Reliable evaluation in reinforcement learning

Olivier Sigaud

Sorbonne Université
http://people.isir.upmc.fr/sigaud

Outline

- Evaluation issues in deep reinforcement learning
- Using appropriate statistical tests
- Better metrics to compare two algorithms
- Hyper-parameter tuning, performance comparisons
- Basics about the computational neuroscience side of RL

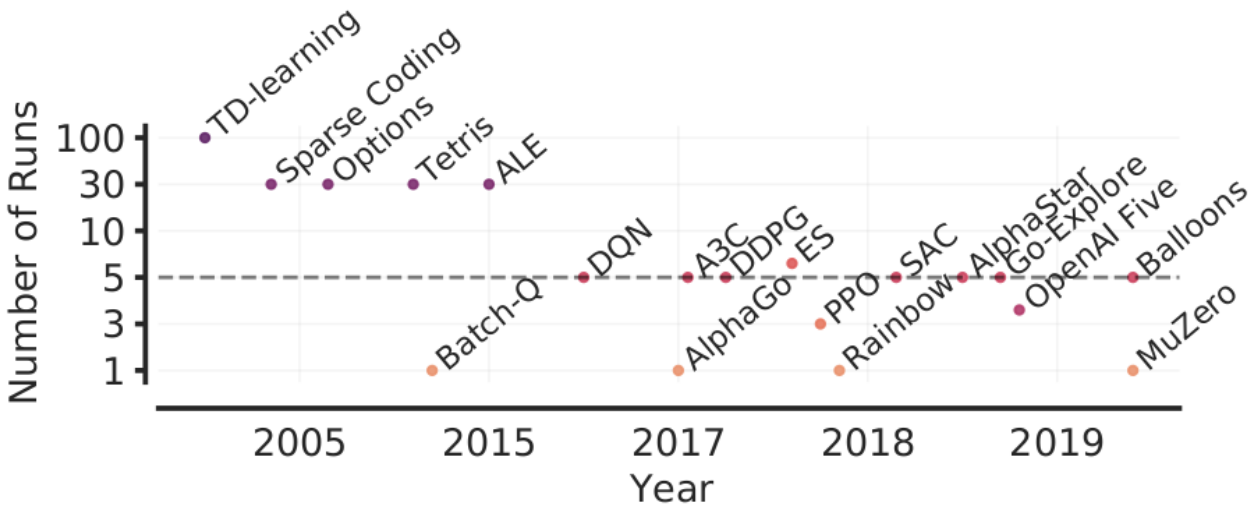
Various research goals

- Exploratory research: reach beyond frontiers, reveal new phenomena
- Theoretical research: prove some properties
- Empirical research: establish some properties from experience
- Empirical research requires a strong empirical methodology
- When results are stochastic, need to use several seeds and aggregate

Bouthillier, X., Laurent, C., and Vincent, P. (2019) Unreproducible research is reproducible. In *International Conference on Machine Learning*, pages 725–734. PMLR

Patterson, A., Neumann, S., White, M., and White, A. (2023) Empirical design in reinforcement learning. *arXiv preprint arXiv:2304.01315*

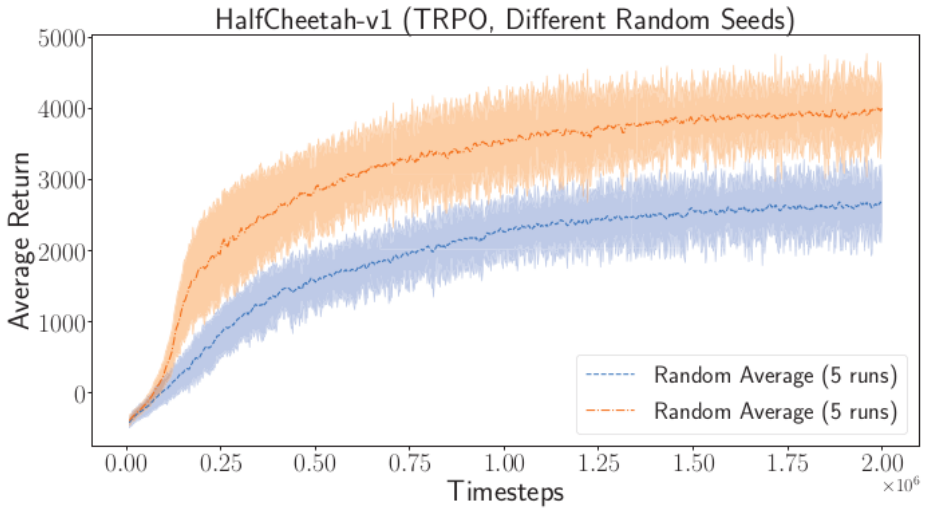
Insufficient number of seeds (common practices)



- With heavier environments, one cannot run enough seeds

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. (2021) Deep reinforcement learning at the edge of the statistical precipice. *Advances in neural information processing systems*, 34:29304–29320

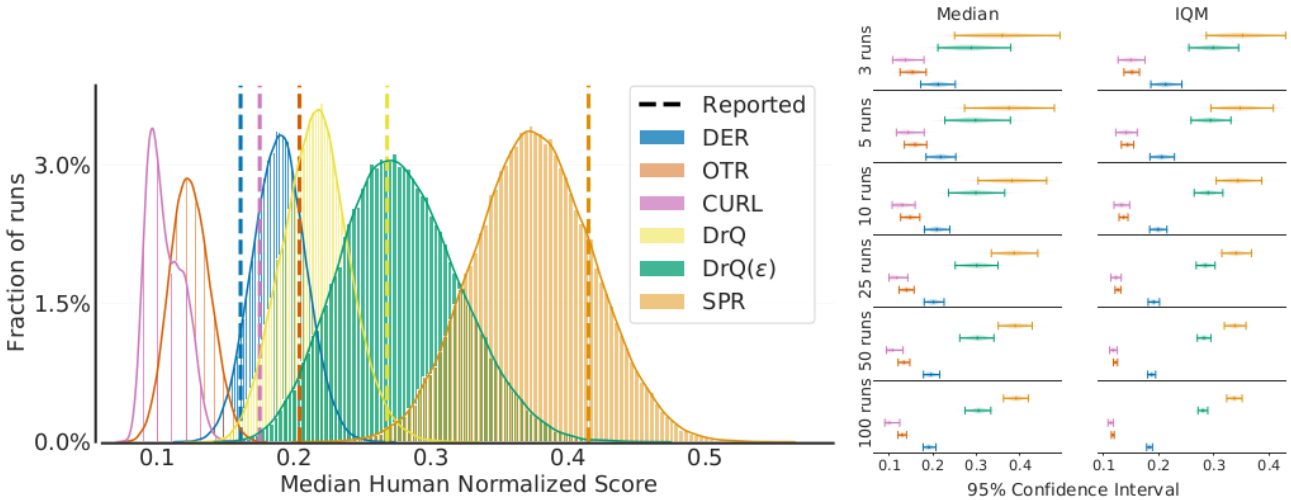
Insufficient number of seeds: the danger



- Without enough seeds, one may wrongly conclude to superiority of a method over another

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018) Deep reinforcement learning that matters. In McIlraith, S. A. and Weinberger, K. Q., editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence*, pp. 3207–3214. AAAI Press

Poor reporting practices



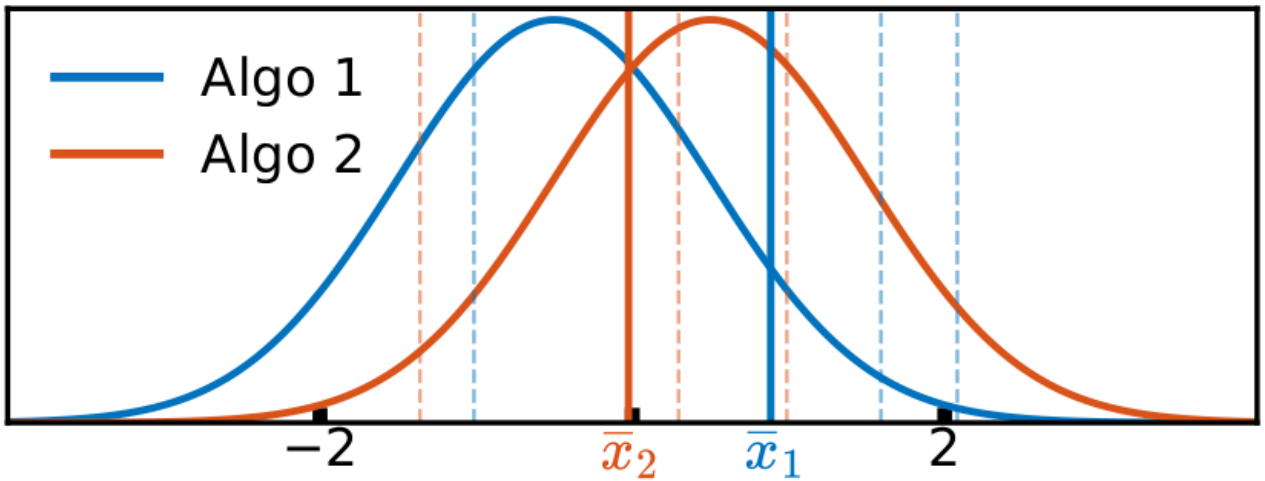
- Authors generally overestimate their method
- Authors generally publish point estimates, they should publish interval estimates
- Reasons for overestimation: selection of seeds, hyperparam overfitting
- Using a seen maximum is a very bad practice. Can be an outlier.
- More problems when comparing to competitors (fair tuning, etc.)
- Note that apart from the pink, the spread looks Gaussian

