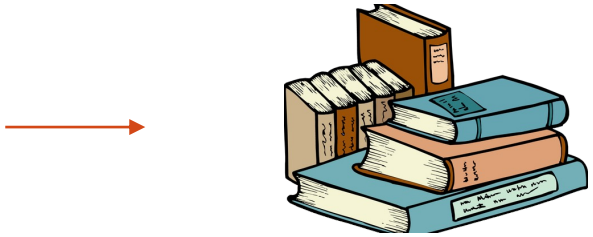
Scientific Experimentation for Reinforcement Learning

SCOTT JORDAN

UNIVERSITY OF ALBERTA

Motivation



Generate Knowledge

Experiments need to produce:

- correct and interpretable results
- knowledge that generalizes outside the specific experimental setting

Are Experiments in RL Sound? No.

Protecting Against Evaluation Overfitting in Empirical Reinforcement Learning

Deep Reinforcement Learning that Matters

Evaluating the Performance of Reinforcement Learning Algorithms

A Novel Benchmark Methodology and Data Repository for Real-life Reinforcement Learning

Re-evaluating Evaluation

Revisiting the Arcade Learning Environment: Evaluation Protocols and Open Problems for General Agents

Reproducibility of Benchmarked Deep Reinforcement Learning Tasks for Continuous Control

EVALUATION METHODS FOR REINFORCEMENT LEARNING

Introduction to the special issue on empirical evaluations in reinforcement learning

How Many Random Seeds? Statistical Power Analysis in Deep Reinforcement Learning Experiments

A Hitchhiker's Guide to Statistical Comparisons of Reinforcement Learning Algorithms

Deep Reinforcement Learning at the Edge of the Statistical Precipice

Issues with Experiments

Insufficient Trials / No quantification of uncertainty
- Henderson et al. 2018
- Colas et al. 2018, 2019
- Jordan et al. 2020
- Agarwal et al. 2021

No accounting for hyperparameter selection
- Dabney 2014
- Jordan et al. 2020

Differing Implementations
- Henderson et al. 2018
- Engstrom et al. 2020

Environment weighting favors one class of algorithm
- Balduzzi et al. 2018
- Jordan et al. 2020

All issues are with benchmarking!

Goals of the Talk

Convince you to:

1. Abandon benchmarking as the primary form of experimentation
2. Perform scientific testing experiments
   - Reveals how an algorithm works
3. Reject papers with only benchmarking experiments

Why is Benchmarking Hard?

Does algorithm X outperform algorithm Y on environments A, B, C

- Not a well-formed question.
- What does better performance mean?

How we construct the evaluation procedure determines what performance means

- Set of environments
- Metric: final performance, average over the agent's lifetime
- Hyperparameter selection process
- Score normalization
- Environment weighting
- Training time

There is no correct procedure! Only different representations of performance.

Specify the specific hypothesis you want to test, then design an evaluation procedure.

How to design an evaluation procedure

Does not matter! (For this talk)

Problem With Benchmarking

1. For a sufficiently diverse set of environments there are multiple best algorithms
2. Performance of an algorithm has many confounding factors
3. Only says which algorithms work well and not why
   - Claims as to why one algorithm is better are only a guess.
4. Performance backed claims can lead to works based on misunderstood concepts.

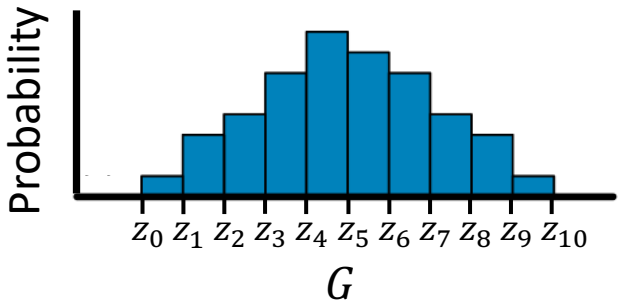
We need to know what makes an algorithm successful or not to build better algorithms.

Benchmarking Leads to Misunderstandings

Benchmarking Leads to Misunderstandings: Distribution RL

Value Predictions

- $\mathrm{G_t} = \sum_{k=0} \gamma^k R_{t+k}$
- Mean prediction: $q(s,a) \approx E[G_t | S_t = s, A_t = a]$
- Categorical prediction $q_i(s,a) \approx \Pr\big(G \in [z_i, z_{i+1})\big)$
- New algorithm C51 (Bellemare et al. 2017)



Motivation

- Better approximation of value
- More stable learning
- Mitigate the effects of learning from a nonstationary policy

Benchmarking Leads to Misunderstandings: Distribution RL

Experiments:

- Vary the expressivity of the distribution to see how much it impacts performance
- Check for State-of-the-art performance!

