Disentangling Epistemic and Aleatoric Uncertainty in Reinforcement Learning

Bertrand Charpentier[1], Ransalu Senanayake[2], Mykel Kochenderfer[2], Stephan Günnemann[1]

Technical University of Munich[1], Stanford University[2]

October 25, 2024

Introduction

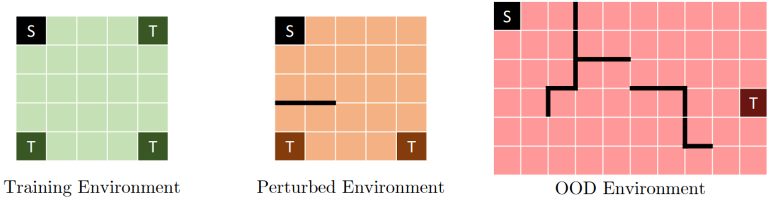


Figure: Reinforcement learning with different environments

Introduction

Reinforcement learning (RL) agents should have:

Three practically desirable properties:

- Learn fast with few episode failures
- Maintain high reward when facing similar environments
- Flag anomalous environment states

Three technical properties:

- High sample efficiency at training time
- High generalization performance on similar environment
- High Out-Of-Distribution (OOD) detection score on unknown environment

Introduction

Key concepts to achieve desired properties:

- **Aleatoric uncertainty:**
  The irreducible and inherent stochasticity of the environment.
- **Epistemic uncertainty:**
  The lack of information for accurate prediction.

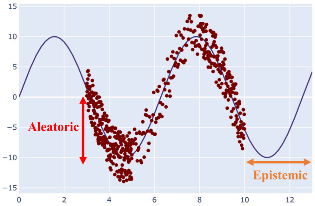


Figure: (M. Abdar et al., 2021)

Study outline

Core Motivation

Disentangle the properties of aleatoric and epistemic uncertainty estimates in RL to build agents with reliable performance in real-world applications.

- **Desiderata:** Define 4 desiderata covering aleatoric and epistemic uncertainty estimates w.r.t. sample efficiency at training time and generalization performance at testing time.
- **Models:** Combine uncertainty estimation to RL agency.
- **Evaluation:** Propose practical evaluation methods based on OOD environment and domain shifts.

Uncertainty in SL

For predicting the output $y^{(i)}$ given an input $\mathbf{x}^{(i)}$,

- **aleatoric uncertainty:**
  $u_{alea}(\mathbf{x}^{(i)}) = \mathbb{H}(\mathbb{P}(y^{(i)}|\boldsymbol{\theta}^{(i)}))$
- **epistemic uncertainty:**
  $u_{epist}(\mathbf{x}^{(i)}) = \mathbb{H}(\mathbb{Q}(\boldsymbol{\theta}^{(i)}|\mathscr{X}^{(i)}))$

To estimate these uncertainty, we have

- **sampling-based methods**: (MC dropout, Ensemble)
  Aggregating statistics from different samples to *implicitly* describe $\mathbb{Q}(\boldsymbol{\theta}^{(i)}|\mathscr{X}^{(i)})$.
- **sampling-free methods**: (DKL, Evidential networks)
  Explicitly parametrizing $\mathbb{Q}(\boldsymbol{\theta}^{(i)}|\mathscr{X}^{(i)})$ with known distribution (e.g., Normal, NIG).

Uncertainty in RL

Learning RL policies with environment at every time step t

- action $a^{(t)}$, state $s^{(t)}$
- reward $r(s^{(t)}, a^{(t)})$
- transition probability $T(s^{(t+1)}|s^{(t)}, a^{(t)})$

Learning Goal

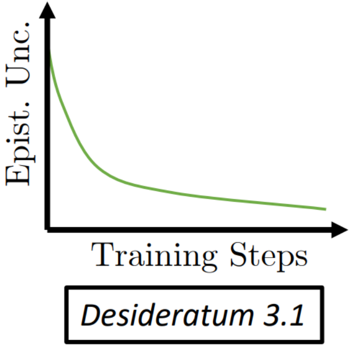
Learn a policy π predicting $a^{(t)}$ leading to the highest reward $y^{(t)} = r(s^{(t)}, a^{(t)})$ given the current state $s^{(t)}$, **in addition to u_{alea} and u_{epist} on the predicted reward**.