General remarks

- The RL research community is lead by Big Tech companies
- Massive use of more and more difficult benchmarks
- More focus on improvements and fancy results than on analysis and understanding
- RL algorithms are slowly moving towards being readily applicable to real-world tasks
- But methodological aspects and understanding are left behind
- This class: start from good practices

Statistical tests

Introduction: the problem



- Usually, RL is stochastic (in the policy and/or in the environment)
- Two episodes can give different results
- A superiority in data can be due to chance
- Need to rigorously compare two algorithms
- Statistical tests are meant to provide this rigor

Statistical tests: the framework

- One wants to compare the (true) central performances (mean or median) μ_1, μ_2 of two algorithms
- The null hypothesis $\mathcal{H}_0 : \mu_1 - \mu_2 = 0$ algorithms perform the same
- Alternative hypothesis $\mathcal{H}_a : |\mu_1 - \mu_2| > 0$ one algorithm is better
- Given a set of realizations, we observe $\bar{x}_1, \bar{x}_2$ (empirical central performances)
- With what confidence can we reject the null hypothesis?
- The confidence level cannot be 100%, would require an infinity of samples

Colas, C., Sigaud, O., and Oudeyer, P.-Y. (2019) A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms. *arXiv preprint arXiv:1904.06979*

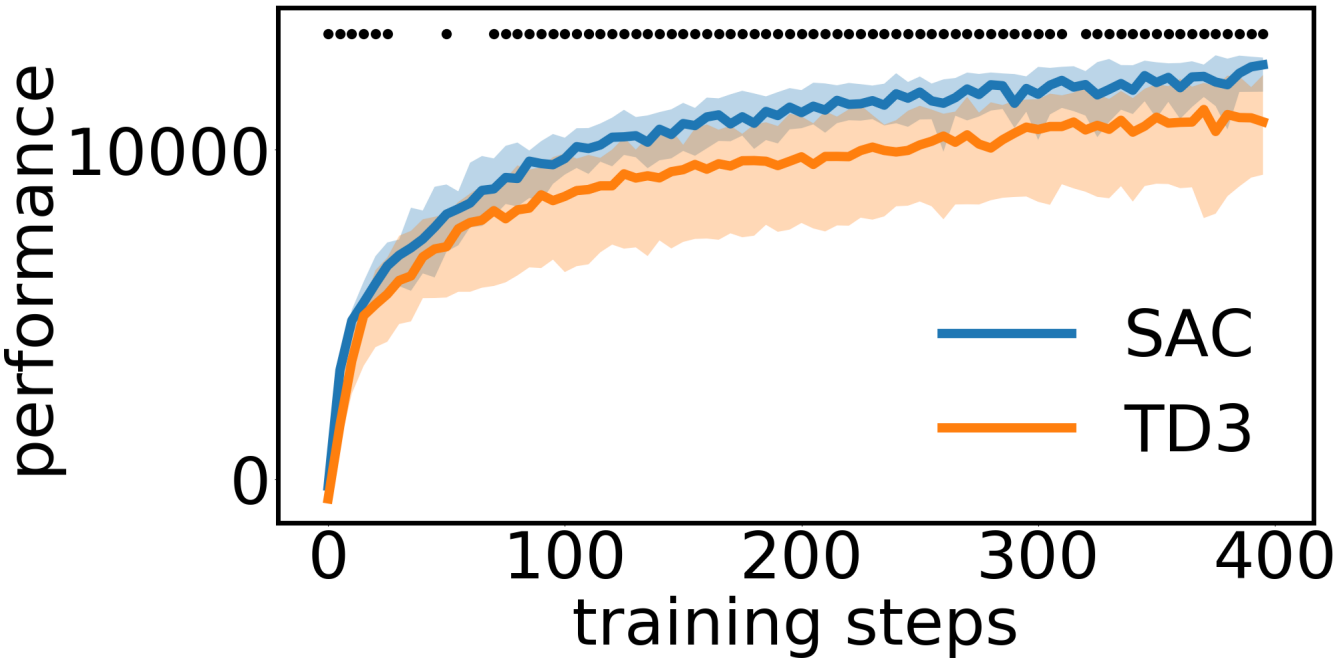
Statistical tests: definitions

- p-value: risk that the test wrongly rejects the null hypothesis
- I.e. probability of a "false positive" (difference found, but there is none)
- Usually, make sure $p - value < 0.05$
- We may claim that there is a (non-existing) difference 1 time out of 20...
- Statistical power: depends on sample size (how many data?) and effect size (how much difference?). The larger, the better

Various statistical tests and their assumptions

- (Student's) T-test: variances are equal (false when comparing two RL algorithms)
- Welch's T-test: variances are not equal (fine!)
- Wilcoxon Mann-Whitney (WMW) rank sum test: distributions are continuous, have the same shape and spread (wrong)
- Ranked T-test: close to MWM, with ranking before T-test
- Bootstrap confidence interval test: no assumptions, but requires large sample size (empirical testing)
- Permutation test: expensive
- From [Colas et al., 2019], use Welch's T-test!

Tests along training



- One can test differences at each evaluation step along training
- The above two algorithms are different most of the time, but not always

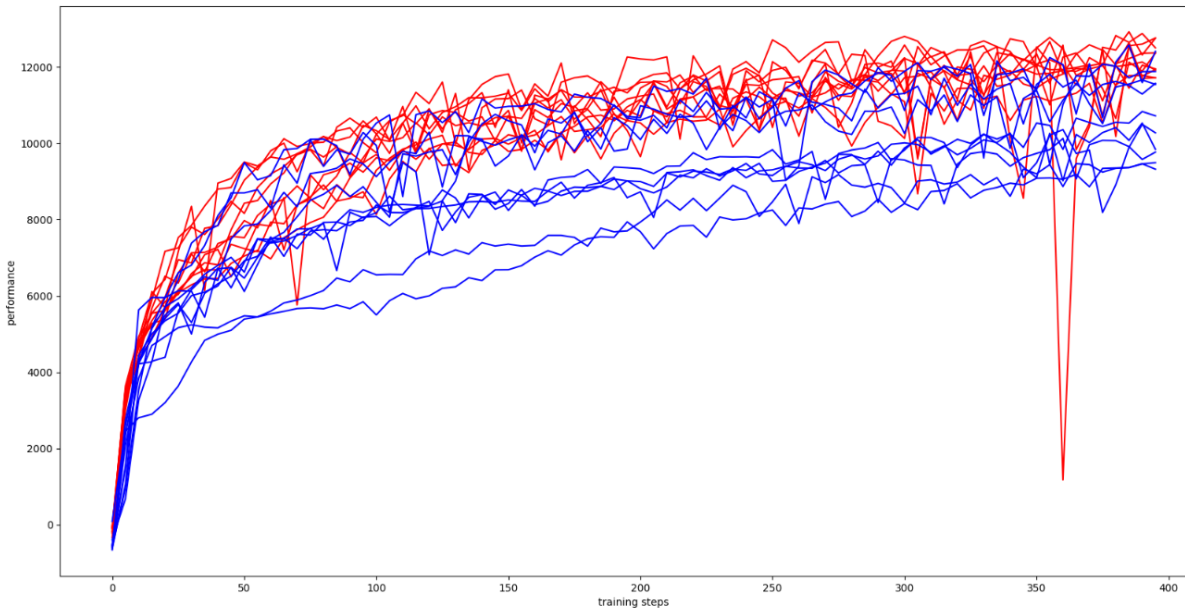
Number of seeds

- In general, 15 seeds is sufficient
- AdaStop: add one seed at a time, until the statistical difference is validated
- When comparing more than two algorithms, more statistical power is needed (Bonferroni correction)
- One may use different seeds for the agent, and the same seed for the environment

Colas, C., Sigaud, O., and Oudeyer, P.-Y. (2018) How many random seeds? statistical power analysis in deep reinforcement learning experiments. *arXiv preprint arXiv:1806.08295*

Mathieu, T., Della Vecchia, R., Shilova, A., de Medeiros, M. C., Kohler, H., Maillard, O.-A., and Preux, P. (2023) AdaStop: sequential testing for efficient and reliable comparisons of deep rl agents. *arXiv preprint arXiv:2306.10882*

Plotting



- Showing the mean/median is never enough (need info about variance)
- The standard deviation is representative only if the spread is Gaussian
- Rather take the $[0.1, 0.9]$ interval of values
- Or the 50% of values around the mean (Inter Quartile Mean, IQM)
- If less than 10 curves, plot them all

Summary

- Use Welch's T-test
- Use the mean rather than the median
- Whenever possible, use at least 15 seeds
- Give p-values, check for statistical power
- Select adapted plots
- When comparing more than two algorithms, add Bonferroni correction
- See `https://github.com/flowersteam/rl_stats`
- Lab: try the notebook