Claims (paraphrasing):

- More expressive distribution always increases performance
- State-of-the-art performance

No substantiation of the motivation

- Implicitly assumed true because of good performance

	Mean	Median	> H.B.	> DQN
DQN	228%	79%	24	0
DDQN	307%	118%	33	43
DUEL.	373%	151%	37	50
PRIOR.	434%	124%	39	48
PR. DUEL.	592%	172%	39	44
C51	**701%**	**178%**	**40**	**50**
UNREAL†	880%	250%	-	-

Figure 6. Mean and median scores across 57 Atari games, measured as percentages of human baseline (H.B., Nair et al., 2015).

New Distributional RL Algorithms

Quantile Regression DQN (Dabney et al. 2018a)

- New distribution representation

Implicit Quantile Networks (Dabney et al. 2018b)

- More expressive quantile approximation

Fully parameterized Quantile Function (Yang et al. 2019)

- More expressive quantile approximation

Distributed Distributional DDPG (Barth-Maron et al. 2018)

- Distributional critic

All cite superior performance of distributional RL as reasons for continued development.

Misunderstood Concept

Is distributional value representation actually important for good performance?

Lyle et al. (2018) showed

- Distributional RL only has benefits for neural networks
- Worse for Tabular and Linear representations

Dabney et al. (2021) suggested that

- Distributonal RL learns representations that are better able to predict future value functions

Better representation learning is likely the source of any performance benefit.

These insights do not come from benchmarking experiments!

Alternative Experiments

Controlled experimentation to understand how an algorithm works

- Scientific testing (Hooker, 1995)

The goal is to produce insightful knowledge that generalizes beyond the specific test setting so that we can construct or select algorithms as to solve specific problems.

Scientific Testing Example

The Mirage of Action-Dependent Baselines in Reinforcement Learning (Tucker et al., 2018)

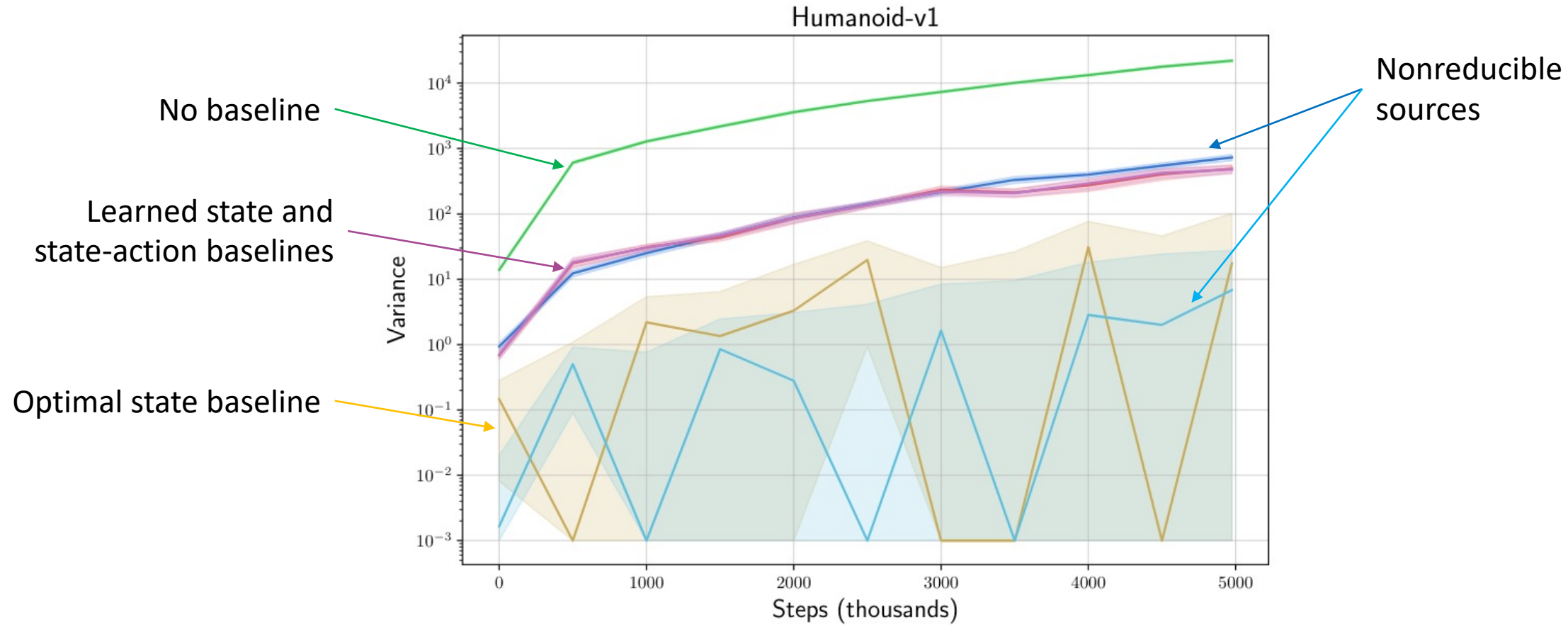
$$\nabla J(\theta) = \mathbf{E}\left[\sum_{t=0} \gamma^t G_t \frac{\partial}{\partial \theta} \ln \pi(S_t, A_t, \theta)\right]$$

