Learning to Plan Optimistically: Uncertainty-Guided Deep Exploration via Latent Model Ensembles

Anonymous Author(s)
Affiliation
Address
email

Abstract: Learning complex robot behaviors through interaction requires structured exploration. Planning should target interactions with the potential to optimize long-term performance, while only reducing uncertainty where conducive to this objective. This paper presents Latent Optimistic Value Exploration (LOVE), a strategy that enables deep exploration through optimism in the face of uncertain long-term rewards. We combine latent world models with value function estimation to predict infinite-horizon returns and recover associated uncertainty via ensembling. The policy is then trained on an upper confidence bound (UCB) objective to identify and select the interactions most promising to improve long-term performance. We apply LOVE to visual robot control tasks in continuous action spaces and demonstrate on average more than 15% improved sample efficiency in comparison to state-of-the-art and other exploration objectives. In sparse and hard to explore environments we achieve an average improvement of over 30%.

Keywords: Learning Control, Sample Efficiency, Exploration

1 Introduction

The ability to learn complex behaviors through interaction will enable the autonomous deployment of various robotic systems in the real world. Reinforcement learning (RL) provides a key framework for realizing these capabilities, but efficiency of the learning process remains a prevalent concern. Real-life applications yield complex planning problems due to high-dimensional environment states, which are further exacerbated by the agent's continuous action space. For RL to enable real-world autonomy, it therefore becomes crucial to determine efficient representations of the underlying planning problem, while formulating interaction strategies capable of exploring these representations efficiently.

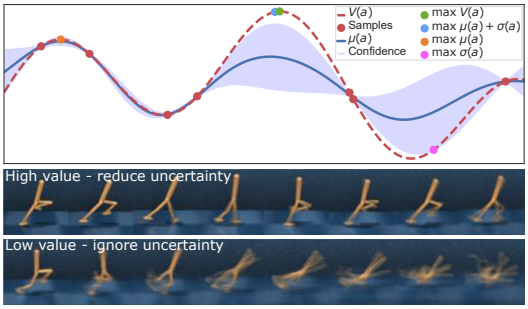

Figure 1: Top - Uncertainty over returns can guide exploration to regions with high potential for improvement (blue dot) during sampling-based value learning (red line). Bottom - uncertainty reduction should focus on expected high-reward behaviors.

In traditional controls, planning problems are commonly formulated based on the underlying state-space representation. This may inhibit efficient learning when the environment states are high-dimensional or their dynamics are susceptible to non-smooth events such as singularities and discontinuities [1, 2, 3]. It may then be desirable for the agent to abstract a latent representation that facilitates efficient learning [4, 5, 6]. The latent representation may then be leveraged either in a model-free or model-based setting. Model-free techniques estimate state-values directly from observed data to distill a policy mapping. Model-based techniques learn an explicit representation of the environment that is leveraged in generating fictitious interactions and enable policy learning in imagination [7]. While the former reduces potential sources of bias, the latter offers a structured representation encoding deeper insights into underlying environment behavior.

Submitted to the 5th Conference on Robot Learning (CoRL 2021). Do not distribute.

The agent should leverage the chosen representation to efficiently identify and explore informative interactions. For example, an engineer would ideally only provide sparse high-level feedback to a robot and the system should learn without getting distracted by environment behavior irrelevant to task completion. This requires models that can disentangle behavioral surprise from expected performance and motivate the agent to investigate the latter. Figure 1 provides a one-dimensional example of a value mapping. The true function and its samples are visualized in red with the true maximum denoted by the green dot. Relying only on the predicted mean can bias policy learning towards local optima (orange dot; Sutton and Barto [8]), while added stochasticity can waste samples on uninformative interactions. Auxiliary information-gain objectives integrate predicted uncertainty, however, uncertain environment behavior does not equate to potential for improvement (pink dot). It is desirable to focus exploration on interactions that harbor potential for improving overall performance. Combining mean performance estimates with their uncertainty into an upper confidence bound (UCB) objective provides a concise method for this (blue dot; Auer et al. [9], Krause and Ong [10]). The uncertainty can be explicitly represented by an ensemble of hypothesis on environment behavior [11, 12]. Figure 1 demonstrates this selective uncertainty reduction by providing forward predictions of a model ensemble on two motion patterns for a Walker. The expected high-reward walking behavior has been sufficiently explored and the models strongly agree, while little effort has been extended to reduce uncertainty and learn the specific details of the expected low-reward falling behavior.

This paper demonstrates that exploring interactions through imagined positive futures can yield information-dense sampling and data-efficient learning. We present latent optimistic value exploration (LOVE), an algorithm that leverages optimism in the face of uncertain long-term rewards in guiding exploration. Potential futures are imagined by an ensemble of latent variable models and their predicted infinite-horizon performance is obtained in combination with associated value function estimates. Training on a UCB objective over imagined futures yields a policy that behaves inherently optimistic and focuses the robot learner on interactions with the potential to improve performance under its current world model. This provides a concise, differentiable framework for driving deep exploration while not relying on stochasticity. LOVE therefore provides an agent with

- an uncertainty-aware model for diverse hypotheses on trajectory evolution and performance,
- intrinsic exploration without reward feedback by active querying of uncertain long-term returns,
- targeted exploration guided by improvement potential, ignoring uncertainty tangential to the task.