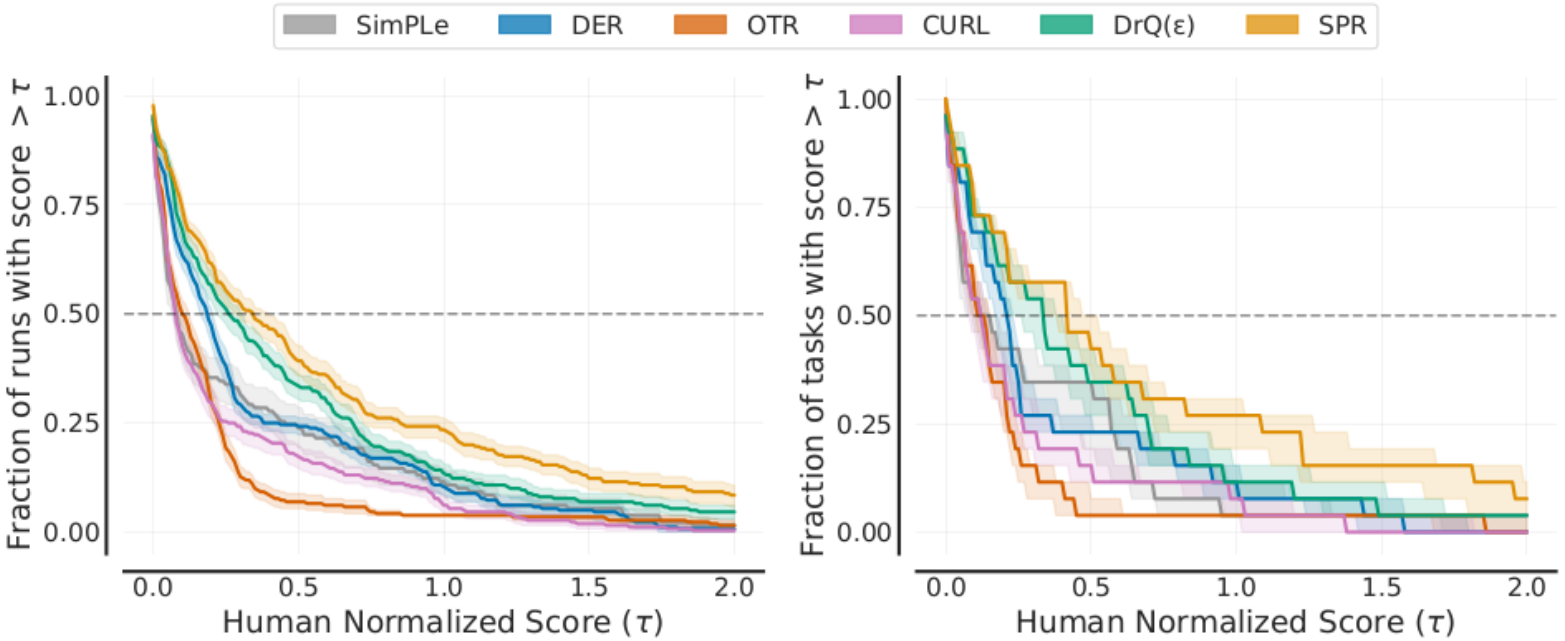
Performance comparisons

Good comparison practices

Desideratum	Current Evaluation Protocol	Our Recommendation
Uncertainty in aggregate performance	**Point estimates** • Ignore statistical uncertainty • Hinder *results reproducibility*	Interval estimates via **stratified bootstrap confidence intervals**
Variability in performance across tasks and runs	**Tables with mean scores per task** • Overwhelming beyond a few tasks • Standard deviations often omitted • Incomplete picture for multimodal and heavy-tailed distributions	**Performance profiles** (*score distributions*) • Show tail distribution of scores on combined runs across tasks • Allow qualitative comparisons • Easily read any score percentile
Aggregate metrics for summarizing performance across tasks	**Mean** • Often dominated by performance on outlier tasks **Median** • Requires large number of runs to claim improvements • Poor indicator of overall performance: zero scores on nearly half the tasks do not affect it	**Interquartile Mean** (IQM) across all runs • Performance on middle 50% of combined runs • Robust to outlier scores but more statistically efficient than median To show other aspects of performance gains, report average *probability of improvement* and *optimality gap*.

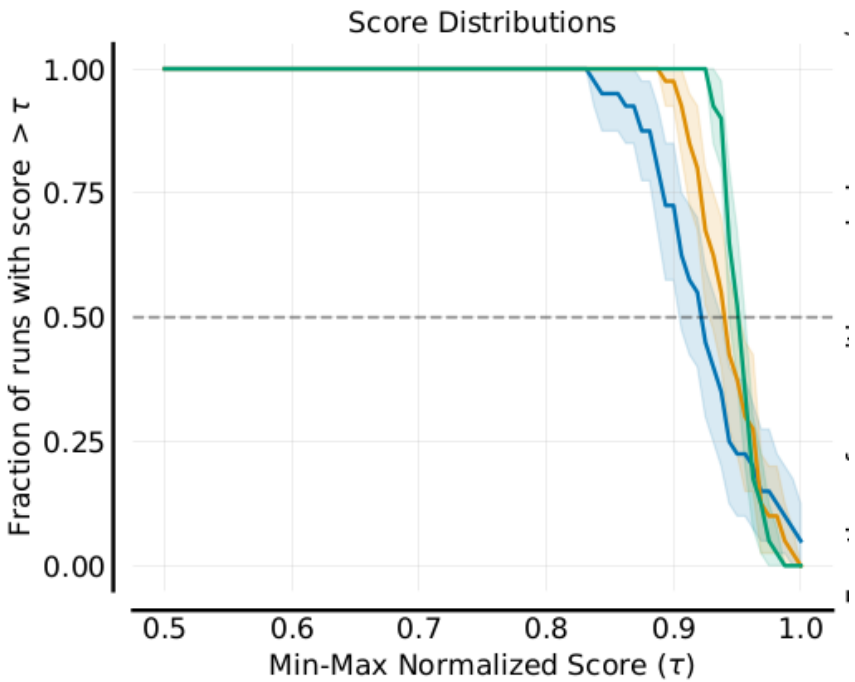
- The authors propose 3 tools to better evaluate algorithms: stratified bootstrap confidence intervals, performance profiles, aggregate metrics
- Focus on performance profiles

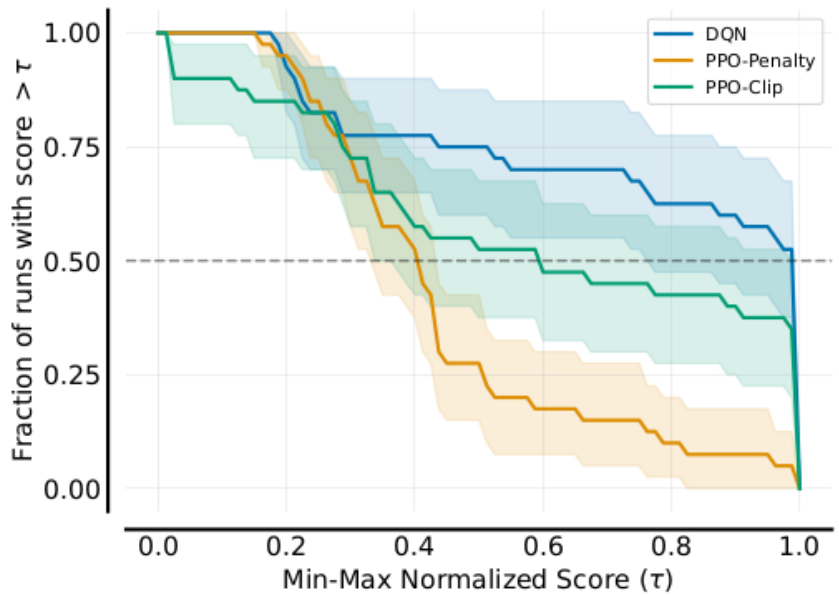
Performance profiles: definition



- Normalize the scores:
  - 0 for some min (e.g. the seen min or the known environment min)
  - 1 for a relevant max (e.g. human performance or the seen max)
- How much % or a number of runs reach performance over the x-axis value?
- An algorithm statistically dominates another if its performance profile is strictly above the other
- The more environment, the less often it happens

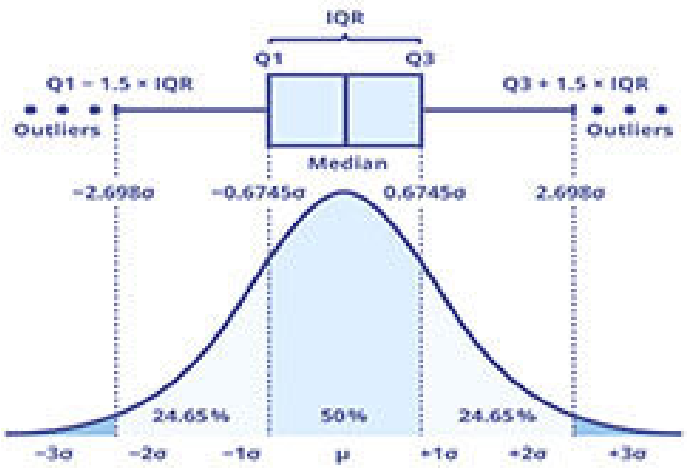
Performance profiles: Rules of thumb





- The worst algorithm should show 100% at 0 and decrease immediately
- The best algorithm should reach 0% shortly before the end of the interval
- This maximizes readability

