Convex Hull Monte-Carlo Tree-Search

Michael Painter, Bruno Lacerda, Nick Hawes
Oxford Robotics Institute, University of Oxford
[mpainter, bruno, nickh]@robots.ox.ac.uk

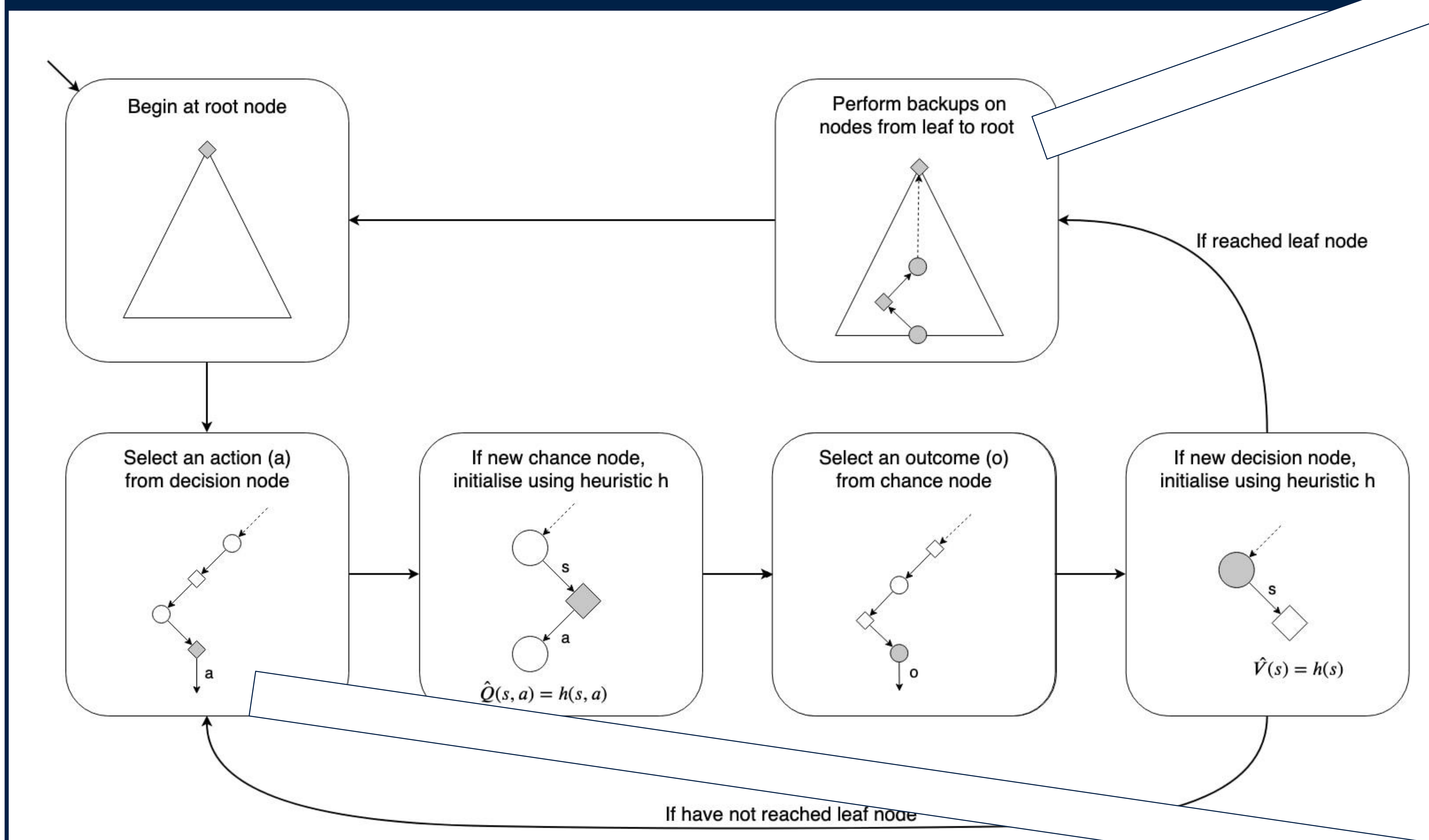
UNIVERSITY OF OXFORD

Contributions

The main contributions of this work are:

- applying the notion of **contextual regret** to multi-objective planning, and justify that exploration policies that achieve low contextual regret explore the trade-offs between objectives appropriately, as opposed to other metrics proposed in the literature;
- proposing **Contextual Zooming for Trees**, that outperforms prior work on this metric.