We present empirical results on challenging visual robot control tasks that highlight the necessity for deep exploration in scenarios with sparse rewards, where we achieve on average more than 30% improvement over state-of-the-art and an exploration baseline. We further demonstrate 15% average improvement on benchmarking environments from the DeepMind Control Suite [13]. We compare to the recent agents Dreamer [7], a curiosity-based agent inspired by Sekar et al. [14], and DrQ [15].

2 Related work

Problem representation Model-free approaches learn a policy by directly estimating performance from interaction data. While their asymptotic performance previously came at the cost of sample complexity [16, 17, 18], recent advances in representation learning through contrastive methods and data augmentation have improved their efficiency [19, 20, 15]. However, their implicit representation of the world can make generalization of learned behaviors under changing task specifications difficult. Model-based techniques leverage a structured representation of their environment that enables them to imagine potential interactions. The nature of the problem hereby dictates model complexity, ranging from linear [21, 22], over Gaussian process models [23, 24] to deep neural networks [25, 26]. In high-dimensional environments, latent variable models can provide concise representations that improve efficiency of the learning process [27, 4, 6, 7].

Planning interactions Model-based approaches leverage their representation of the world in predicting the performance of action sequences. The agent may then either solve a model predictive control-style optimization [28, 25, 29] or train a policy in simulation [30, 26]. The resulting finite-horizon formulations can be extended by value function estimates to approximate an infinite-horizon planning problem [31, 7, 32]. When considering learned models, ensembling the model predictions may further be leveraged in debiasing the actor during training [30, 25, 26, 32]. Both explicit and implicit model rollouts together with value estimation can accelerate model-free learning [33, 34, 35].

Directed exploration Directed exploration can improve over random exploration by focusing on information-dense interactions [36]. These methods are commonly driven by uncertainty estimates. Information gain techniques define an auxiliary objective that encourages exploration of unexpected environment behavior or model disagreement and have been applied in discrete [37, 38, 39] and continuous actions spaces [40, 41, 42]. When focusing on improving knowledge of the environment dynamics, effort may be wasted on uncertain interactions that are tangential to the objective. Alternatively, exploration can be driven by uncertainty over expected performance as encoded by value functions [11, 43, 44, 45, 46], multi-step imagined returns [47, 48] or their combination [31, 1, 32].

Model-ensemble agents Related work on ensemble agents has demonstrated impressive results. We note key differences to our approach. ME-TRPO [30] leverages a dynamics ensemble to debias policy optimization on finite-horizon returns under a known reward function and random exploration. MAX [49] and Amos et al. [50] explore via finite-horizon uncertainty in a dynamics ensemble. RP1 [51] explores in reward space via finite-horizon returns, but assumes access to the nominal reward function and full proprioceptive feedback. Seyde et al. [32] leverage full proprioception and embed optimism into the value function, which prohibits adjustment of the exploration trade-off during learning and limits transferability. Exploring uncertain dynamics samples interactions orthogonal to task completion and finite-horizon objectives limit exploration locally, while full-observability and access to the reward function are strong assumptions. We learn latent dynamics, reward and value functions under partial observability to explore uncertain infinite-horizon returns. This enables backpropagation through imagined trajectories to recover analytic policy gradients, while enabling guided deep exploration based on expected potential for long-term improvement.

3 Preliminaries

In the following, we first introduce the problem definition and associated nomenclature. We then provide an overview of the representation learning framework used to abstract environment behavior, which leverages the recurrent state space model (RSSM) proposed by Hafner et al. [7].

3.1 Problem formulation

We formulate the underlying optimization problem as a partially observable Markov decision process (POMDP) defined by the tuple $\{\mathcal{X}, \mathcal{A}, T, R, \Omega, \mathcal{O}, \gamma\}$, where $\mathcal{X}$, $\mathcal{A}$, $\mathcal{O}$ denote the state, action and observation space, respectively, $T\colon \mathcal{X} \times \mathcal{A} \to \mathcal{X}$ signifies the transition mapping, $R\colon \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ the reward mapping, $\Omega\colon \mathcal{X} \to \mathcal{O}$ the observation mapping, and $\gamma \in [0, 1)$ is the discount factor. We define x_t and a_t to be the state and action at time t, respectively, and use the notation $r_t = R(x_t, a_t)$. Let $\pi_\phi(a_t|o_t)$ denote a policy parameterized by ϕ and define the discounted infinite horizon return $G_t = \sum_{\tau=t}^{\infty} \gamma^{\tau-t} R(x_\tau, a_\tau)$, where $x_{t+1} \sim T(x_{t+1}|x_t, a_t)$ and $a_t \sim \pi_\phi(a_t|o_t)$. The goal is then to learn the optimal policy maximizing G_t under unknown nominal dynamics and reward mappings.

3.2 Planning from pixels

We build on the world model introduced in Hafner et al. [29] and refined in Hafner et al. [7]. High-dimensional image observations are first embedded into a low-dimensional latent space using a neural network encoder. An RSSM then serves as a probabilistic transition model and defines the model state s. Together, the encoder and RSSM define the representation model. The agent therefore abstracts observation o_t of environment state x_t into model state s_t, which is leveraged for planning. Consistency of the learned representations is enforced by minimizing the reconstruction error of a decoder network in the observation model and the ability to predict rewards of the reward model. For details, we refer the reader to Hafner et al. [7], and provide their definitions of the models as

$$\begin{aligned} &\text{Representation model:} && p_\theta(s_t|s_{t-1}, a_{t-1}, o_t) \\ &\text{Transition model:} && q_\theta(s_t|s_{t-1}, a_{t-1}) \\ &\text{Observation model:} && q_\theta(o_t|s_t) \\ &\text{Reward model:} && q_\theta(r_t|s_t), \end{aligned} \tag{1}$$

where p and q denote distributions in latent space, with θ as their joint parameterization. The action model $\pi_\phi(a_t|s_t)$ is then trained to optimize the predicted return of imagined world model rollouts.