Action selection strategies with uncertainty

- **epsilon-greedy strategy**
  $a^{(t)} = max_a y^{(t)}$ with $(1-\varepsilon)$ probability
- **sampling-aleatoric strategy**
  $a^{(t)} = max_a y^{(t)}$, where $y^{(t)} \sim \mathbb{P}(y^{(t)}|\theta^{(t)})$
- **sampling-epistemic strategy**
  $a^{(t)} = max_a \mathbb{E}_{\mathbb{P}(y^{(t)}|\theta^{(t)})}[y^{(t)}]$, where $\theta^{(t)} \sim \mathbb{Q}(\theta^{(t)}|\mathscr{X}^{(t)})$

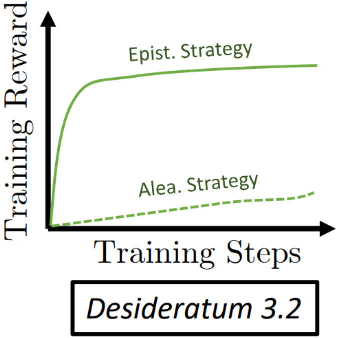
Training time

Desiderata 1. *An agent training longer on states sampled from one specific environment should become more* ***epistemically confident*** *when predicting actions on states sampled from the* ***same specific environment****.*



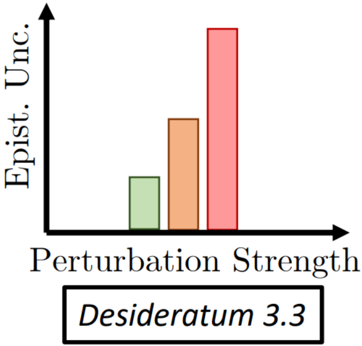
Training time

Desiderata 2. *All else being equal, an agent selecting actions with the* ***sampling-aleatoric strategy*** *at training time should achieve* ***lower sample efficiency*** *than an agent selecting actions with the sampling-epistemic strategy.*



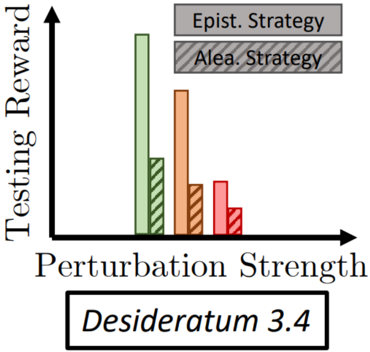
Testing time

Desiderata 3. *At testing time, **epistemic uncertainty** should be greater in environments that are very **different** from the original training environments.*



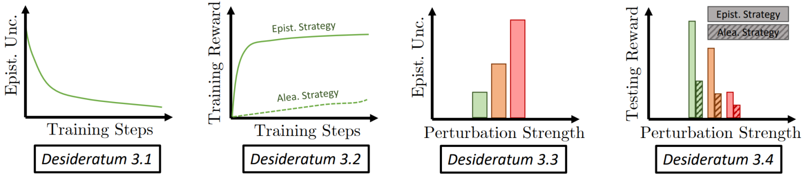
Testing time

Desiderata 4. *All else being equal, an agent* ***sampling actions from the epistemic uncertainty*** *at training and testing time should* ***generalize*** *better at testing time than an agent sampling actions from the aleatoric uncertainty.*



Desiderata

- Exploration-exploitation trade-off. (**Des. 2**)
- Trade-off between high uncertainty and generalizing to new test environments. (**Des. 3** vs. **Des. 4**)



Deep Q-Networks (DQN)

Model-free RL agent π:

Optimal Bellman equation

$$Q^{\pi*}(\mathbf{s}^{(t)}, a^{(t)}) = r(\mathbf{s}^{(t)}, a^{(t)}) + \gamma \mathbb{E}_T[\max_{a^{(t+1)}} Q^{\pi*}(\mathbf{s}^{(t+1)}, a^{(t+1)})]$$

Deep RL agents:

Minimize temporal difference error

$$\|r(\mathbf{s}^{(t)}, a^{(t)}) + \gamma \max_{a^{(t+1)}} f_{\theta'}(\mathbf{s}^{(t+1)}, a^{(t+1)}) - f_{\theta}(\mathbf{s}^{(t)}, a^{(t)})\|_2$$

