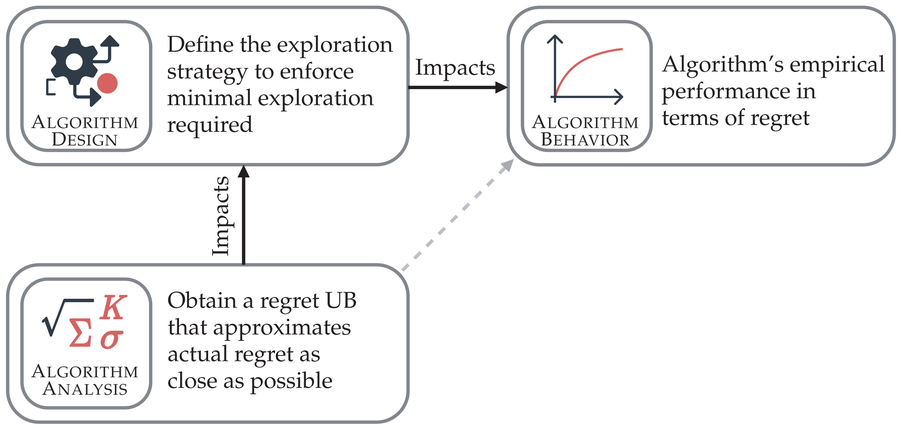
Position: Constants are Critical in Regret Bounds for Reinforcement Learning

Simone Drago, Marco Mussi, and Alberto Maria Metelli
{simone.drago, marco.mussi, albertomaria.metelli}@polimi.it

Our Position

Ignoring multiplicative constants when evaluating whether an algorithm matches the problem's lower bound may lead to disappointing results when *using* such an algorithm

Discussion



Key Takeaway

More effort in designing the exploration method of an algorithm w.r.t. lower order and constant terms can lead to significant performance improvements

Effects are clear in tabular reinforcement learning

Idea for Further Inquiries

We conjecture that the performance difference ascribable to tighter constants and lower order terms will be even higher in more complex scenarios

Research Directions

Study theoretical guarantees of existing algorithms in more complex settings ⇨ Gain deeper knowledge of inner workings of complex settings

Research novel mathematical tools to analyze complex settings

Alternative Views

Theoretical Researchers

Algorithms' significance can extend beyond the tightness of their theoretical guarantees

Algorithms may introduce novel technical tools and algorithmic solutions

Results of independent interest

Experimentalists

Pursuing algorithm with tight theoretical guarantees may not be the optimal research path

Worst-case guarantees can be overly conservative as they rarely occur

Focus on the empirical performance of algorithms

Case Study: the UCBVI Algorithm

Tabular Reinforcement Learning

MDP with S states, A actions, horizon H, K episodes, $T = HK$ time steps

Lower Bound: $\Omega(\sqrt{HSAT})$

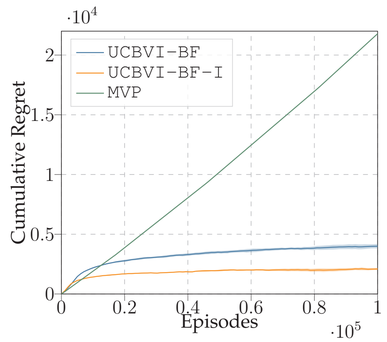
Upper Bounds:

`UCBVI-BF` $\tilde{\mathcal{O}}(\sqrt{HSAT})$ For "sufficiently large" T → Refined Analysis → `UCBVI-BF-I` $\tilde{\mathcal{O}}(\sqrt{HSAT})$ For "sufficiently large" T — 2x Empirical Improvement

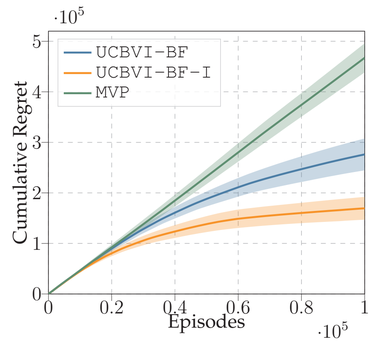
`MVP` $\tilde{\mathcal{O}}(\sqrt{HSAT})$ For all T — Theoretical Improvement

`MVP` employs a *doubling trick*, known to be convenient in terms of regret analysis, but negatively affects the empirical efficiency

`UCBVI-BF-I` improves upon `UCBVI-BF`, yet is still theoretically suboptimal w.r.t. `MVP`



Riverswim environment with $S = 5$, $H = 10$



Toy environment with $S = 5$, $A = 10$, and $H = 10$

References

Y. Abbasi-Yadkori, D. Pál, and C. Szepesvári. Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, 2011.

P. Auer, N. Cesa-Bianchi, and P. Fischer. Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 2002.

M. G. Azar, I. Osband, and R. Munos. Minimax regret bounds for reinforcement learning. In *International Conference on Machine Learning*, 2017.

S. Bubeck. *Bandits games and clustering foundations*. PhD thesis, Université des Sciences et Technologie de Lille-Lille, 2010.

V. Dani, T. P. Hayes, and S. M. Kakade. Stochastic linear optimization under bandit feedback. In *Conference on Learning Theory*, 2008.

Z. Zhang, Y. Chen, J. D. Lee, and S. S. Du. Settling the sample complexity of online reinforcement learning. In *Conference on Learning Theory*, 2024.