Interquartile Mean (IQM)



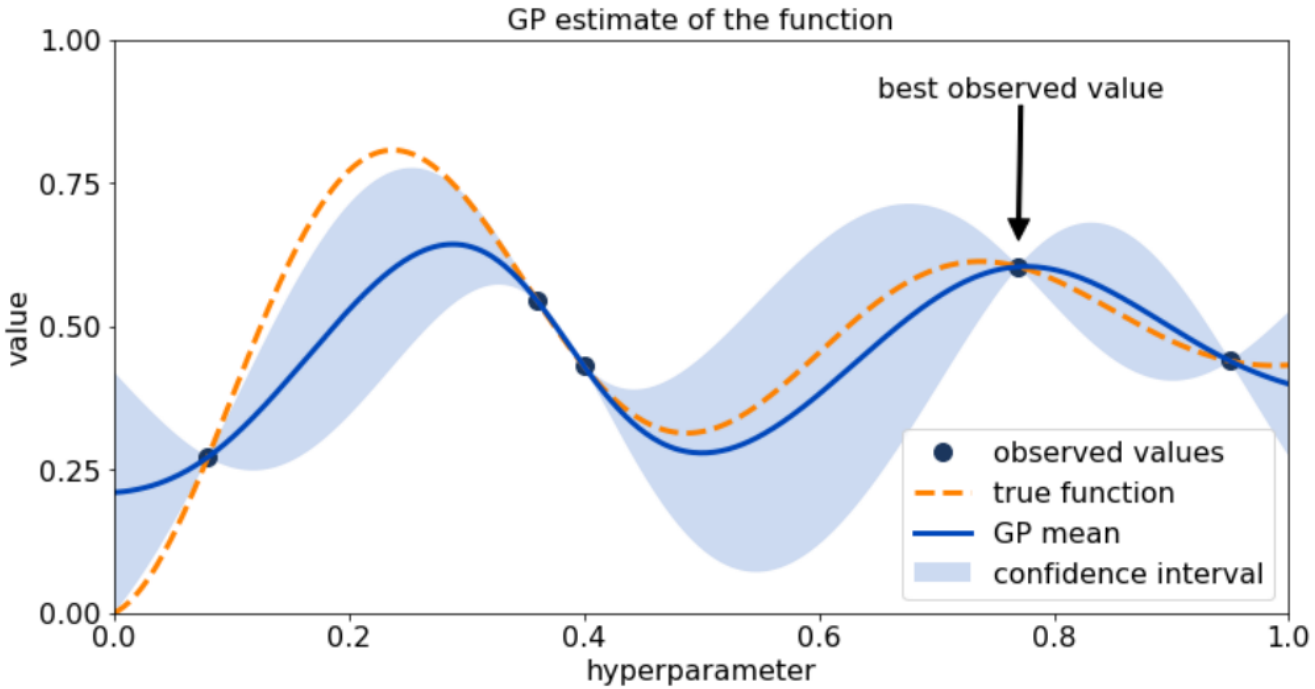
- Remove the 25% worst and the 25% best scores, show the mean and the interval
- Better than median (would be biased if close to 50% runs get 0 value)
- Better than mean: less sensitive to outliers
- Better at finding a true difference (empirical study)

Conclusion

- Other measures: probability of improvement, optimality gap, ...
- Try it at `https://github.com/google-research/rliable`
- An easy to use notebook, with an ATARI example

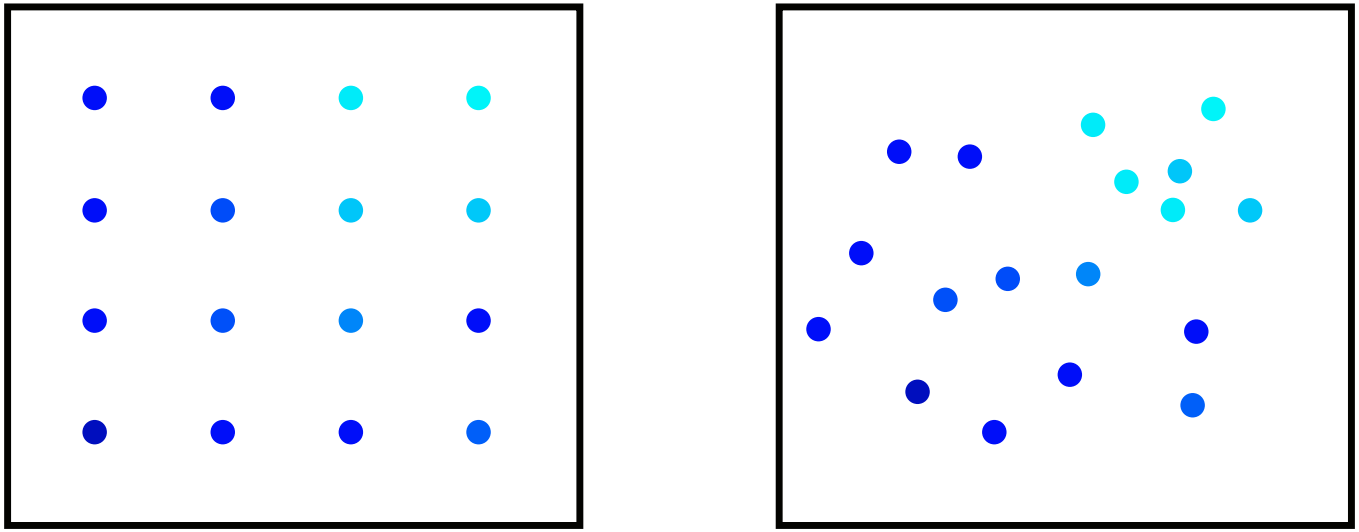
Tuning methods

Bayesian optimization



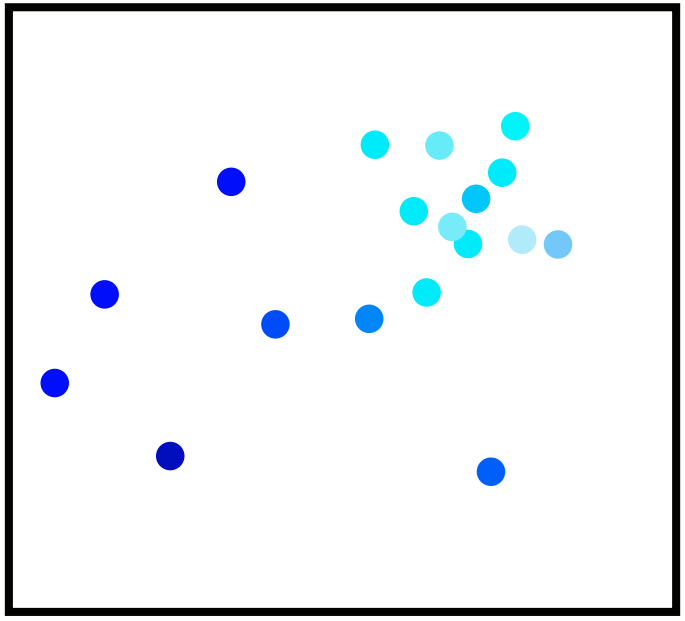
- The choice of h_{t+1} depends on information gathered with h_t
- No room for parallel search
- Key example: optuna

Grid search or random search



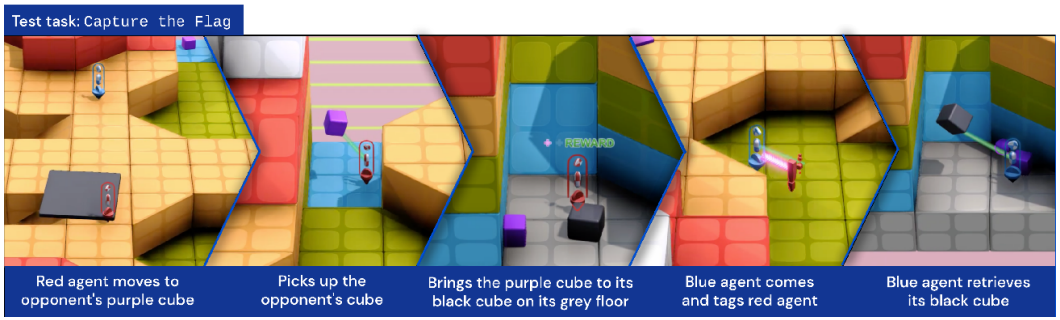
- All evaluations are independent and can be performed in parallel
- Random search is generally better than grid search
- But no mechanism to insist on promising areas

Evolutionary methods



- The best of both worlds:
  - Parallel search at each generation
  - Convergence to sweet spots
- PBT uses an evolutionary approach
- But hyperparameter search is performed during the training of agents
- A more adaptive dynamics