The world model is only rolled-out over a finite horizon H, but complemented by predictions from the value model $v_\psi(s_t)$ at the terminal state s_{t+H} to approximate the infinite horizon return. The policy and value function are trained jointly using policy iteration on the objective functions

$$\max_\phi E_{q_\theta, q_\phi}\left(\sum_{\tau=t}^{t+H} V_\lambda(s_\tau)\right), \qquad \min_\psi E_{q_\theta, q_\phi}\left(\sum_{\tau=t}^{t+H} \frac{1}{2}\|v_\psi(s_\tau) - V_\lambda(s_\tau)\|^2\right), \tag{2}$$

respectively. Here, $V_\lambda(s_\tau)$ represents an exponentially recency-weighted average of the k-step value estimates $V_N^k(s_\tau)$ along the trajectory to stabilize the learning [8] as further specified in Appendix K.

4 Uncertainty-guided latent exploration

The world model introduced in Section 3.2 can be leveraged in generating fictitious interactions for the policy to train on. However, the learned model will exhibit bias in uncertain regions where insufficient samples are available. Training on imagined model rollouts then propagates simulation bias into the policy. Here, we address model bias by ensembling our belief over environment behavior. We can leverage the underlying epistemic uncertainty in formulating a UCB objective for policy learning that focuses exploration on regions with high predicted potential for improvement.

4.1 Model learning with uncertainty estimation

The model parameters are only weakly constrained in regions where interaction data is scarce and random influences strongly impact prediction performance. In order to prevent the agent from learning to exploit these model mismatches, we consider predictions from an ensemble. Individual predictions will align in regions of high data support and diverge in regions of low support. The ensemble mean then serves as a debiased estimator of environment behavior, while the epistemic uncertainty is approximated via model disagreement [12]. We consider an ensemble of M latent-space particles. Each particle is represented by a unique pairing of a transition, reward and value model to yield,

$$\text{Particles: } \{(q_{\theta_i}(s_t|s_{t-1}, a_{t-1}),\ q_{\theta_i}(r_t|s_t),\ v_{\psi_i}(s_t))\}_{i=1}^M. \tag{3}$$

The encoder and decoder remain shared to ensure that all particles leverage the same latent space, while the internal transition dynamics retain the ability of expressing distinct hypothesis over environment behavior. For particle i, we define the predicted infinite-horizon trajectory return as

$$G_{t,i}(\theta_i, \psi_i, \phi) = \sum_{\tau=t}^{t+H} V_{\lambda,i}(s_\tau), \tag{4}$$

where $V_{\lambda,i}(s_\tau)$ is computed via Eq. (8) with the particle's individual transition, reward and value models. Distinctness of the particles is encouraged by varying the initial network weights between ensemble members and shuffling the batch order during training. Predicted trajectory returns with corresponding uncertainty estimates are then obtained by considering the ensemble mean and variance,

$$\mu_G(\theta, \psi, \phi) = \frac{1}{M}\sum_{i=1}^M G_{t,i}(\theta_i, \psi_i, \phi), \tag{5}$$

$$\sigma_G^2(\theta, \psi, \phi) = \frac{1}{M}\sum_{i=1}^M (G_{t,i}(\theta_i, \psi_i, \phi) - \mu_G(\theta, \psi, \phi))^2. \tag{6}$$

4.2 Policy learning with directed exploration

The policy learning objective in Eq. (2) could be replaced by the ensemble mean in Eq. (6). This would reduce model bias in the policy, but require an auxiliary objective to ensure sufficient exploration. We consider exploration to be desirable when it reduces uncertainty over realizable task performance. The trajectory return variance in Eq. (6) encodes uncertainty over long-term performance of actions. In combination with the expected mean, we recover estimated performance bounds. During data acquisition, we explicitly leverage the epistemic uncertainty in identifying interactions with potential for improvement and define the acquisition policy objective via an upper confidence bound (UCB) as

$$G_{aq}(\theta, \psi, \phi) = \mu_G(\theta, \psi, \phi) + \beta\sigma_G(\theta, \psi, \phi), \tag{7}$$

where the scalar variable β quantifies the exploration-exploitation trade-off. For $\beta < 0$ we recover a safe-interaction objective, while $\beta > 0$ translates to an inherent optimism that uncertainty harbors potential for improvement. Here, we learn an optimistic policy $\pi_{\phi_{aq}}$ that is intrinsically capable of deep exploration and focuses interactions on regions with high information-density. Furthermore, in the absence of dense reward signals, the acquisition policy can leverage prediction uncertainty in driving exploration. This behavior is not limited to the preview window, as the value function ensemble projects long-term uncertainty into the finite-horizon model rollouts. While training in imagination, we leverage the optimistic policy to update our belief in regions that the acquisition policy intends to visit. In parallel, we train an evaluation policy $\pi_{\phi_{ev}}$ that aims to select the optimal actions under our current belief. The evaluation policy optimizes the expected mean return ($\beta = 0$).

4.3 Latent optimistic value exploration (LOVE)

In the following, we provide a high-level description of the proposed algorithm, LOVE, together with implementation details of the overall training process and the associated pseudo-code in Algorithm 1.

