pritzel, and Benjamin Van Roy. Deep exploration via bootstrapped DQN. In *Advances in neural information processing systems*, pages 4026–4034, 2016a.

Ian Osband, Benjamin Van Roy, and Zheng Wen. Generalization and exploration via randomized value functions. *International Conference on Machine Learning*, page 2377–2386, 2016b.

Ian Osband, Benjamin Van Roy, Daniel Russo, and Zheng Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 2017.

Georg Ostrovski, Marc G Bellemare, Aäron van den Oord, and Rémi Munos. Count-based exploration with neural density models. In *International Conference on Machine Learning*, pages 2721–2730, 2017.

Brendan O'Donoghue, Ian Osband, Remi Munos, and Volodymyr Mnih. The Uncertainty Bellman Equation and Exploration. In *International Conference on Machine Learning*, pages 3836–3845, 2018.

Deepak Pathak, Pulkit Agrawal, Alexei A Efros, and Trevor Darrell. Curiosity-driven exploration by self-supervised prediction. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

Tim Pearce, Mohamed Zaki, Alexandra Brintrup, and Andy Neel. Uncertainty in Neural Networks: Bayesian Ensembling. *arXiv preprint arXiv:1810.05546*, 2018.

Tim Seyde, Jan Carius, Farbod Farshidian, and Marco Hutter. Locomotion Planning through a Hybrid Bayesian Trajectory Optimization. In *2019 IEEE International Conference on Robotics and Automation (ICRA)*, pages 5544–5550. IEEE, 2019.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of Go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

Bradly C Stadie, Sergey Levine, and Pieter Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

Susanne Still and Doina Precup. An information-theoretic approach to curiosity-driven reinforcement learning. *Theory in Biosciences*, 131(3):139–148, 2012.

Tingwu Wang, Xuchan Bao, Ignasi Clavera, Jerrick Hoang, Yeming Wen, Eric Langlois, Shunshi Zhang, Guodong Zhang, Pieter Abbeel, and Jimmy Ba. Benchmarking model-based reinforcement learning. *arXiv preprint arXiv:1907.02057*, 2019.