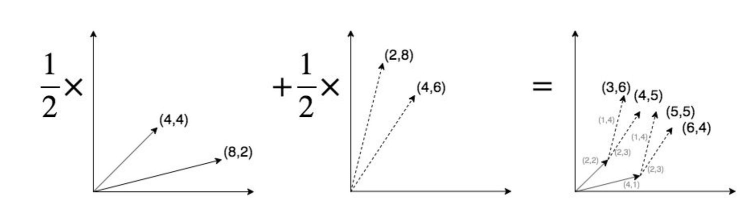
Trial-based Heuristic Tree-Search [1]

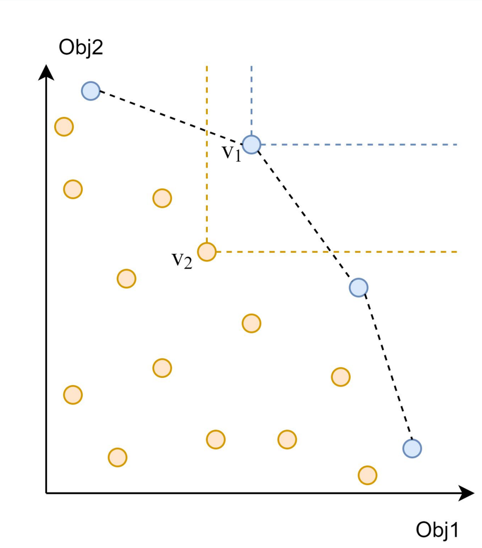


Backup Functions

Arithmetic over sets of vectors $\hat{C}$ and $\hat{D}$:

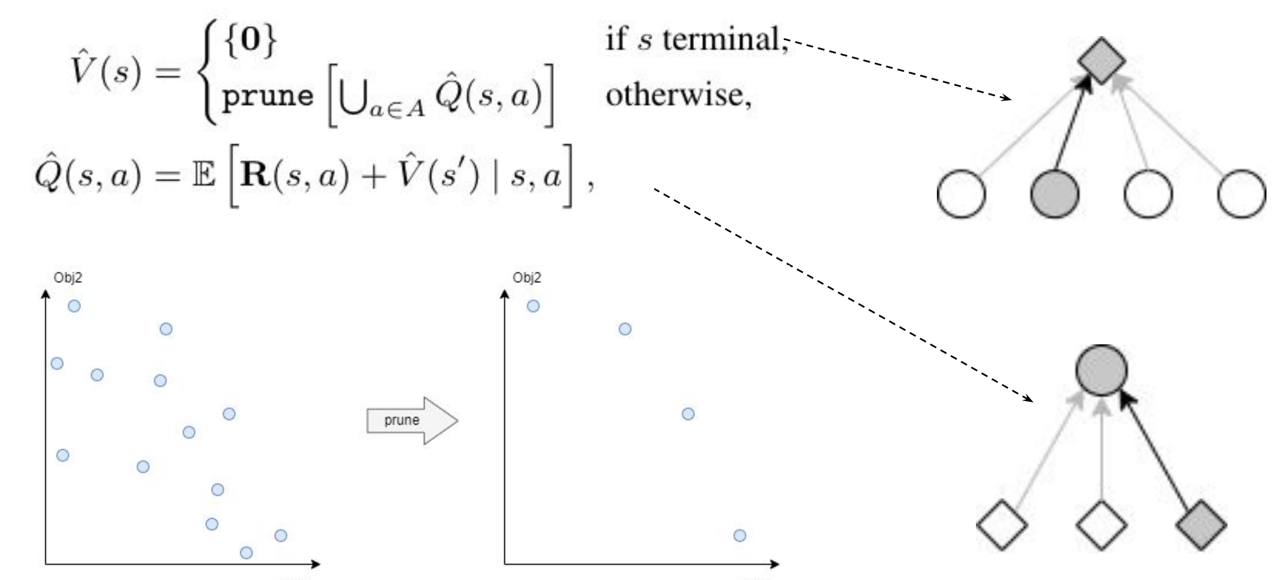
$$\hat{C} + \hat{D} = \{\mathbf{c} + \mathbf{d} \mid \mathbf{c} \in \hat{C}, \mathbf{d} \in \hat{D}\},$$
$$\mathbf{b} + k\hat{C} = \{\mathbf{b} + k\mathbf{c} \mid \mathbf{c} \in \hat{C}\}.$$