Summary LOVE leverages an ensemble of latent variable models in combination with value function estimates to predict infinite-horizon trajectory performance and associated uncertainty. The acquisition policy is trained on a UCB objective to imagine positive futures and focus exploration on interactions with high predicted potential for long-term improvement. The ensemble members are constrained to operate over the same latent space to encourage learning of abstract representations conducive to the objective, while ensuring prediction consistency.

Algorithm 1: LOVE

Initialize : random parameters $\{\theta_i, \psi_i, \phi_{aq}, \phi_{ev}\}$
for *episode in episodes* **do**
  for *timestep* $t = 1$ ***to*** T **do**
    $a_t \sim \pi_\phi(\cdot|s_t)$, $s_t \sim p_{\theta_i}(\cdot|s_{t-1}, a_{t-1}, o_t)$
    Observe environment transition and add to $\mathcal{D}$
  for *trainstep* $s = 1$ ***to*** S **do**
    for *particle* $i = 1$ ***to*** M **do**
      Sample sequence batch $\{(o_t, a_t, r_t)\}_{t=b}^{b+L} \sim D$
      Compute states $s_t \sim p_{\theta_i}(s_t|s_{t-1}, a_{t-1}, o_t)$
      Estimate values $V_{\lambda,i}(s_\tau) \leftarrow$ **rollout**(s_t, i)
      Representation learning of θ_i on $\{(r_t, o_{t+1})\}_{t=b}^{b+L}$
      Regression of ψ_i on $V_{\lambda,i}(s_\tau)$ targets in (2)
    Sample sequences $\{(o_t, a_t, r_t)\}_{t=b}^{b+L} \sim D$
    for *particle* $i = 1$ ***to*** M **do**
      Compute states $s_t \sim p_{\theta_i}(s_t|s_{t-1}, a_{t-1}, o_t)$
      Estimate values $V_{\lambda,i}(s_\tau) \leftarrow$ **rollout**(s_t, i)
    Compute ensemble statistics μ_G, σ_G via (6)
    Update ϕ by optimizing UCB objective in (7)

Implementation The algorithm proceeds in two alternating phases. In the online phase, the agent leverages its acquisition policy to explore interactions optimistically and resulting transitions are appended to memory $\mathcal{D}$. In the offline phase, the agent first updates its belief about environment behavior and then adjusts its policy accordingly. The representation learning step extends the procedure introduced in Hafner et al. [7] to ensemble learning and groups each model with a unique value function estimator into a particle. The batch order is varied between particles during training to ensure variability in the gradient updates and to prevent collapse into a single mode. The policy learning step combines particle predictions to generate value targets in Eq. (4) by simulating ensemble rollouts from the same initial conditions. The trajectory return statistics of Eq. (6) are combined into the UCB objective of Eq. (7), which the acquisition policy optimizes. We provide an overview in Algorithm 1, where we consider the acquisition policy $\pi_\phi = \pi_{\phi_{aq}}$ and evaluation policy training has been omitted for brevity.

5 Experiments

We evaluate LOVE on continuous visual control tasks. First, we showcase intrinsic exploration in the absence of reward signals. Then, we illustrate interpretable uncertainty estimation of long-term returns with sparse reward signals. Finally, we benchmark performance on tasks from the DeepMind Control Suite [13]. We use a single set of hyperparameters throughout, as detailed in Appendix A.

5.1 Intrinsic exploration without reward feedback

In the absence of reward feedback an agent relies on intrinsic motivation to explore. LOVE enables intrinsic exploration by considering uncertainty over infinite-horizon returns. Without informative mean performance estimates, optimism acts as a proxy to guide exploration (see Eq. 7).

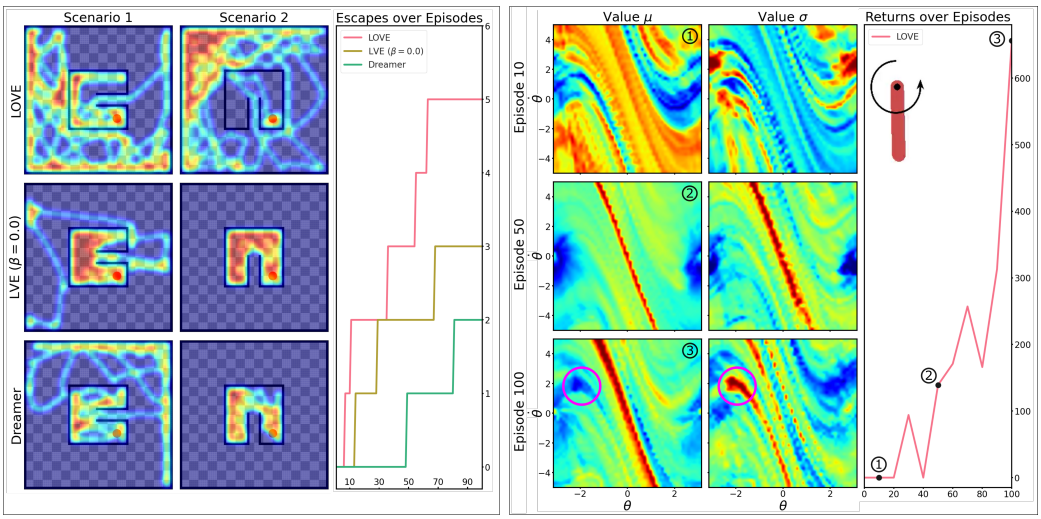

Figure 2: **Left** - the agent starts inside a bug trap (red dot) and does not receive reward feedback. We provide occupancy maps for LOVE, LVE (no optimism), and Dreamer with escape counts on 3 seeds. LOVE's ability to consider uncertain long-term performance enables consistent exploration in the absence of reward signals yielding the highest escape rate and area coverage. **Right** - Pendulum initialized at the bottom with sparse rewards at the top. Predicted ensemble mean (left) and standard deviation (right) of the value function. The agent first learns to quantify model uncertainty (Ep. 10), and then discovers high reward potential at the top (Ep. 50). Once the agent learned to reach the top, remaining uncertainty with low potential for improvement is not further explored (Ep. 100, circle).

