Information-Directed Exploration for Deep Reinforcement Learning

Nikolay Nikolov [1,2] Johannes Kirschner [1] Felix Berkenkamp [1] Andreas Krause [1]

[1]ETH Zurich [2]Imperial College London

Introduction

Problem

- Popular exploration strategies, e.g., UCB/Thompson Sampling (TS), account only for parametric uncertainty and assume identically distributed (*homoscedastic*) returns.
- Variability of the returns in RL depends on the current state and action, and is heteroscedastic.

Approach

- Use *Information-Directed Sampling* (IDS) [1, 2] for exploration in RL.
- Develop tractable approximation of IDS for deep RL, based on *distributional RL*.
- Explicitly account both for parametric uncertainty and heteroscedastic observation noise (return uncertainty).

Background

Information-Directed Sampling

- Bandit algorithm; Focus on Deterministic Frequentist IDS [1]
- Balance between incurring *regret* $\Delta_t(\mathbf{a}) = \mathbb{E}\left[R(\mathbf{a}^*) - R(\mathbf{a})\right]$ and acquiring *information gain* $I_t(\mathbf{a})$

$$\Psi_t(\mathbf{a}) := \frac{\Delta_t(\mathbf{a})^2}{I_t(\mathbf{a})} \qquad \text{(regret-information ratio)}$$

- Select action that minimizes the *regret-information ratio*

$$\mathbf{a}_t^{\text{IDS}} \in \arg\min_{\mathbf{a}\in\mathcal{A}} \frac{\Delta_t(\mathbf{a})^2}{I_t(\mathbf{a})}.$$

Choose information gain $I_t(\mathbf{a}) = \log\left(1 + \sigma_t(\mathbf{a})^2/\rho_t(\mathbf{a})^2\right)$, where

- $\sigma_t(\mathbf{a})^2$ is the variance in the parametric estimate
- $\rho_t(\mathbf{a})^2$ is the variance of the heteroscedastic return observation

Gaussian Process Bandit Example

- UCB and TS account only for parametric uncertainty and sample very noisy actions.
- IDS accounts for heteroscedastic noise and shrinks parametric uncertainty by sampling nearby points with low noise.