$$\widehat{\nabla} J_s = \sum_{t=0} \gamma^t (G_t - v(S_t)) \frac{\partial}{\partial \theta} \ln \pi(S_t, A_t, \theta)$$

$$\widehat{\nabla} J_{s,a} = \sum_{t=0} \gamma^t (G_t - q(S_t, A_t)) \frac{\partial}{\partial \theta} \ln \pi(S_t, A_t, \theta) + \gamma^t \frac{\partial}{\partial \theta} \mathbf{E}[q(S_t\ A_t) |\ S_t]$$

Claim that $\mathrm{Var}\left(\widehat{\nabla} J_{s,a}\right) < \mathrm{Var}\left(\widehat{\nabla} J_s\right)$ and estimators using $\widehat{\nabla} J_{s,a}$ have better performance

Scientific Testing Example



Scientific Testing Example

Lessons learned:

1. Benchmarking led to misleading claims
2. Scientific testing revealed more insightful information

Comparison of Experiment Styles

BENCHMARKING

1. Shows one algorithm outperforms other algorithms
2. The experiment is valuable only if the new method is better
3. Hard to design proper experiments with interpretable results.

SCIENTIFIC TESTING

1. Produce knowledge about how one algorithm works
2. Experiment is valuable regardless of the outcome
3. Need to be creative in designing experiments

For Better Experiments

1. Stop using benchmarking to justify new algorithms
2. Start using scientific testing to teach each other how algorithms work
3. For real change we need to change reviewer expectations.

New Reviewing Criteria

Reject papers that only have benchmarking experiments:

- We do not conduct these correctly, so they have almost zero value
- Do not provide insights for future algorithm development

Reject papers that do not validate that design decisions were behaving as intended

- showing a distributional value representation leads to better approximation
- Showing the amount of variance reduction achieved by new control variates

Ignore reviews that ask for benchmark comparisons.

Conclusion

Focus on generating knowledge not algorithms.

Questions?

Contact me: sjordan@ualberta.ca

Bibliography

Bellemare, Marc G., Will Dabney, and Rémi Munos. "A distributional perspective on reinforcement learning." In *International Conference on Machine Learning*, pp. 449-458. PMLR, 2017.

Dabney, Will, Mark Rowland, Marc Bellemare, and Rémi Munos. "Distributional reinforcement learning with quantile regression." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1. 2018.

Dabney, Will, Georg Ostrovski, David Silver, and Rémi Munos. "Implicit quantile networks for distributional reinforcement learning." In *International conference on machine learning*, pp. 1096-1105. PMLR, 2018.

Yang, Derek, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. "Fully parameterized quantile function for distributional reinforcement learning." *Advances in neural information processing systems* 32 (2019).

Barth-Maron, Gabriel, Matthew W. Hoffman, David Budden, Will Dabney, Dan Horgan, Dhruva Tb, Alistair Muldal, Nicolas Heess, and Timothy Lillicrap. "Distributed distributional deterministic policy gradients." *arXiv preprint arXiv:1804.08617* (2018).

Lyle, Clare, Marc G. Bellemare, and Pablo Samuel Castro. "A comparative analysis of expected and distributional reinforcement learning." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, no. 01, pp. 4504-4511. 2019.

Dabney, Will, André Barreto, Mark Rowland, Robert Dadashi, John Quan, Marc G. Bellemare, and David Silver. "The value-improvement path: Towards better representations for reinforcement learning." In *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 8, pp. 7160-7168. 2021.

Hooker, John N. "Testing heuristics: We have it all wrong." *Journal of heuristics* 1, no. 1 (1995): 33-42.

Tucker, George, Surya Bhupatiraju, Shixiang Gu, Richard Turner, Zoubin Ghahramani, and Sergey Levine. "The mirage of action-dependent baselines in reinforcement learning." In *International conference on machine learning*, pp. 5015-5024. PMLR, 2018.