We consider a reward-free bug trap environment. The agent starts inside the trap and can exit through a narrow passage. There is no reward feedback to guide planning and escapes need to arise implicitly through exploration. The agent is represented as a pointmass under continuous actuation-limited acceleration control and observes a top-down view of the environment. The ratio of agent diameter to passage width is 0.9 and collisions are modelled as inelastic with the coefficient of restitution set to 0.3. The relative size constrains the set of policies allowing for entry into the passage, while inelastic collisions reduce random motions.

We compare performance of LOVE to both Dreamer and LVE, a variation that ablates on LOVE's optimism ($\beta = 0$). Two variations of the environment are run on 3 random seeds for 100k timesteps. Exemplary occupancy maps are provided in Figure 2 (left). We note that LOVE escapes the bug trap in both scenarios to explore a much larger fraction of the state space than either LVE or Dreamer. This can be attributed to LOVE's ability to envision uncertain rewards beyond the preview horizon, which optimistically drives deep exploration in the absence of reward signals (row 1). Removing long-term optimism by only guiding interactions through mean performance estimates can lead to prediction collapse and the agent assuming the absence of reward without explicitly querying the environment for confirmation (row 2). We observe similar behavior for the Dreamer agent, which employs a single latent model and leverages random noise for exploration (row 3). The total escape count confirms this insight (column 3), where the remaining occupancy maps are provided in Appendix E.

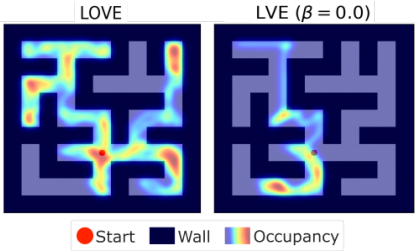

Figure 3: Reward-free maze exploration.

We further highlight the differences in intrinsic exploration between LOVE and LVE on a reward-free maze exploration task. The environment follows the same physics as the bug trap and the ratio of agent diameter to corridor width is 0.33. The occupancy maps in Figure 3 exemplify how optimism over returns can set its own exploration goals in domains without reward feedback. These observations further underline the potential of combining model ensembling for policy debiasing with optimism over infinite-horizon returns in guiding deep exploration.

5.2 Targeted exploration with sparse rewards

Efficient planning under sparse reward feedback requires long-term coordination. LOVE focuses on interactions with high potential for improving infinite-horizon performance by trading-off predicted returns with their uncertainty. The agent then only reduces uncertainty when it aligns with successful task completion. We consider a sparse reward pendulum task, as its functional representations are readily interpretable. The pendulum is always initialized in a downward configuration and only receives reward at the top. The agent can therefore not rely on random initialization to observe initial reward signals and needs to actively explore the upright configuration.

Figure 2 (right) visualizes the mean and variance of the value ensemble during training. We generate predictions in state-space by providing 5 contextual images for velocity estimation. Initially, the agent builds a world model within which it learns to quantify uncertainty (Ep. 10). The agent then leverages the model to explore the top. This yields a spinning behavior and reward propagates into the value mean, while uncertainty is allocated for trajectories that could yield stabilization at the top (Ep. 50). The agent queries these promising motions and refines its behavior to reach the upright, reducing associated uncertainty (Ep. 100). The UCB objective allows the agent to ignore remaining uncertainty tangential to successful task completion (circle). LOVE therefore guides exploration via potential for long-term improvement while offering interpretable uncertainty estimates.

5.3 Performance with mixed sparse and dense rewards

The previous sections highlighted LOVE's intrinsic motivation to explore potential for long-term improvement based on its uncertainty estimation over returns. In the following, we compare performance on a variety of visual control tasks from the DeepMind Control Suite and sparsified variations.

The environments feature $(64, 64)$ image observations, continuous actions, and dense or sparse rewards. The sparse Cheetah Run, Walker Run and Walker Walk tasks are generated by zeroing rewards below a threshold (0.25, 0.25 and 0.7, respectively; rescaled to [0, 1]). We use the same set of parameters throughout all experiments mirroring those of Dreamer. However, we do not use random exploration noise. We furthermore use an ensemble of $M = 5$ latent models, an initial UCB trade-off of $\beta = 0.0$, an episodic growth of $\delta = 10^{-3}$, and alter the policy learning rate and training steps to account for the more complex UCB objective. The changes to the policy learning rate and training steps are also applied to Dreamer to yield ΔDreamer, as we found this to improve performance (see Appendix G). We furthermore introduce an exploration baseline, ΔDreamer+Curious, inspired by the curiosity bonus in Sekar et al. [14] over an ensemble of RSSM dynamics.