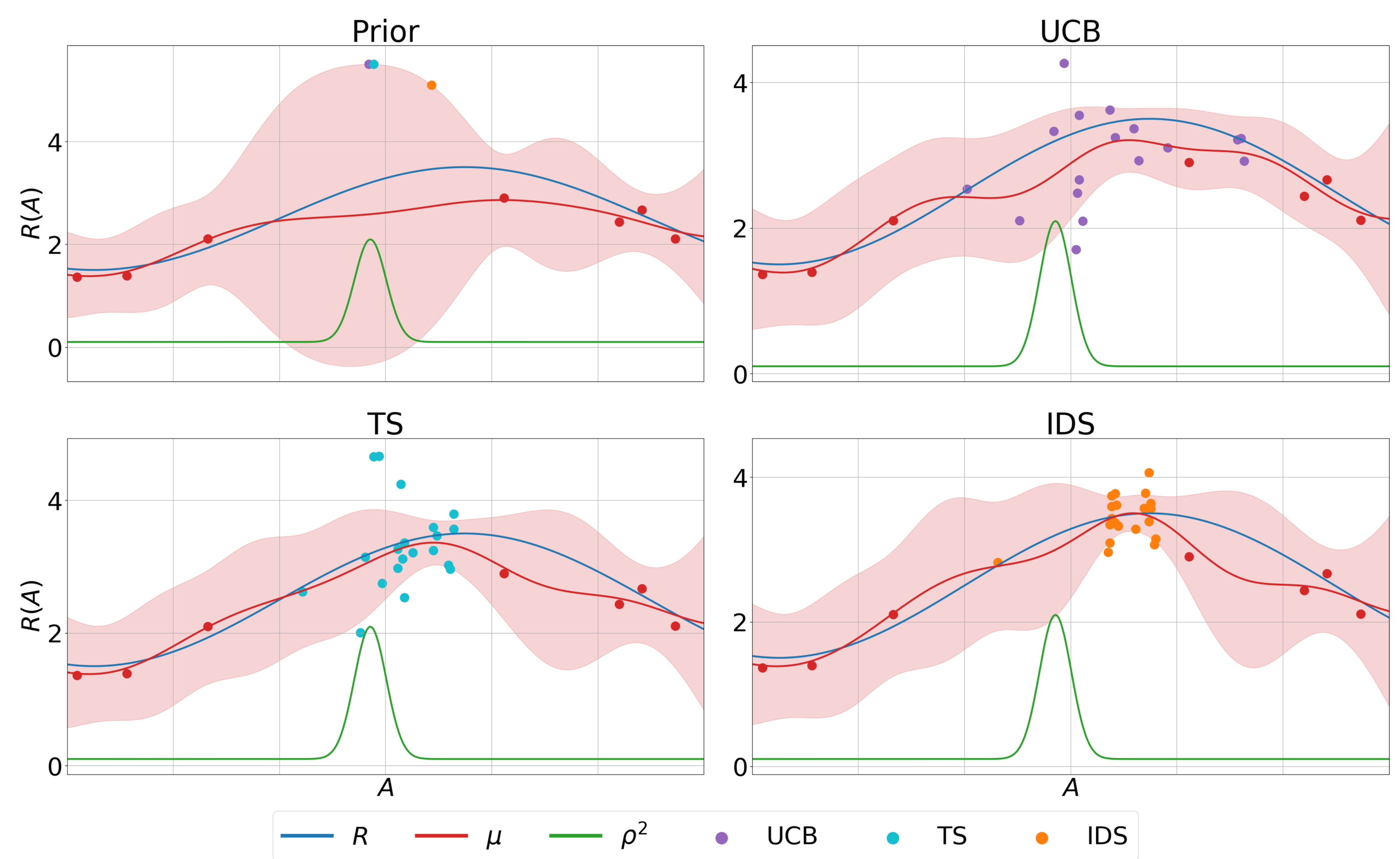


Figure 1. UCB, TS, IDS in Gaussian Processes. R: true function, μ mean estimate, ρ^2: observation noise,

Information-Directed Sampling for Reinforcement Learning

DQN with ϵ-greedy	→	UCB/TS exploration with Bootstrap-DQN
↓		↓
Distributional RL with ϵ-greedy (e.g. C51)	→	C51-IDS (ours)

Regret

Estimate parametric uncertainty of $Q(\mathbf{s},\mathbf{a})$ via Bootstrap-DQN [3] mean and variance

- Use a neural net with K bootstrap heads $\{Q_k\}_{k=1}^K$

$$\mu(\mathbf{s},\mathbf{a}) = \frac{1}{K}\sum_{k=1}^K Q_k(\mathbf{s},\mathbf{a}), \qquad \sigma(\mathbf{s},\mathbf{a})^2 = \frac{1}{K}\sum_{k=1}^K \left(Q_k(\mathbf{s},\mathbf{a}) - \mu(\mathbf{s},\mathbf{a})\right)^2. \tag{1}$$

Use bootstrap confidence intervals for conservative *regret* estimate:

$$\hat{\Delta}_t^\pi(\mathbf{s},\mathbf{a}) = \max_{\mathbf{a}'\in\mathcal{A}} \underbrace{\left(\mu_t(\mathbf{s},\mathbf{a}') + \lambda_t\sigma_t(\mathbf{s},\mathbf{a}')\right)}_{ucb(\mathbf{s},\mathbf{a}')} - \underbrace{\left(\mu_t(\mathbf{s},\mathbf{a}) - \lambda_t\sigma_t(\mathbf{s},\mathbf{a})\right)}_{lcb(\mathbf{s},\mathbf{a})} \tag{2}$$

Information Gain

Use distributional RL (e.g. C51) to estimate heteroscedastic noise in $Q(\mathbf{s},\mathbf{a}) = \mathbb{E}[Z(\mathbf{s},\mathbf{a})]$

$$Z^\pi(\mathbf{s},\mathbf{a}) \stackrel{D}{=} R(\mathbf{s},\mathbf{a}) + \gamma Z^\pi(\mathbf{s}',\mathbf{a}').$$