Introduction to PBT



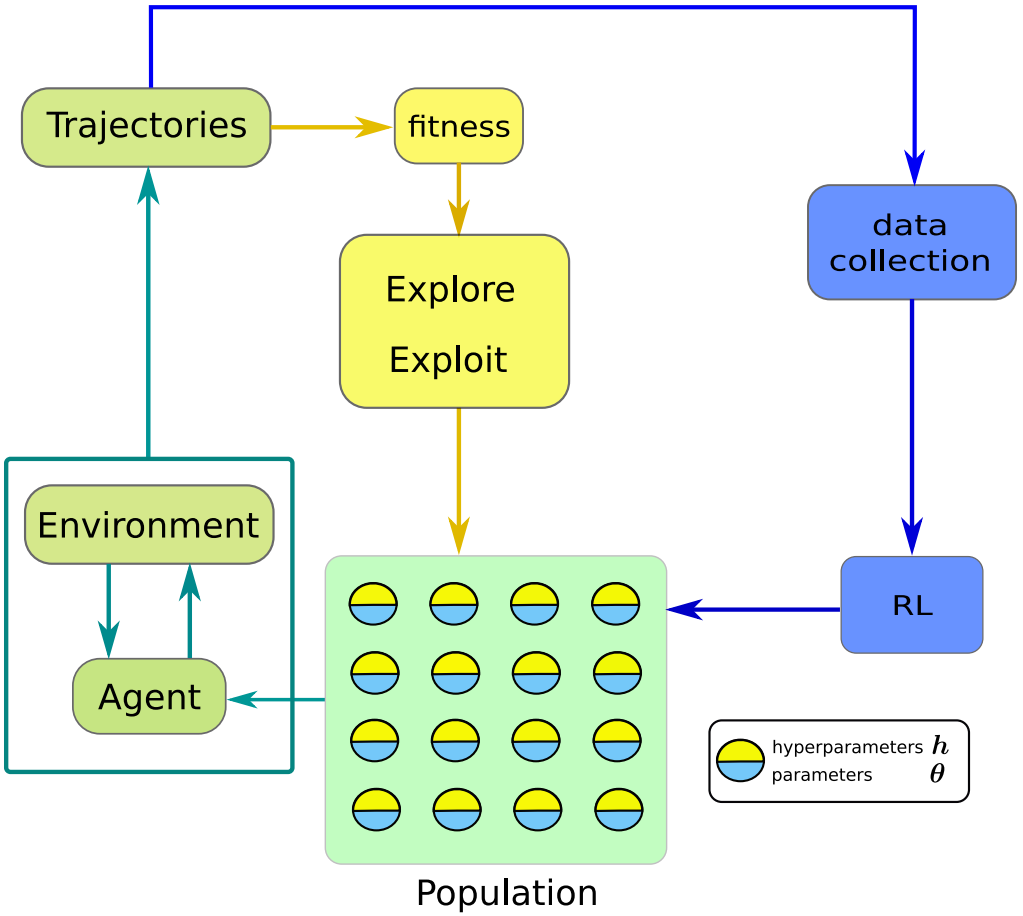
- Hyperparameter search is crucial in Deep RL
- Population-Based Training (PBT) provides an efficient solution to this problem
- It has been used in several notorious applications of Deep RL
- We note $\mathbf{h}$ the hyperparameter vector and $\boldsymbol{\theta}$ the parameter vector

Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. (2017) Population-based training of neural networks. *arXiv preprint arXiv:1711.09846*

Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., Ruderman, A., et al. (2019) Human-level performance in 3D multiplayer games with population-based reinforcement learning. *Science*, 364(6443), 859–865

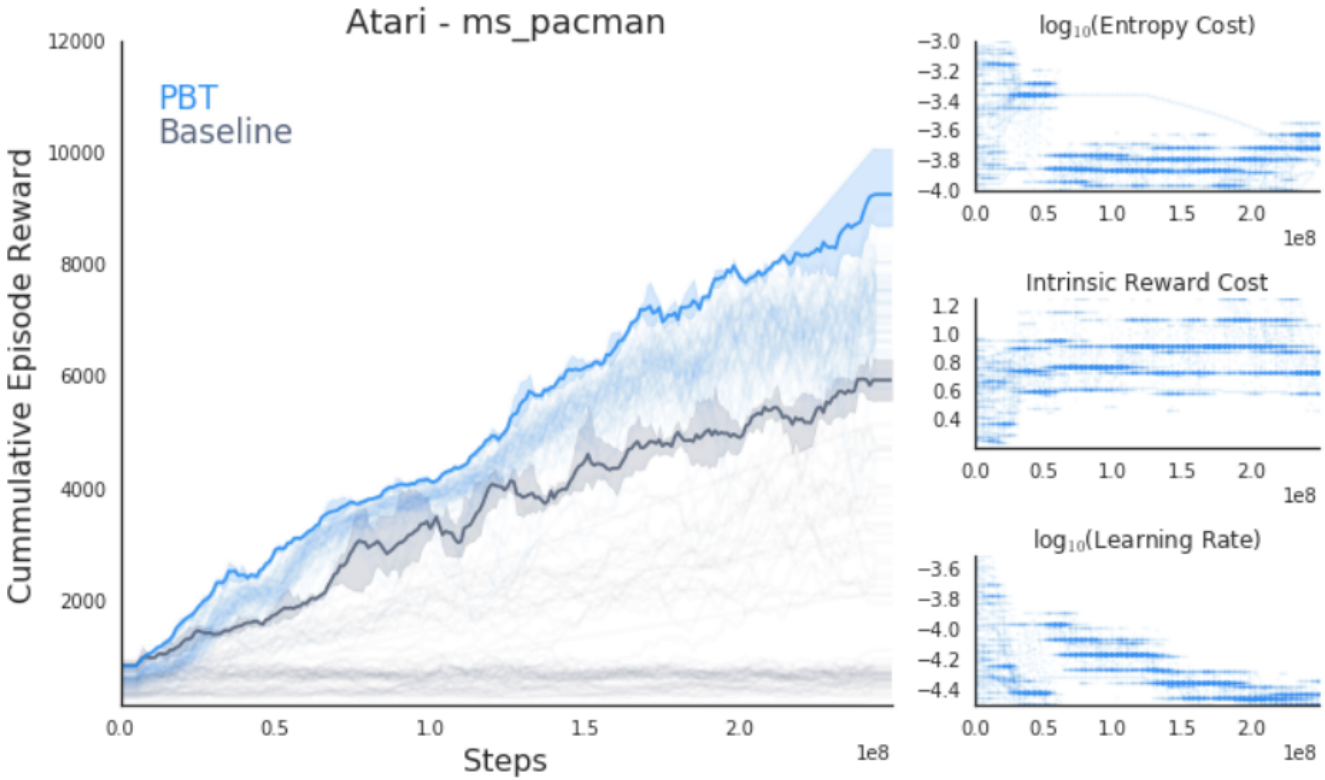
Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., et al. (2021) Open-ended learning leads to generally capable agents. *arXiv preprint arXiv:2107.12808*

The PBT architecture



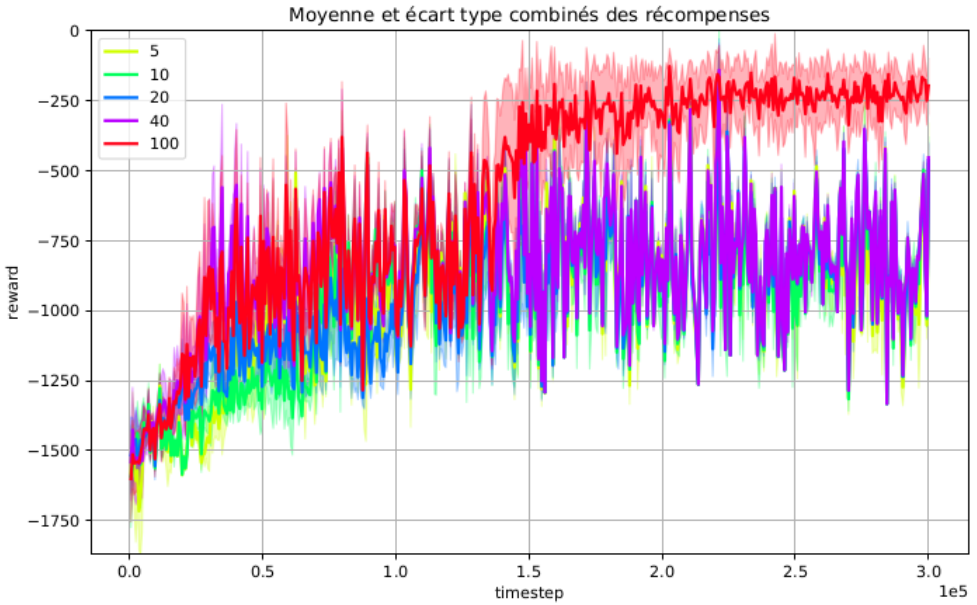
- The PBT approach is applied to more than RL (GAN, supervised learning...) but here we focus on RL.
- The variation-selection operators (**Exploit**, **Explore**) are applied to both parameters and hyperparameters

Variation of hyperparameters over time



- We can see the $\mathbf{h}$ drifting over time
- Does not converge to a single value

A PBT project



- Tested on pendulum with various population sizes
- Not convincing with a small population
- A larger population can find the right hyper-parameters
- The evolution part is very naive, could be much improved