Table 1 provides benchmarking performance on 5 seeds over 300 episodes (see also Appendix F). On the sparsified environments, LOVE achieves 40% average improvement over the state-of-the-art ΔDreamer agent and 30% over both the exploration baseline ΔDreamer+Curious and ensembling baseline LVE. On the other environments, LOVE improves by 26%, 37%, and 9%, respectively. This suggests that LOVE's combination of latent model ensembling with directed exploration can aid in identifying interactions conducive to task completion. Ensembling reduces propagation of model-specific biases into the policy. Optimistic exploration of uncertain long-term returns focuses sampling on regions with promising performance estimates while ignoring uncertain regions that are tangential to task completion. This is an advantage of formulating the exploration objective in terms of uncertain rewards and not uncertain dynamics, as shown by the comparison to ΔDreamer+Curious.

We ablate the performance of LOVE on the UCB trade-off parameter β and compare against LVE ($\beta = 0$). We note that while LOVE outperforms LVE on the majority of tasks, both reach similar terminal performance in several instances. However, LOVE provides clear improvements on the fully-sparse Cartpole Swingup and the partially-sparse Hopper Hop tasks, as well as the sparse locomotion tasks (Table 1, top). These environments initialize the agent in configurations that provide no reward feedback, thereby forcing the agent to actively explore. LOVE leverages uncertainty-guided exploration and gains an advantage under these conditions. Similarly, this can explain performance on the Pendulum Swingup task. While the task only provides sparse reward feedback, random environment initializations offer sufficient visitation to the non-zero reward states, removing the need for active exploration. LOVE also improves performance on the dense Walker tasks. These environments require learning of stable locomotion patterns under complex dynamics, where directed exploration efficiently identifies tweaks to the gait. Similar to the bug trap environment of Section 5.1,

300k steps	LOVE	LVE	ΔDreamer	+Curious	DrQ	A3C	D4PG
Cheetah Sparse (R)	$614_{\pm 262}$	$606_{\pm 117}$	$259_{\pm 280}$	$514_{\pm 302}$	–	–	–
Walker Sparse (R)	$234_{\pm 129}$	$164_{\pm 151}$	$78_{\pm 60}$	$106_{\pm 120}$	–	–	–
Walker Sparse (W)	$935_{\pm 24}$	$864_{\pm 135}$	$584_{\pm 315}$	$639_{\pm 277}$	–	–	–
Hopper Sparse (S)	$831_{\pm 72}$	$778_{\pm 142}$	$649_{\pm 212}$	$530_{\pm 246}$	–	–	–
Avg. Diff. to LOVE	+0%	-11%	-46%	-35%	–	–	–
Cartpole Dense	$688_{\pm 124}$	$666_{\pm 78}$	$512_{\pm 253}$	$539_{\pm 101}$	$781_{\pm 100}$	$558_{\pm 7}$	$862_{\pm 1}$
Cartpole Sparse	$631_{\pm 259}$	$192_{\pm 132}$	$388_{\pm 245}$	$219_{\pm 199}$	$231_{\pm 337}$	$180_{\pm 6}$	$482_{\pm 57}$
Cheetah Run	$771_{\pm 118}$	$753_{\pm 120}$	$655_{\pm 93}$	$760_{\pm 73}$	$533_{\pm 143}$	$214_{\pm 2}$	$524_{\pm 7}$
Finger Spin	$584_{\pm 302}$	$605_{\pm 306}$	$348_{\pm 156}$	$380_{\pm 173}$	$898_{\pm 131}$	$129_{\pm 2}$	$986_{\pm 1}$
Hopper Hop	$203_{\pm 104}$	$165_{\pm 96}$	$118_{\pm 95}$	$83_{\pm 87}$	$151_{\pm 51}$	$1_{\pm 0}$	$242_{\pm 2}$
Pendulum	$641_{\pm 298}$	$639_{\pm 332}$	$320_{\pm 353}$	$177_{\pm 208}$	$399_{\pm 298}$	$49_{\pm 5}$	$681_{\pm 42}$
Walker Run	$528_{\pm 123}$	$506_{\pm 67}$	$438_{\pm 85}$	$451_{\pm 119}$	$338_{\pm 82}$	$192_{\pm 2}$	$567_{\pm 19}$
Walker Walk	$947_{\pm 36}$	$914_{\pm 52}$	$902_{\pm 94}$	$873_{\pm 122}$	$815_{\pm 184}$	$311_{\pm 2}$	$968_{\pm 2}$
Avg. Diff. to LOVE	+0%	-12%	-29%	-35%	-18%	-70%	+9%

Table 1: DeepMind Control Suite. Mean and standard deviation on 9 seeds after 3×10^5 timesteps. LOVE improves sample efficiency particularly under sparse reward feedback. Temporally-extended optimism helps LOVE in actively exploring uncertain returns, providing an advantage over LVE. Formulating intrinsic motivation in reward-space enables LOVE to identify uncertain interactions conducive to solving the task, providing an advantage over the curiosity baseline ΔDreamer+Curious. *D4PG, A3C: converged results at 10^8 timesteps for reference.

we observe that optimistic exploration is especially favoured by objectives that provide sparse reward feedback, while enabling efficient discovery of tweaks to motion patterns under complex dynamics.