Use the normalized return uncertainty

$$\rho(\mathbf{s},\mathbf{a})^2 = \frac{\operatorname{Var}\left(Z(\mathbf{s},\mathbf{a})\right)}{\epsilon_1 + \frac{1}{|\mathcal{A}|}\sum_{\mathbf{a}'\in\mathcal{A}}\operatorname{Var}\left(Z(\mathbf{s},\mathbf{a}')\right)}, \tag{3}$$

and the parametric uncertainty $\sigma(\mathbf{s},\mathbf{a})^2$ to compute the *information gain*

$$I(\mathbf{s},\mathbf{a}) = \log\left(1 + \frac{\sigma(\mathbf{s},\mathbf{a})^2}{\rho(\mathbf{s},\mathbf{a})^2}\right) + \epsilon_2. \tag{4}$$

Algorithm

Algorithm 1 Deterministic Information-Directed Q-learning

Input: λ, action-value function Q with K outputs $\{Q_k\}_{k=1}^K$, action-value return distribution Z
for episode $i = 1 : M$ do
 for step $t = 0 : T$ do
  $\mu(\mathbf{s}_t,\mathbf{a}) = \frac{1}{K}\sum_{k=1}^K Q_k(\mathbf{s}_t,\mathbf{a})$
  $\sigma(\mathbf{s}_t,\mathbf{a})^2 = \frac{1}{K}\sum_{k=1}^K \left[Q_k(\mathbf{s}_t,\mathbf{a}) - \mu(\mathbf{s}_t,\mathbf{a})\right]^2$
  $\hat{\Delta}(\mathbf{s}_t,\mathbf{a}) = \max_{\mathbf{a}'\in\mathcal{A}}\left[\mu(\mathbf{s}_t,\mathbf{a}') + \lambda\sigma(\mathbf{s}_t,\mathbf{a}')\right] - \left[\mu(\mathbf{s}_t,\mathbf{a}) - \lambda\sigma(\mathbf{s}_t,\mathbf{a})\right]$
  $\rho(\mathbf{s}_t,\mathbf{a})^2 = \operatorname{Var}\left(Z(\mathbf{s}_t,\mathbf{a})\right) / \left(\epsilon_1 + \frac{1}{|\mathcal{A}|}\sum_{\mathbf{a}'\in\mathcal{A}}\operatorname{Var}\left(Z(\mathbf{s}_t,\mathbf{a}')\right)\right)$
  $I(\mathbf{s}_t,\mathbf{a}) = \log\left(1 + \frac{\sigma(\mathbf{s}_t,\mathbf{a})^2}{\rho(\mathbf{s}_t,\mathbf{a})^2}\right) + \epsilon_2$
  Compute regret-information ratio: $\hat{\Psi}(\mathbf{s}_t,\mathbf{a}) = \frac{\hat{\Delta}(\mathbf{s}_t,\mathbf{a})^2}{I(\mathbf{s}_t,\mathbf{a})}$
  Execute action $\mathbf{a}_t = \arg\min_{\mathbf{a}\in\mathcal{A}}\hat{\Psi}(\mathbf{s}_t,\mathbf{a})$
 end for
end for

Experiments

We evaluate two versions:

- DQN-IDS: *homoscedastic* version, $\rho(\mathbf{s},\mathbf{a})^2 = \text{const}$
- C51-IDS: *heteroscedastic* version, use C51 [4] to estimate $\rho(\mathbf{s},\mathbf{a})^2$
- For fair comparison, *no gradients* flow from distributional head $Z(\mathbf{s},\mathbf{a})$ into convolutional layers