The **Convex Hull Value Iteration** [2] backup equations:

$$\hat{V}(s) = \begin{cases} \{\mathbf{0}\} & \text{if } s \text{ terminal,} \\ \texttt{prune}\left[\bigcup_{a \in A} \hat{Q}(s,a)\right] & \text{otherwise,} \end{cases}$$
$$\hat{Q}(s,a) = \mathbb{E}\left[\mathbf{R}(s,a) + \hat{V}(s') \mid s, a\right],$$

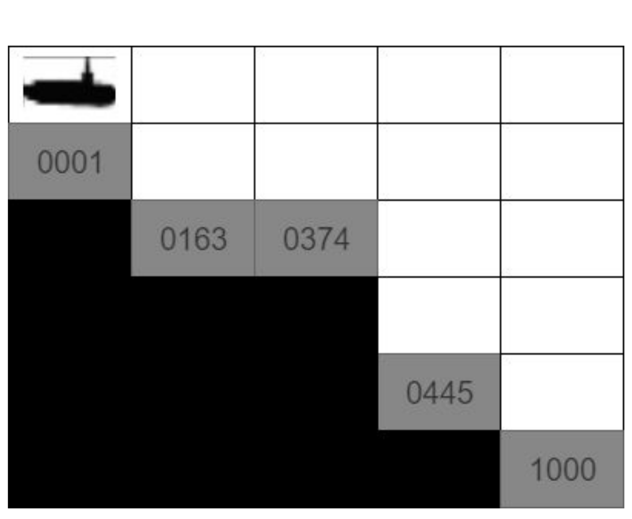


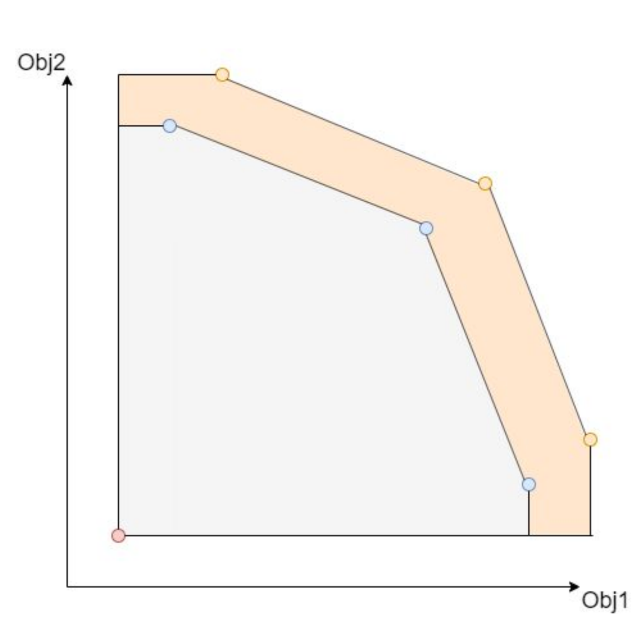
Action Selection

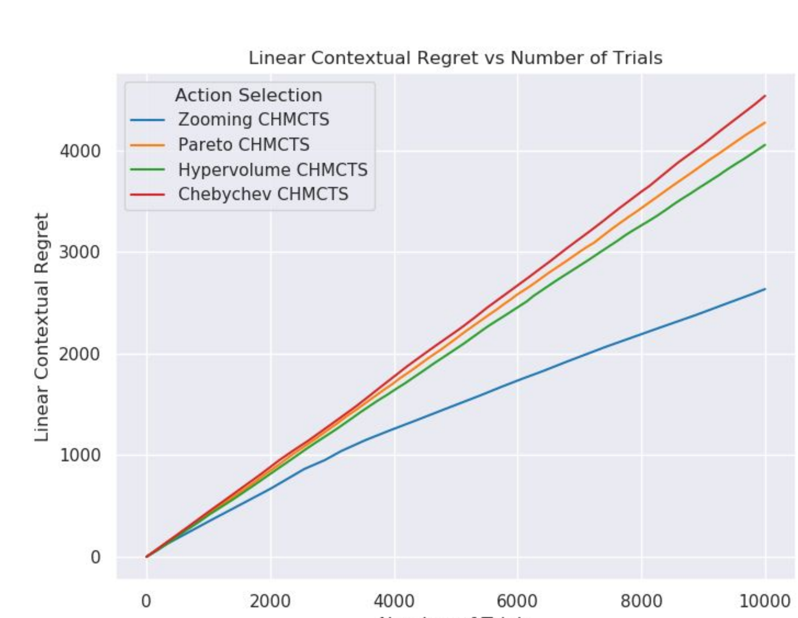
- We can analyse action selection using the **Contextual Multi-Armed Bandit Problem** and **Contextual Regret**.
- UCB1 commonly used for action selection in single-objective problems.
  - The decision problem at a decision node is a non-stationary multi-armed bandit problem.
- In CHMCTS we associate a context weight vector with each trial.
  - Non-stationary multi-armed bandit → contextual non-stationary multi-armed bandit.
- **Contextual Zooming** [3] used for contextual multi-armed bandit problems.
- Can also use action selection from prior multi-objective MCTS work.