We additionally compare to Data Regularized Q (DrQ), a concurrent model-free approach. DrQ updates its actor-critic models online giving it an advantage over LOVE and Dreamer. LOVE performs favourably on the majority of environments with significant differences on the sparse tasks (Pendulum, Cartpole Sparse) and the locomotion tasks (Hopper, Walker). DrQ outperforms LOVE on Finger Spin and Cartpole with dense rewards. On these tasks, learning an explicit world model may actually be detrimental to attaining performance quickly. The former task features high-frequency behaviors that may induce aliasing, while the latter task allows the agent to leave the frame yielding transitions with no visual feedback. We additionally provide converged performance results for pixel-based D4PG [52] and proprioception-based A3C [53] at 10^8 environment steps to put the results into perspective.

6 Conclusion

We propose Latent Optimistic Value Exploration (LOVE), a model-based reinforcement learning algorithm that leverages long-term optimism to guide exploration for continuous visual control. LOVE leverages finite-horizon rollouts of a latent model ensemble in combination with value function estimates to predict long-term performance of candidate action sequences. The ensemble predictions are then combined into an upper confidence bound objective for policy learning. Training on this objective yields a policy that optimistically explores interactions that have the potential of improving task performance while ignoring uncertain interactions tangential to task completion. These aspects are particularly important for robot control in complex environments, where the engineer aims to provide sparse high-level feedback without intensive reward engineering and the system should efficiently learn without getting distracted by irrelevant environment behavior. To this end, we evaluate LOVE regarding its exploration capabilities and performance on a variety of tasks. In the absence of reward signals, LOVE demonstrates an intrinsic motivation to explore interactions based on their information density. Empirical results on DeepMind Control Suite tasks showcase LOVE's competitive performance and ability to focus exploration on interactions conducive to task completion, particularly when reward signals are sparse. Lifting intrinsic motivation into reward-space is then preferred over dynamics-space, as shown by comparison to a curiosity baseline. Furthermore, LOVE demonstrates improved sample efficiency over the recent model-based Dreamer agent and competitiveness with the model-free DrQ agent.

References

[1] J. Schrittwieser, I. Antonoglou, T. Hubert, K. Simonyan, L. Sifre, S. Schmitt, A. Guez, E. Lockhart, D. Hassabis, T. Graepel, et al. Mastering atari, go, chess and shogi by planning with a learned model. *arXiv preprint arXiv:1911.08265*, 2019.

[2] J. Hwangbo, J. Lee, A. Dosovitskiy, D. Bellicoso, V. Tsounis, V. Koltun, and M. Hutter. Learning agile and dynamic motor skills for legged robots. *Science Robotics*, 4(26), 2019.

[3] Y. Yang, K. Caluwaerts, A. Iscen, T. Zhang, J. Tan, and V. Sindhwani. Data efficient reinforcement learning for legged robots. *arXiv preprint arXiv:1907.03613*, 2019.

[4] D. Ha and J. Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

[5] M. Zhang, S. Vikram, L. Smith, P. Abbeel, M. Johnson, and S. Levine. Solar: Deep structured representations for model-based reinforcement learning. In *International Conference on Machine Learning*, pages 7444–7453, 2019.

[6] A. X. Lee, A. Nagabandi, P. Abbeel, and S. Levine. Stochastic latent actor-critic: Deep reinforcement learning with a latent variable model. *arXiv preprint arXiv:1907.00953*, 2019.

[7] D. Hafner, T. Lillicrap, J. Ba, and M. Norouzi. Dream to control: Learning behaviors by latent imagination. *arXiv preprint arXiv:1912.01603*, 2019.

[8] R. S. Sutton and A. G. Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[9] P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine learning*, 47(2-3):235–256, 2002.

[10] A. Krause and C. S. Ong. Contextual gaussian process bandit optimization. In *Advances in neural information processing systems*, pages 2447–2455, 2011.

[11] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy. Deep exploration via bootstrapped DQN. In *Advances in neural information processing systems*, pages 4026–4034, 2016.

[12] B. Lakshminarayanan, A. Pritzel, and C. Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems*, pages 6402–6413, 2017.

[13] Y. Tassa, Y. Doron, A. Muldal, T. Erez, Y. Li, D. de Las Casas, D. Budden, A. Abdolmaleki, J. Merel, A. Lefrancq, T. Lillicrap, and M. Riedmiller. DeepMind Control Suite. Technical report, DeepMind, 2018.

[14] R. Sekar, O. Rybkin, K. Daniilidis, P. Abbeel, D. Hafner, and D. Pathak. Planning to explore via self-supervised world models. *arXiv preprint arXiv:2005.05960*, 2020.

[15] I. Kostrikov, D. Yarats, and R. Fergus. Image augmentation is all you need: Regularizing deep reinforcement learning from pixels. *arXiv preprint arXiv:2004.13649*, 2020.

[16] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra. Continuous control with deep reinforcement learning. *arXiv preprint arXiv:1509.02971*, 2015.

[17] S. Fujimoto, H. Hoof, and D. Meger. Addressing Function Approximation Error in Actor-Critic Methods. In *International Conference on Machine Learning*, pages 1582–1591, 2018.

[18] T. Haarnoja, A. Zhou, P. Abbeel, and S. 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.

[19] A. Srinivas, M. Laskin, and P. Abbeel. Curl: Contrastive unsupervised representations for reinforcement learning. *arXiv preprint arXiv:2004.04136*, 2020.

[20] M. Laskin, K. Lee, A. Stooke, L. Pinto, P. Abbeel, and A. Srinivas. Reinforcement learning with augmented data. *arXiv preprint arXiv:2004.14990*, 2020.