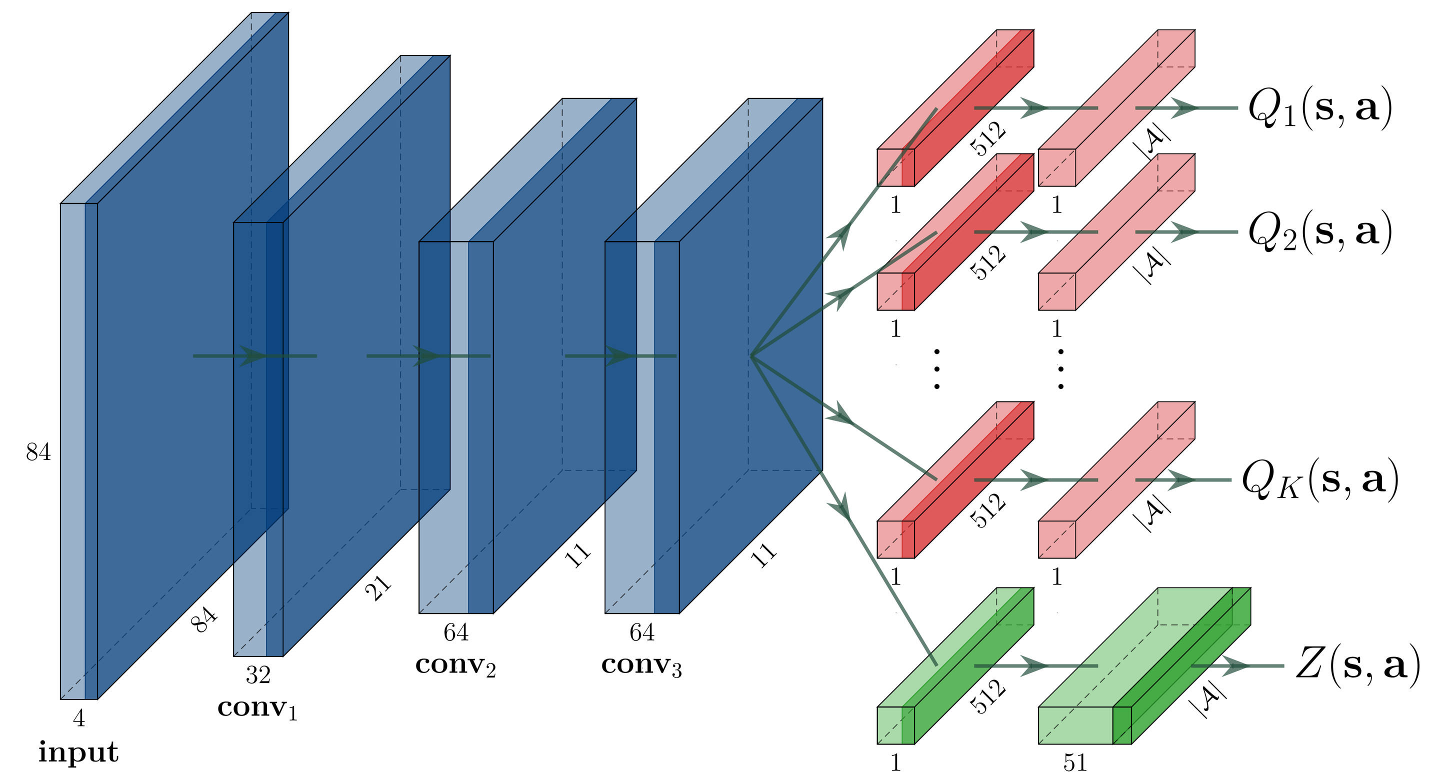
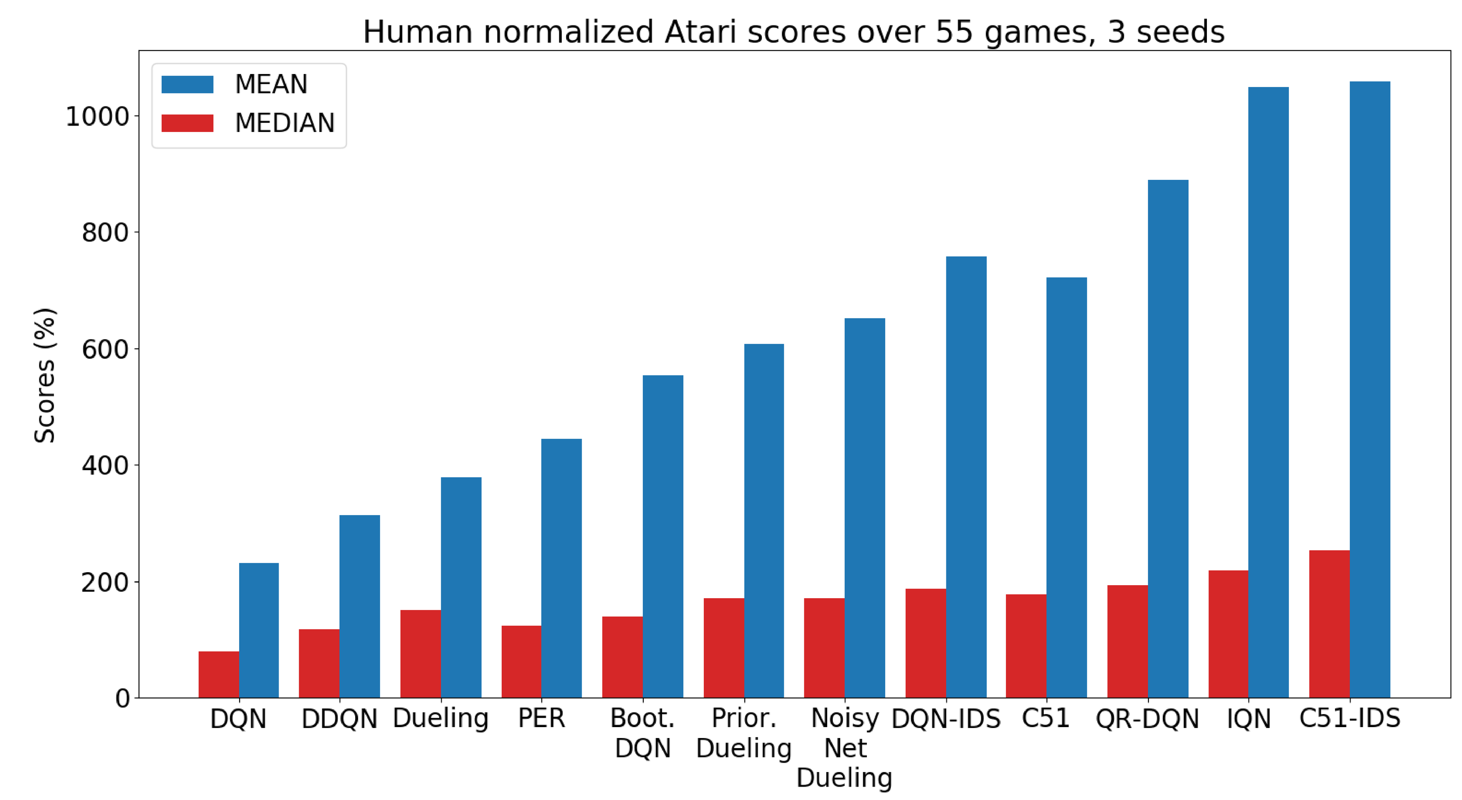


Figure 2. C51-IDS architecture

Results



- DQN-IDS outperforms Bootstrap-DQN (TS) by $\approx 200\%$
- C51-IDS improves on DQN-IDS and highlights the importance of considering *heteroscedastic* returns
- C51-IDS achieves best performance among C51, QR-DQN, IQN

References

[1] J. Kirschner and A. Krause, "Information directed sampling and bandits with heteroscedastic noise," 2018.

[2] D. Russo and B. Van Roy, "Learning to optimize via information-directed sampling," 2014.

[3] I. Osband, C. Blundell, A. Pritzel, and B. Van Roy, "Deep exploration via bootstrapped dqn," 2016.

[4] M. G. Bellemare, W. Dabney, and R. Munos, "A distributional perspective on reinforcement learning," 2017.