MC Dropout & Ensemble

(1) K independent set of parameters

- Dropout: Dropping activations
- Ensemble: Train models with different parameters

(2) K forward passes: $\mu_k, \sigma_k = f_{\theta_k}(s^{(t)}, a^{(t)})$

(3) Aggregate predictions

- Mean prediction:
  $\mu(s^{(t)}, a^{(t)}) = \frac{1}{K}\sum_{k=1}^{K} \mu_k$
- Aleatoric uncertainty estimate:
  $u_{\text{alea}}(s^{(t)}, a^{(t)}) = \frac{1}{K}\sum_{k=1}^{K} \sigma_s$
- epistemic uncertainty estimate:
  $u_{\text{epist}}(s_t, a_t) = \frac{1}{K}\sum_{k=1}^{K} (\mu_k - \mu(s^{(t)}, a^{(t)}))^2$

MC Dropout & Ensemble

Limitations

- May not concentrate with more observed data (violating **Des. 1**).
- No guarantee to produce meaningful uncertainty estimates for extreme input states with a finite number of samples K (violating **Des. 2**).
- Computationally expensive for large K values.

Deep Kernel Learning (DKL)

(1) Latent presentation of each input state: $z^{(t)} = f_\theta(s^{(t)})$
(2) Predict Normal distribution of $\mu(s^{(t)}, a)$ and $\sigma(s^{(t)}, a)$

- K inducing points $\{\phi_{a,k}\}_{k=1}^{K}$
- predefined positive definite kernel $\kappa(\cdot,\cdot)$
- Gaussian process

(3) Epistemic uncertainty estimate:

- $u_{\text{epist}}(s_t, a_t) = \mathbb{H}(\mathcal{N}(\mu(s^{(t)}, a^{(t)}), \sigma(s^{(t)}, a^{(t)})))$

Limitation

Does not disentangle aleatoric and epistemic uncertainty.

Evidential Networks

(1) Latent presentation of each input state: $z^{(t)} = f_\theta(s^{(t)})$

(2) Predict Normal Inverse-Gamma distribution of $\mathbb{Q}(\mathscr{X}(s^{(t)}, a), n(s^{(t)}, a))$

- $\mathscr{X}(s^{(t)}, a) = g_{\psi_a}(z^{(t)})$, g_{ψ_a} is linear decoder.
- $n(s^{(t)}, a) \propto \mathbb{P}(z^{(t)} | \omega_a)$, $\mathbb{P}(.|\omega_a)$ is density estimator.
- $\mathscr{X}^{\text{post}}(s^{(t)}, a) = \frac{n^{\text{prior}} \mathscr{X}^{\text{prior}} + n(s^{(t)}, a)) \mathscr{X}(s^{(t)}, a)}{n^{\text{prior}} + n(s^{(t)}, a))}$
- $n^{\text{post}}(s^{(t)}, a)) = n^{\text{prior}} + n(s^{(t)}, a)$

(3) Epistemic uncertainty estimate:

- $u_{\text{epist}}(s_t, a_t) = \mathbb{H}(\mathscr{N}\Gamma^{-1}(\mathscr{X}(s^{(t)}, a^{(t)}), n(s^{(t)}, a^{(t)})))$

Aleatoric uncertainty estimate:

- $u_{\text{alea}}(s_t, a_t) = \mathbb{H}(\mathscr{N}(\mu(s^{(t)}, a^{(t)}), \sigma(s^{(t)}, a^{(t)})))$

Uncertainty models

Table 1: Summary of the uncertainty properties of the models.

	DropOut	Ensemble	Deep Kernel Learning	Evidential Networks
Uncertainty concentration (Des. 3.1)	✗	✗	✗	✓
Alea. vs epist. sampling at training time (Des. 3.2)	✓	✓	✗	✓
OOD detection (Des. 3.3)	✗	✗	✓	✓
Alea. vs epist. sampling at testing time (Des. 3.2)	✓	✓	✗	✓

Figure: Summary of the uncertainty properties of the models

Environments

Training environments:
CartPole, Acrobot, LunarLander

OOD environments

Input state is composed of Gaussian noise at every time step.