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

[22] V. Kumar, E. Todorov, and S. 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.

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

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

[25] K. Chua, R. Calandra, R. McAllister, and S. Levine. Deep reinforcement learning in a handful of trials using probabilistic dynamics models. In *Advances in Neural Information Processing Systems*, pages 4754–4765, 2018.

[26] I. Clavera, J. Rothfuss, J. Schulman, Y. Fujita, T. Asfour, and P. Abbeel. Model-based reinforcement learning via meta-policy optimization. In *Conference on Robot Learning*, pages 617–629, 2018.

[27] M. Watter, J. Springenberg, J. Boedecker, and M. Riedmiller. Embed to control: A locally linear latent dynamics model for control from raw images. In *Advances in neural information processing systems*, pages 2746–2754, 2015.

[28] A. Nagabandi, G. Kahn, R. S. Fearing, and S. 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.

[29] D. Hafner, T. Lillicrap, I. Fischer, R. Villegas, D. Ha, H. Lee, and J. Davidson. Learning Latent Dynamics for Planning from Pixels. In *International Conference on Machine Learning*, pages 2555–2565, 2019.

[30] T. Kurutach, I. Clavera, Y. Duan, A. Tamar, and P. Abbeel. Model-ensemble trust-region policy optimization. In *International Conference on Learning Representations*, 2018.

[31] K. Lowrey, A. Rajeswaran, S. Kakade, E. Todorov, and I. Mordatch. Plan Online, Learn Offline: Efficient Learning and Exploration via Model-Based Control. In *International Conference on Learning Representations*, 2019.

[32] T. Seyde, W. Schwarting, S. Karaman, and D. Rus. Learning to plan via deep optimistic value exploration. In *Proceedings of the 2nd Annual Conference on Learning for Dynamics and Control*, 2020.

[33] J. Oh, S. Singh, and H. Lee. Value prediction network. In *Advances in Neural Information Processing Systems*, pages 6118–6128, 2017.

[34] V. Feinberg, A. Wan, I. Stoica, M. I. Jordan, J. E. Gonzalez, and S. Levine. Model-based value estimation for efficient model-free reinforcement learning. *arXiv preprint arXiv:1803.00101*, 2018.

[35] J. Buckman, D. Hafner, G. Tucker, E. Brevdo, and H. Lee. Sample-efficient reinforcement learning with stochastic ensemble value expansion. In *Advances in Neural Information Processing Systems*, pages 8224–8234, 2018.

[36] J. Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2(3):230–247, 2010.

[37] B. C. Stadie, S. Levine, and P. Abbeel. Incentivizing exploration in reinforcement learning with deep predictive models. *arXiv preprint arXiv:1507.00814*, 2015.

[38] G. Ostrovski, M. G. Bellemare, A. van den Oord, and R. Munos. Count-based exploration with neural density models. In *International Conference on Machine Learning*, pages 2721–2730, 2017.

[39] D. Pathak, P. Agrawal, A. A. Efros, and T. Darrell. Curiosity-driven exploration by self-supervised prediction. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 16–17, 2017.

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

[41] R. Houthooft, X. Chen, Y. Duan, J. Schulman, F. De Turck, and P. Abbeel. Vime: Variational information maximizing exploration. In *Advances in Neural Information Processing Systems*, pages 1109–1117, 2016.

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

[43] I. Osband, B. Van Roy, D. Russo, and Z. Wen. Deep exploration via randomized value functions. *Journal of Machine Learning Research*, 2017.

[44] R. Y. Chen, S. Sidor, P. Abbeel, and J. Schulman. UCB exploration via Q-ensembles. *arXiv preprint arXiv:1706.01502*, 2017.

[45] B. O'Donoghue, I. Osband, R. Munos, and V. Mnih. The Uncertainty Bellman Equation and Exploration. In *International Conference on Machine Learning*, pages 3836–3845, 2018.

[46] K. Lee, M. Laskin, A. Srinivas, and P. Abbeel. Sunrise: A simple unified framework for ensemble learning in deep reinforcement learning. *arXiv preprint arXiv:2007.04938*, 2020.

[47] S. Depeweg, J.-M. Hernandez-Lobato, F. Doshi-Velez, and S. Udluft. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In *International Conference on Machine Learning*, pages 1184–1193, 2018.

[48] M. Henaff, Y. LeCun, and A. Canziani. Model-predictive policy learning with uncertainty regularization for driving in dense traffic. In *7th International Conference on Learning Representations, ICLR 2019*, 2019.

[49] P. Shyam, W. Jaśkowski, and F. Gomez. Model-based active exploration. In *International Conference on Machine Learning*, pages 5779–5788. PMLR, 2019.

[50] B. Amos, L. Dinh, S. Cabi, T. Rothörl, S. G. Colmenarejo, A. Muldal, T. Erez, Y. Tassa, N. de Freitas, and M. Denil. Learning awareness models. *arXiv preprint arXiv:1804.06318*, 2018.

[51] P. Ball, J. Parker-Holder, A. Pacchiano, K. Choromanski, and S. Roberts. Ready policy one: World building through active learning. *arXiv preprint arXiv:2002.02693*, 2020.

[52] G. Barth-Maron, M. W. Hoffman, D. Budden, W. Dabney, D. Horgan, D. Tb, A. Muldal, N. Heess, and T. Lillicrap. Distributed distributional deterministic policy gradients. *arXiv preprint arXiv:1804.08617*, 2018.

[53] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937, 2016.

[54] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.