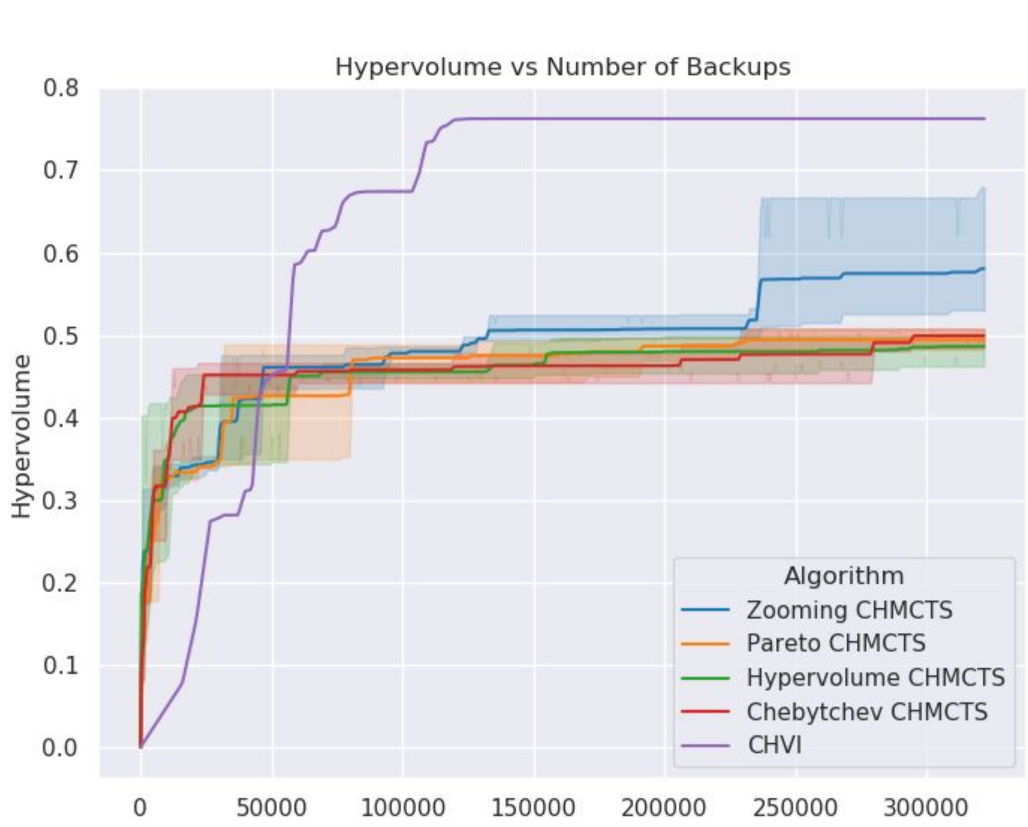
Results

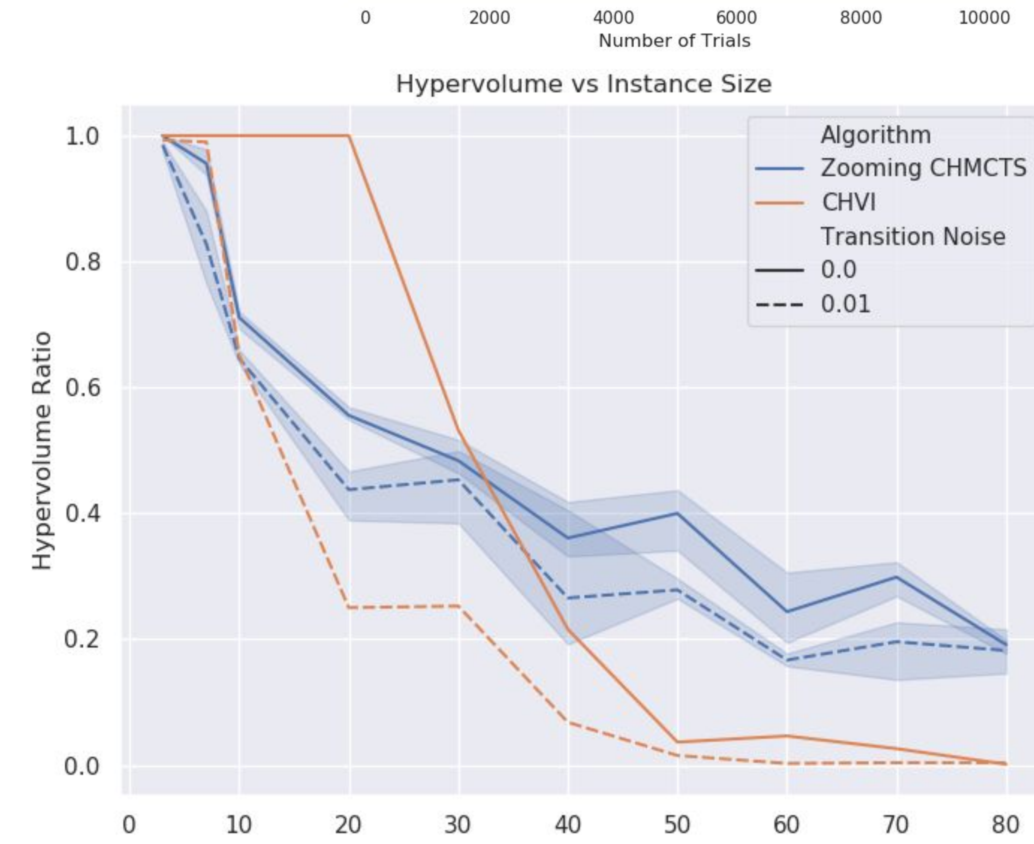
- Our algorithm outperforms prior state-of-the-art multi-objective monte-carlo tree-search methods, and can additionally handle stochastic environments.
- Our algorithm also empirically obtains a sub-linear contextual regret.











References

[1] Barrett, L., and Narayanan, S. 2008. Learning all optimal policies with multiple criteria. In Proceedings of the 25th international conference on Machine learning, 41–47. ACM.
[2] Keller, T., and Helmert, M. 2013. Trial-Based Heuristic Tree Search for Finite Horizon MDPs. In ICAPS.
[3] Slivkins, A. 2014. Contextual bandits with similarity information. The Journal of Machine Learning Research 15(1):2533–2568.