Fair tuning

Methodological requirements

- Tuning should be automatized to remove human biases
- Competitors should be allocated the same tuning budget
- Automatic tuning should start from a similar initial performance

Practical methodology

- The seed is NOT a tunable hyperparameter. Random noise must keep random.
- Tune hyper-parameters of competitors by hand to reach a similar start performance
- Start an automated tuning framework (optuna) from there
- Define a time budget or computational budget
- Tune competitors with the allocated budget
- Generate enough performance results to run statistical tests
- Use baselines to contextualize performance: random, oracle
- Using appropriate tests, conclude about performance disparities

Evaluating a fully specified algorithm

- Separate training and evaluation
- Reporting mean or median performance is not enough
- Do not report standard errors, based on wrong Gaussian assumption
- Use steps rather than episodes (episodes are of varying length)
- Do not run an incomplete experiment due to insufficient resources: calibrate your experiment depending on your resources
- E.g.: Choose to study the speed of early learning or the optimal performance (depending on your budget)

Sensitivity curves

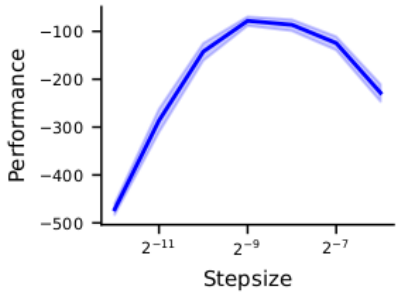


Figure 9: A good sensitivity curve that captures a wide range of the variable of interest and illustrates that performance changes smoothly as the hyperparameter changes.

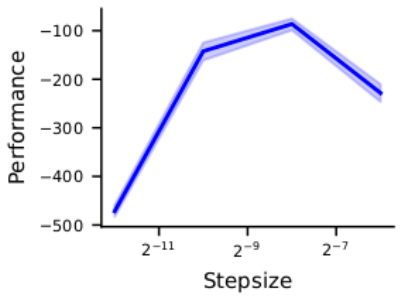


Figure 10: A sensitivity curve where the range of tested values may be too wide, instead of being focused in the region of interest. We lose some information around the peak performance and the algorithm appears quite sensitive. This sensitivity might be an artefact of the plot—testing insufficiently many values—rather than a property of the algorithm.

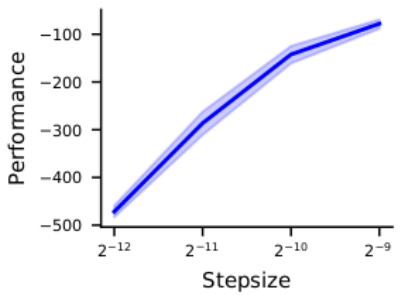


Figure 11: A sensitivity curve where we potentially missed the best performance. The best performing hyperparameter may be outside the range or may be the endpoint of the range, but we cannot tell with the presented information.

- Use parameter sensitivity plots to find adequate parameter ranges

Patterson, A., Neumann, S., White, M., and White, A. (2023) Empirical design in reinforcement learning. *arXiv preprint arXiv:2304.01315*

Sensitivity regions

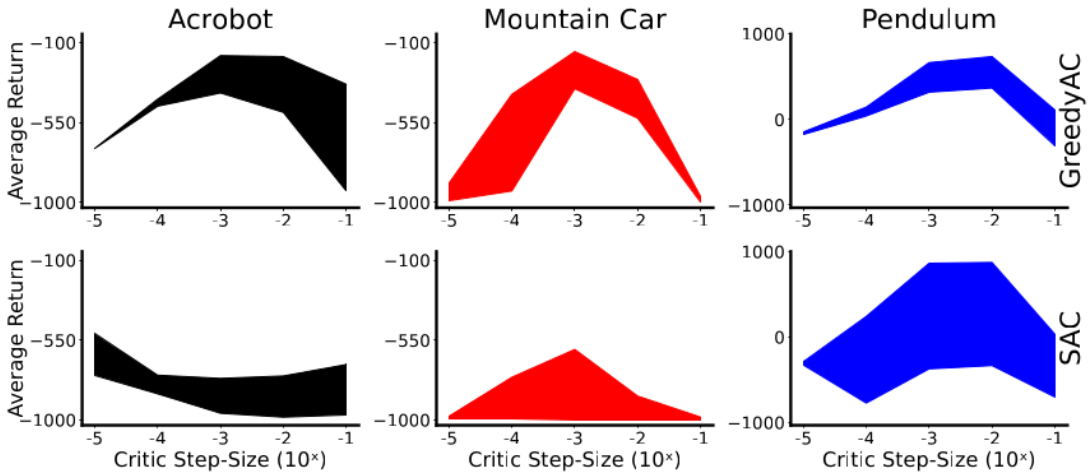


Figure 4: A **sensitivity region** plot for entropy, for GreedyAC (top row) and SAC (bottom row) in the continuous action problems.

- More information when sampling many hyper-parameter sets
- Hyper-parameters with narrow sensitivity at peak performance should be set first (?)

Neumann, S., Lim, S., Joseph, A. G., Pan, Y., White, A., and White, M. (2023) Greedy actor-critic: A new conditional cross-entropy method for policy improvement. In *The Eleventh International Conference on Learning Representations*

A first example

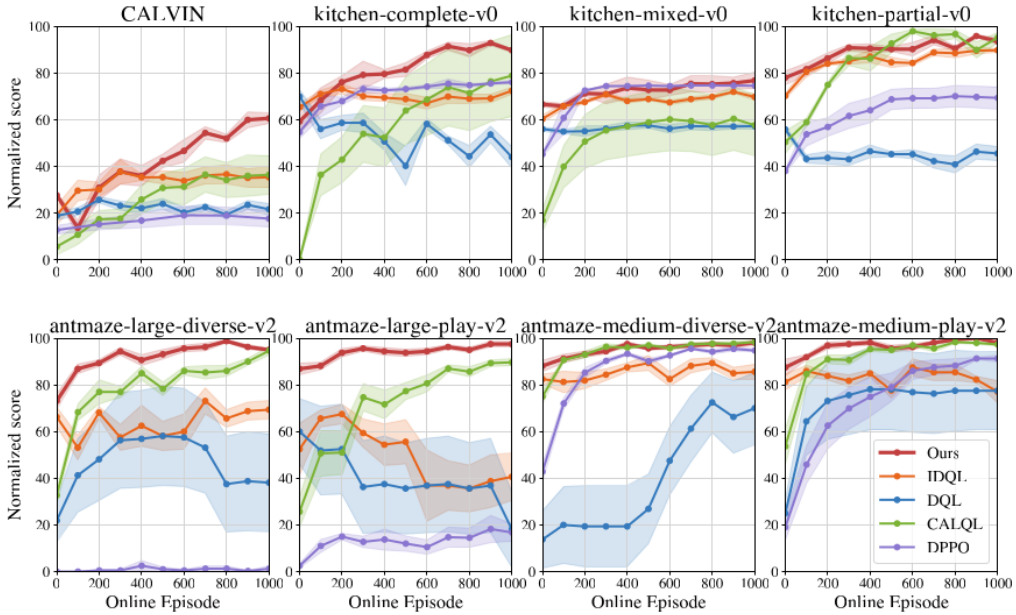


Figure 3: Learning curves of online fine-tuning with various methods. Observe that ***PA-RL*** + Cal-QL (red) largely always dominates or attains similar performance to the next best method. Other methods for fine-tuning diffusion policies (IDQL, DQL, DPPO) are a bit unstable, and perform substantially worse. Since DPPO is substantially more data inefficient, we plot it with different x-axis units: for kitchen each unit is 500 episodes (axis goes from 0 to 500k), for antmaze each unit is 100 episodes (axis goes from 0 to 100k) and for calvin each unit is 10 episodes (axis goes until 10k).

- How many seeds?
- What is the measure of variability?
- Episodes rather than steps

Mark, M. S., Gao, T., Sampaio, G. G., Srirama, M. K., Sharma, A., Finn, C., and Kumar, A. (2024) Policy agnostic RL: Offline RL and online RL fine-tuning of any class and backbone. *arXiv preprint arXiv:2412.06685*

