Context-Dependent Upper-Confidence Bounds for Directed Exploration

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.

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. $\hat{Q}(S_t,a) + \hat{U}(S_t,a)$, which provides a high-confidence upper-bound for the best possible action in the state S_t.

Let $\tilde{Q}_t = \hat{Q}_t + \hat{U}_t$, and let π_t be the policy induced by greedy action selection on $\tilde{Q}_t$. Then, this process of action selection converges to a policy that is optimal under a defined density:

$$Q^* = \underset{Q \in \mathcal{Q}}{\text{argmax}} \int_{\mathcal{S}\times\mathcal{A}} d(s,a) Q(s,a) ds da$$

This requires three key assumptions:

- *Stochastic Optimism:* After some point $T > 0$, for all $t \geq T$, $\mathbb{E}[\tilde{Q}_t(S,A)] \geq \mathbb{E}[Q^*(S,A)]$.
- *Shrinking Confidence Interval Radius:* $\mathbb{E}[\hat{U}_t(S,A)] \leq f(t)$ for some non-negative function f with $f(t) \to 0$.
- *Convergent Action Values:* $\left|\mathbb{E}[\hat{Q}_t(S,A) - Q^{\pi_t}(S,A)]\right| \leq g(t)$ for some non-negative function g with $g(t) \to 0$.

Optimistic Values Theorem states that under these 3 assumptions:

$$\text{Regret}(T) \stackrel{\text{def}}{=} \sum_{t=1}^{T} \mathbb{E}[Q^*(S,A)] - \mathbb{E}[Q^{\pi_t}(S,A)] \leq \sum_{t=1}^{T} f(t) + g(t)$$

Confidence Interval Bounds for Policy Evaluation

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

$$r_{t+1} = (\mathbf{x}_t - \gamma \mathbf{x}_{t+1})^\top \mathbf{w}^* + \nu_t$$

and a finite set of samples, T, with $\bar{\boldsymbol{\nu}}_T \stackrel{\text{def}}{=} \frac{1}{T}\sum_{t=0}^{T-1} \mathbf{z}_t \nu_t$, we show that the following holds with probability at least $1-p$:

$$\mathbf{x}^\top \mathbf{w}^* \leq \mathbf{x}^\top \mathbf{w}_T + \sqrt{\frac{p+1}{p}} \sqrt{\mathbf{x}^\top \mathbb{E}[\mathbf{A}_T^+ \bar{\boldsymbol{\nu}}_T \bar{\boldsymbol{\nu}}_T^\top \mathbf{A}_T^{+\top}] \mathbf{x}} + O\left(\mathbb{E}[(\mathbf{x}^\top \boldsymbol{\epsilon}_T^*)^2]\right) \quad (1)$$

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

$$\mathbf{x}^\top \mathbf{w}^* \leq \mathbf{x}^\top \mathbf{w}_T + \sigma \sqrt{\frac{p+1}{p}} \sqrt{\mathbf{x}^\top \mathbb{E}[\mathbf{A}_T^+ \bar{\mathbf{z}}_T \bar{\mathbf{z}}_T^\top \mathbf{A}_T^{+\top}] \mathbf{x}} + O\left(\mathbb{E}[(\mathbf{x}^\top \boldsymbol{\epsilon}_T^*)^2]\right) \quad (2)$$

Control with Confidence Interval Bounds

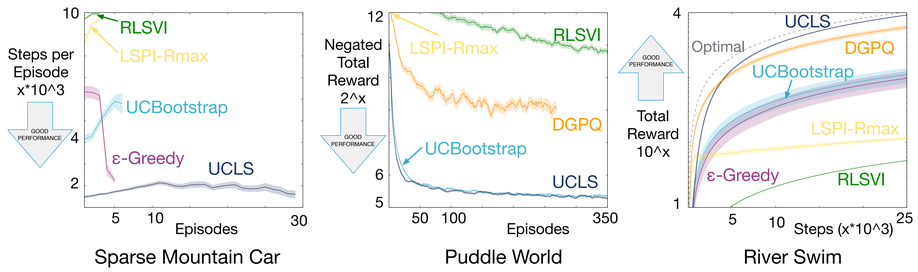
The bounds derived are for a stationary policy. But during control, the policy is slowly changing, and therefore, we slowly track these upper-bounds resulting in:

- 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

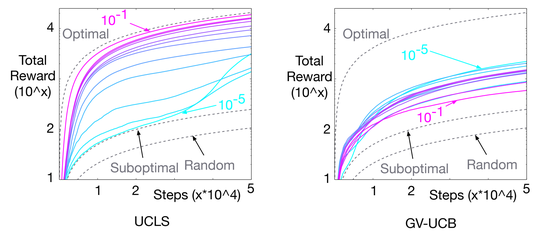
We compare UCLS against algorithms that use other approaches to estimate confidence intervals:

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

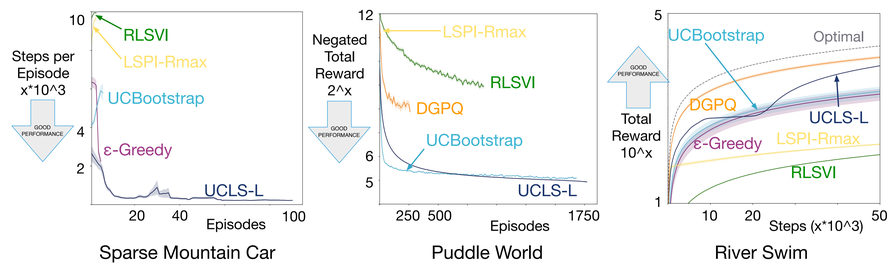


Results: Advantage of Contextual Variance

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-L 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

- R Grande, T Walsh, and J How. "Sample Efficient Reinforcement Learning with Gaussian Processes". In: *International Conference on Machine Learning*. 2014.
- L Li, ML Littman, and CR Mansley. "Online exploration in least-squares policy iteration". In: *International Conference on Autonomous Agents and Multiagent Systems*. 2009.
- I Osband, B Van Roy, and Z Wen. "Generalization and Exploration via Randomized Value Functions.". In: *International Conference on Machine Learning*. 2016.
- M White and A White. "Interval estimation for reinforcement-learning algorithms in continuous-state domains". In: *Advances in Neural Information Processing Systems*. 2010.

Raksha Kumaraswamy[1], Matthew Schlegel[1], Adam White[1,2], Martha White[1]

kumarasw@ualberta.ca, mkschleg@ualberta.ca, adamwhite@google.com, whitem@ualberta.ca