We propose an incremental, model-free exploration algorithm with fast-converging upper-confidence bounds, called UCLS.

Motivation for Directed Exploration

Common approaches to exploration like optimistic initialization are not always viable. Hence, we would like a mechanism for directed exploration in reinforcement learning. For instance, if we have access to uncertainty, \( \hat{U}(s,a) \), around mean estimates \( \hat{Q}(s,a) \), action selection can be greedy w.r.t. \( \hat{Q}(s,a) + \hat{U}(s,a) \), which provides a high-confidence upper-bound for the best possible action in the state \( s \).

Let \( \hat{Q} = \hat{Q} + \hat{U} \), and let \( \pi_c \) be the policy induced by greedy action selection on \( \hat{Q} \). Then, this process of action selection converges to a policy that is optimal under a defined confidence upper-bound for the best possible action in the state \( s \).

Confidence Interval Bounds for Policy Evaluation

Given the noise w.r.t. the optimal estimator \( \mathbf{w}^* \),

\[
r_{t+1} = (x_t - y_t x_{t+1})' \mathbf{w}^* + \nu_t
\]

and a finite set of samples, \( T \), with \( \nu_T \sim \sum_{t=0}^{T-1} \mathbf{z}_t \nu_t \), we show that the following holds with probability at least \( 1 - p \):

\[
x' \mathbf{w}^* \leq x' \mathbf{w}_T
\]

\[
+ \frac{p+1}{p} \sqrt{x' \mathbb{E}[A_T' \mathbf{z}_T \mathbf{z}_T' A_T]} \mathbf{x} + O \left( \mathbb{E}[(x' \mathbf{e}_T)^2] \right)
\]

Further, if we assume \( \nu_t \sim N(0, \sigma^2) \), and \( \mathbf{z}_T \sim \sum_{t=0}^{T-1} \mathbf{z}_t \), we show that the following holds with probability at least \( 1 - p \):

\[
x' \mathbf{w}^* \leq x' \mathbf{w}_T
\]

\[
+ \sigma \frac{p+1}{p} \sqrt{x' \mathbb{E}[A_T' \mathbf{z}_T \mathbf{z}_T' A_T]} \mathbf{x} + O \left( \mathbb{E}[(x' \mathbf{e}_T)^2] \right)
\]

Control with Confidence Interval Bounds

The bounds derived are for a stationary policy. But during control, the policy is slowly changing, and therefore, we slowly track these upper-bounds resulting in:

- Upper-Confidence Least Squares (UCLS) - for bound in Equation (1).
- Global Variance-Upper Confidence Bound (GV-UCB) - for bound in Equation (2).

Results: Comparing State-of-the-Art Exploration Methods

We compare UCLS against algorithms that use other approaches to estimate confidence intervals:

- DGPQ - using GPs. [1]
- LSPI-Rmax - using a measure of knownness. [2]
- RLSVI - using Bayesian Linear Regression. [3]
- UCBBootstrap - using bootstrapped confidence intervals. [4]

Overview

- We propose an incremental, model-free exploration algorithm with fast-converging upper-confidence bounds, called UCLS.
- We derive confidence intervals around action-values for LSTD, and use it to provide a directed exploration signal during control by tracking the bounds with a slowly changing policy.

Results: Advantage of Contextual Variance

This study contrasts the advantage of contextual variance estimates (UCLS) over global variance estimates (GV-UCB).

UCLS-Linear: An Effective Linear Complexity Variant

- Pros: (1) lower computational complexity, (2) may be more amenable to changing representations.
- Cons: (1) while the performance of the two algorithms is comparable, UCLS-Linear experiences more regret (e.g. River Swim, Puddle World), (2) two hyper parameters that need to be tuned.

Conclusion & Future Work

- Context-based exploration is a promising direction for designing sample-efficient learning algorithms. UCLS is a principled application of this motivation within the LSTD learning framework.
- Further lines of research include: (1) UCLS/UCLS-L with other stationary and changing representations, (2) other approaches to promote context-based exploration at a grounded or abstract level.

References