A better example

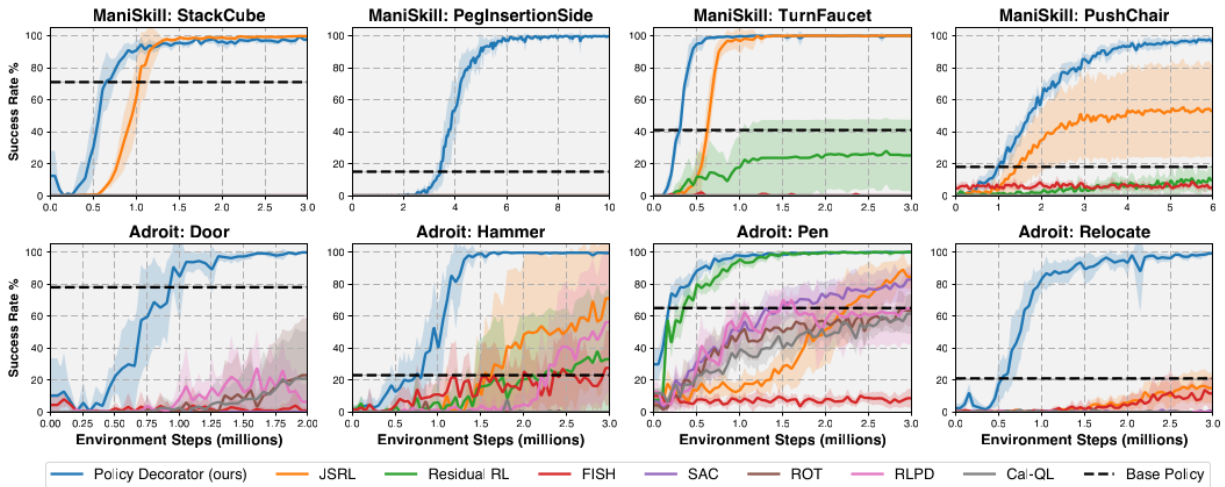
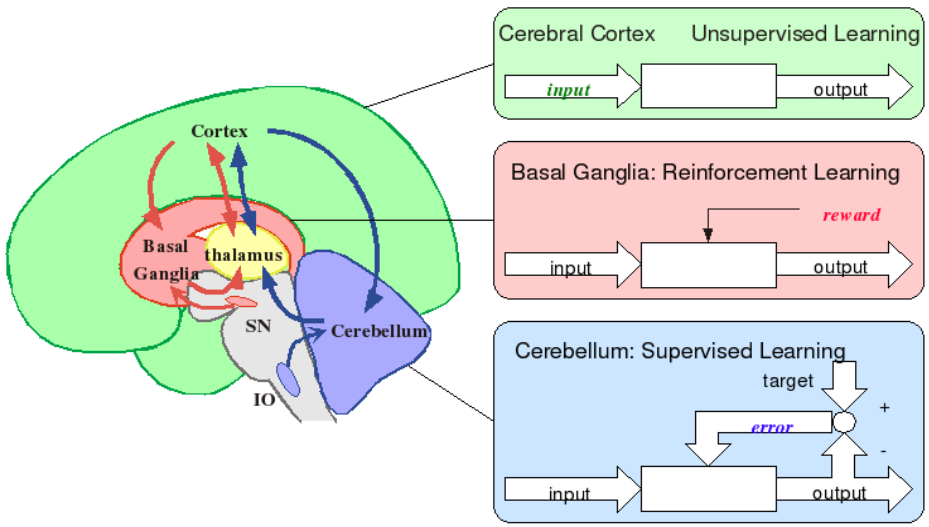


Figure 6: **Results (with Behavior Transformer):** During training, we evaluate the agent for 50 episodes every 50K environment steps. The curves depict the evaluation success rates averaged over ten seeds, and the shaded areas represent standard deviations. Our method consistently improves the base policy and outperforms all other baselines.

- 10 seeds, steps rather than episodes, standard deviation, conclusion given...

Biological counterpart

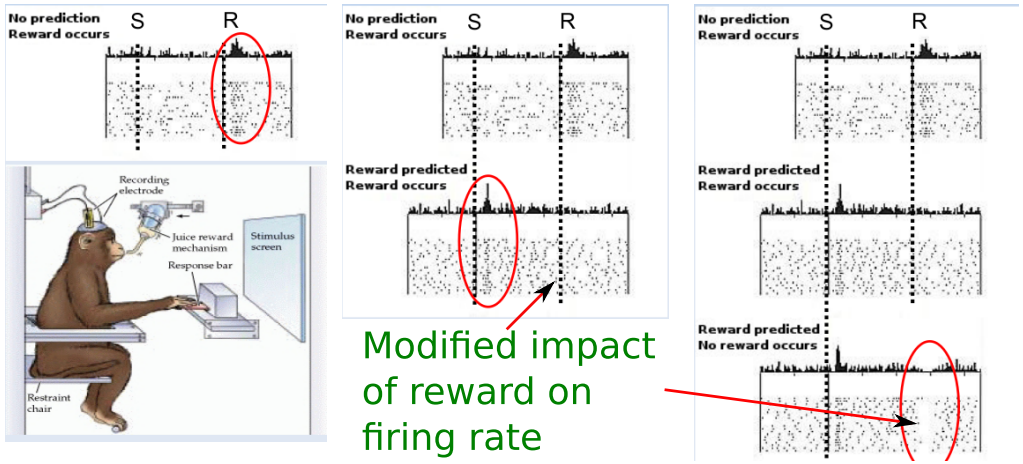
Overview: the brain



- Assumption: model-free RL takes place in basal ganglia
- A place with many dopaminergic neurons

Doya, K. (2000) Complementary roles of basal ganglia and cerebellum in learning and motor control. *Current Opinion in Neurobiology*, 10:732–739

TD rule in dopaminergic neurons

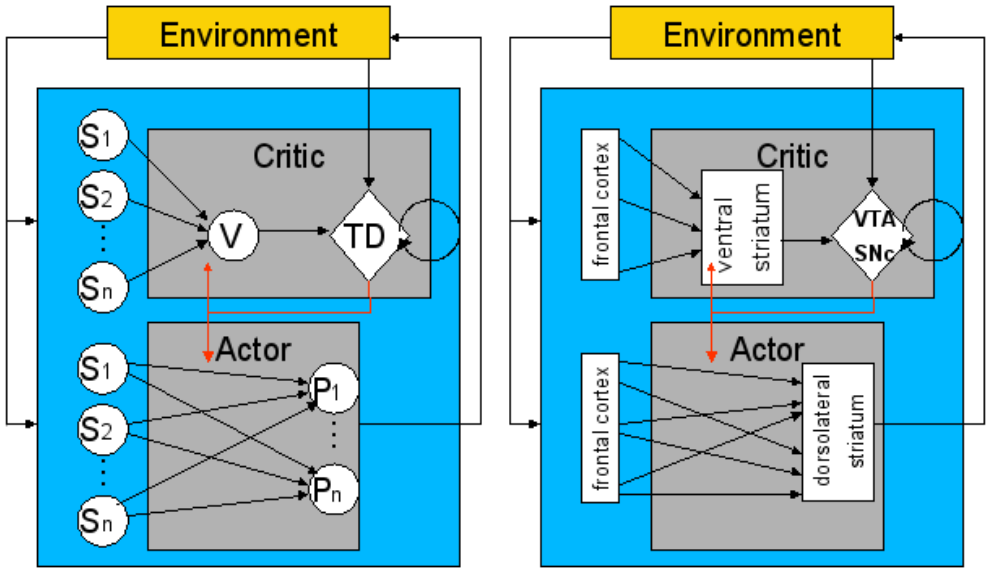


$$V(s_t) \leftarrow V(s_t) + \alpha[r_{t+1} + \gamma V(s_{t+1}) - V(s_t)]$$

- The firing rates of dopaminergic neurons reflect the TD error (or RPE)

Schultz, W., Dayan, P., and Montague, P. R. (1997) A neural substrate of prediction and reward. *Science*, 275(5306):1593–1599

A potential architecture

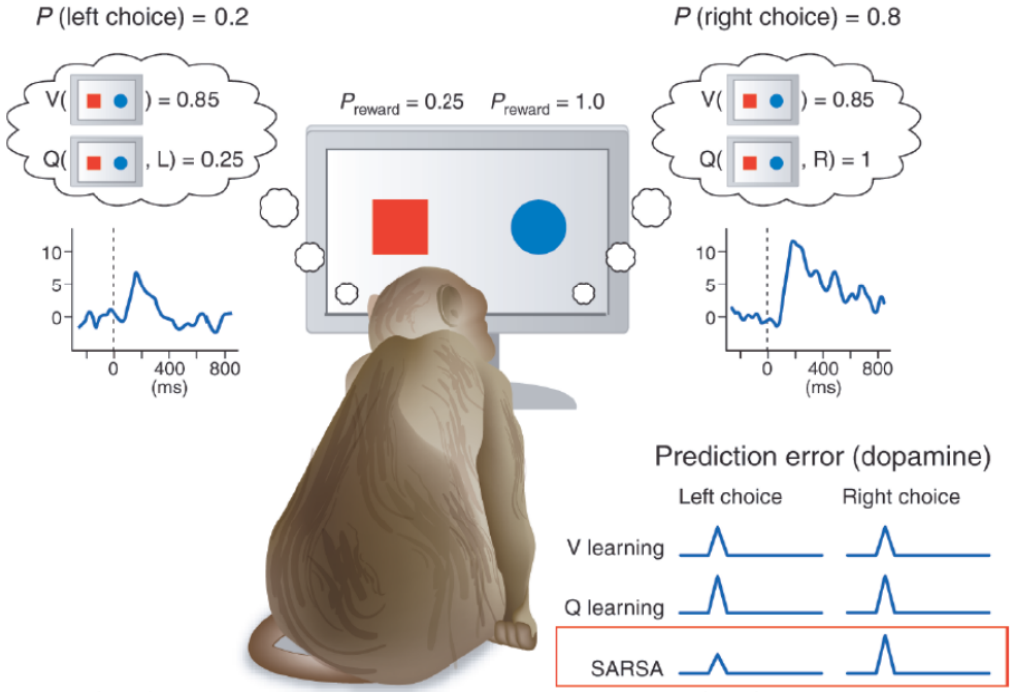


From Takahashi, Schoenbaum and Niv, Frontiers in Neurosciences, pp. 86-97, july 2008

- Basal ganglia: ventral, dorsal and dorsolateral striatum
- The actor would be the dorsolateral striatum and the critic the ventral striatum
- Even more sophisticated views have emerged
- Question: which algorithm could it be?