Perturbed environments

Separately perturb the state space, the action space, and the transition dynamics with different strengths of Gaussian noise.

Training time

Desiderata 1

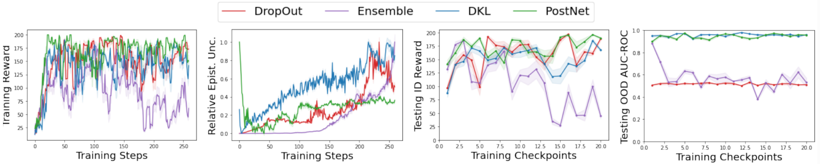


Figure: Comparison of the training performance (a, b) and testing performance (c, d) using epsilon-greedy strategies on CartPole.

Training time

Desiderata 2

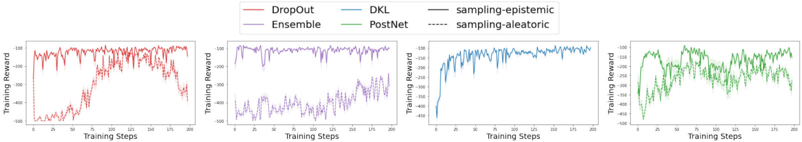


Figure: Comparison of the training performance using sampling-aleatoric or sampling epistemic at training time on Acrobot.

Testing time

Desiderata 3

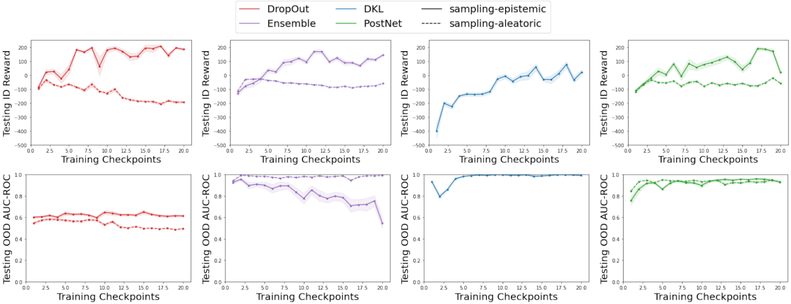


Figure: Comparison of the testing reward and OOD performance using sampling-aleatoric or sampling epistemic at both training and testing time on LunarLander.

Testing time

Desiderata 4

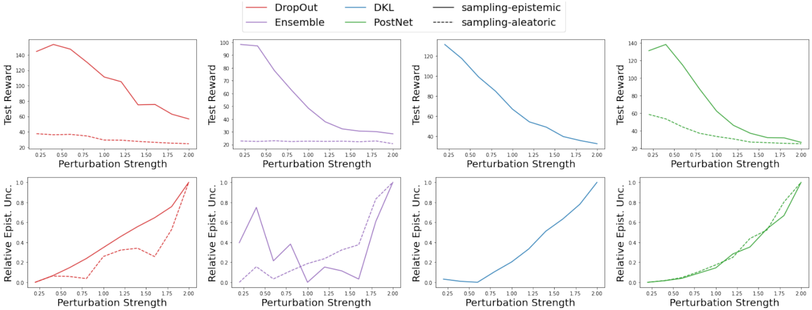


Figure: Comparison of the testing performance and the epistemic uncertainty predictions on CartPole with perturbed states, using the epsilon-greedy strategy at training time and the sampling-aleatoric or sampling-epistemic strategy at testing time.

Conclusion

- Introduce a new framework to characterize aleatoric and epistemic uncertainty estimation in RL.
- Explicitly define four desiderata of uncertainty estimates during both training and testing.
- Integrate DQN with sampling-based and sampling-free uncertainty methods.
- Give theoretical and empirical evidence that these methods can fulfill the desiderata.
- Evaluate on sample efficiency, generalization and OOD detection tasks.

Conclusion

Limitations

- Desiderata should be instantiated with formal definitions in practice.
- Potential to adapt uncertainty methods to other model-free RL methods.

Recommendations

- **Worth reading?** Yes. Well structured. Covers essential concepts in RL and uncertainty estimation.
- **Worth implementing?** Yes. The paper show both theoretically and empirically for the benefit of disentangling epistemic and aleatoric uncertainty.