Takahashi, Y., Schoenbaum, G., and Niv, Y. (2008) Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model. *Frontiers in neuroscience*, 2:282

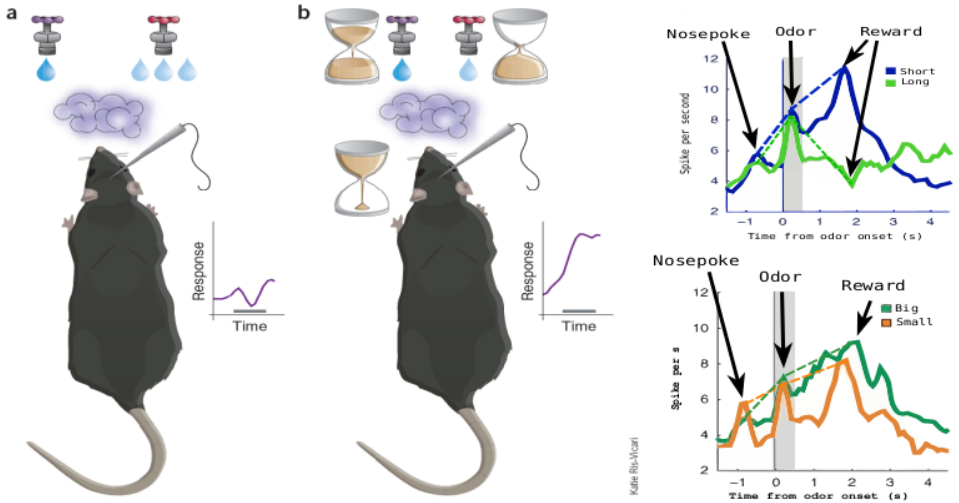
In favor of SARSA



- If basal ganglia perform TD learning, which algorithm do they use?
- With SARSA: $\delta_t = r_{t+1} + \gamma Q(s_{t+1}, a_{t+1}) - Q(s_t, a_t)$
- With Q-LEARNING: $\delta_t = r_{t+1} + \max_a \gamma Q(s_{t+1}, a) - Q(s_t, a_t)$
- Does the RPE depend on the next action?
- According to Morris et al.'s experiments, yes

Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006) Midbrain dopamine neurons encode decisions for future action. *Nature neuroscience*, 9(8):1057–63

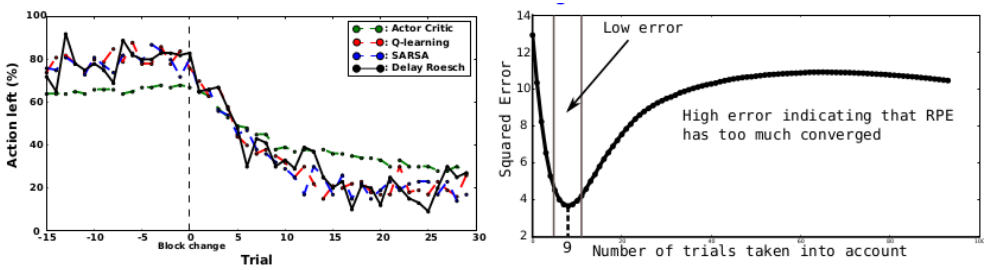
In favor of Q-LEARNING



- Does the RPE depend on the next action?
- According to Roesch et al.'s experiments, no

Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007) Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards. *Nature neuroscience*, 10(12):1615–24

Computational study



Left: best reproduction of the behavior of the rats obtained with the different algorithms (illustrated for the delay case).
Right: fitting error when comparing $aRPE_{QL} + b$ and DA activity (recording during the delay case), in function of the number of trials taken into account.

- A computational studied showed that neither Q-LEARNING nor SARSA do fit well
- Could be actor-critic?
- More consistent with architecture-oriented knowledge
- Rather, dopamine seems to encode for both RPE and value

Bellot, J., Sigaud, O., Roesch, M. R., Schoenbaum, G., Girard, B., and Khamassi, M. (2012) Dopamine neurons activity in a multi-choice task: reward prediction error or value function? In *Proceedings of the French Computational Neuroscience NeuroComp'12 workshop*, pages 1–7

Bellot, J., Khamassi, M., Sigaud, O., and Girard, B. (2013) Which temporal difference learning algorithm best reproduces dopamine activity in a multi-choice task? *BMC Neuroscience*, 14:1–2

Any question?

Send mail to: Olivier.Sigaud@isir.upmc.fr

Agarwal, R., Schwarzer, M., Castro, P. S., Courville, A. C., and Bellemare, M. (2021).

Deep reinforcement learning at the edge of the statistical precipice.
Advances in neural information processing systems, 34:29304–29320.

Bellot, J., Khamassi, M., Sigaud, O., and Girard, B. (2013).

Which temporal difference learning algorithm best reproduces dopamine activity in a multi-choice task?
BMC Neuroscience, 14:1–2.

Bellot, J., Sigaud, O., Roesch, M. R., Schoenbaum, G., Girard, B., and Khamassi, M. (2012).

Dopamine neurons activity in a multi-choice task: reward prediction error or value function?
In *Proceedings of the French Computational Neuroscience NeuroComp'12 workshop*, pages 1–7.

Bouthillier, X., Laurent, C., and Vincent, P. (2019).

Unreproducible research is reproducible.
In *International Conference on Machine Learning*, pages 725–734. PMLR.

Colas, C., Sigaud, O., and Oudeyer, P.-Y. (2018).

How many random seeds? statistical power analysis in deep reinforcement learning experiments.
arXiv preprint arXiv:1806.08295.

Colas, C., Sigaud, O., and Oudeyer, P.-Y. (2019).

A hitchhiker's guide to statistical comparisons of reinforcement learning algorithms.
arXiv preprint arXiv:1904.06979.

Doya, K. (2000).

Complementary roles of basal ganglia and cerebellum in learning and motor control.
Current Opinion in Neurobiology, 10:732–739.

Henderson, P., Islam, R., Bachman, P., Pineau, J., Precup, D., and Meger, D. (2018).

Deep reinforcement learning that matters.
In McIlraith, S. A. and Weinberger, K. Q., editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 3207–3214. AAAI Press.


Jaderberg, M., Czarnecki, W. M., Dunning, I., Marris, L., Lever, G., Castaneda, A. G., Beattie, C., Rabinowitz, N. C., Morcos, A. S., Ruderman, A., et al. (2019).
Human-level performance in 3d multiplayer games with population-based reinforcement learning.
Science, 364(6443):859–865.


Jaderberg, M., Dalibard, V., Osindero, S., Czarnecki, W. M., Donahue, J., Razavi, A., Vinyals, O., Green, T., Dunning, I., Simonyan, K., et al. (2017).
Population-based training of neural networks.
arXiv preprint arXiv:1711.09846.


Mark, M. S., Gao, T., Sampaio, G. G., Srirama, M. K., Sharma, A., Finn, C., and Kumar, A. (2024).
Policy agnostic RL: Offline RL and online RL fine-tuning of any class and backbone.
arXiv preprint arXiv:2412.06685.


Mathieu, T., Della Vecchia, R., Shilova, A., de Medeiros, M. C., Kohler, H., Maillard, O.-A., and Preux, P. (2023).
AdaStop: sequential testing for efficient and reliable comparisons of deep rl agents.
arXiv preprint arXiv:2306.10882.


Morris, G., Nevet, A., Arkadir, D., Vaadia, E., and Bergman, H. (2006).
Midbrain dopamine neurons encode decisions for future action.
Nature neuroscience, 9(8):1057–63.


Neumann, S., Lim, S., Joseph, A. G., Pan, Y., White, A., and White, M. (2023).
Greedy actor-critic: A new conditional cross-entropy method for policy improvement.
In *The Eleventh International Conference on Learning Representations*.


Patterson, A., Neumann, S., White, M., and White, A. (2023).
Empirical design in reinforcement learning.
arXiv preprint arXiv:2304.01315.


Roesch, M. R., Calu, D. J., and Schoenbaum, G. (2007).
Dopamine neurons encode the better option in rats deciding between differently delayed or sized rewards.
Nature neuroscience, 10(12):1615–24.

Schultz, W., Dayan, P., and Montague, P. R. (1997).

A neural substrate of prediction and reward.
Science, 275(5306):1593–1599.

Stooke, A., Mahajan, A., Barros, C., Deck, C., Bauer, J., Sygnowski, J., Trebacz, M., Jaderberg, M., Mathieu, M., et al. (2021).

Open-ended learning leads to generally capable agents.
arXiv preprint arXiv:2107.12808.

Takahashi, Y., Schoenbaum, G., and Niv, Y. (2008).

Silencing the critics: understanding the effects of cocaine sensitization on dorsolateral and ventral striatum in the context of an actor/critic model.
Frontiers in neuroscience, 2:282.