Reliable RL: An Algorithmic Perspective

Logan Engstrom

(with A. Ilyas*, S. Santurkar, D. Tsipras, F. Janoos, L. Rudolph, and A. Madry)





Reinforcement Learning (RL)



Reinforcement Learning (RL)





Reinforcement Learning (RL)



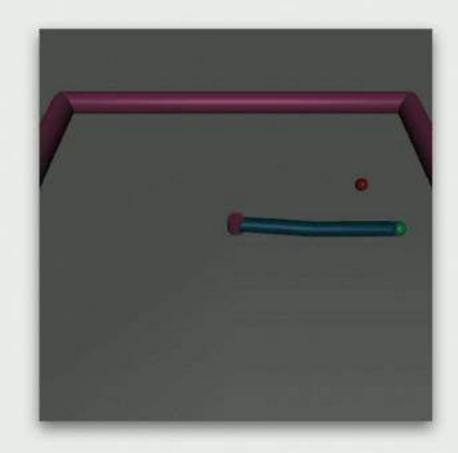
Is RL "real-world-ready"?

Is RL "real-world-ready"?

(Spoiler: No)

Deep RL is unreliable even in simple settings...





How do we get reliable RL?

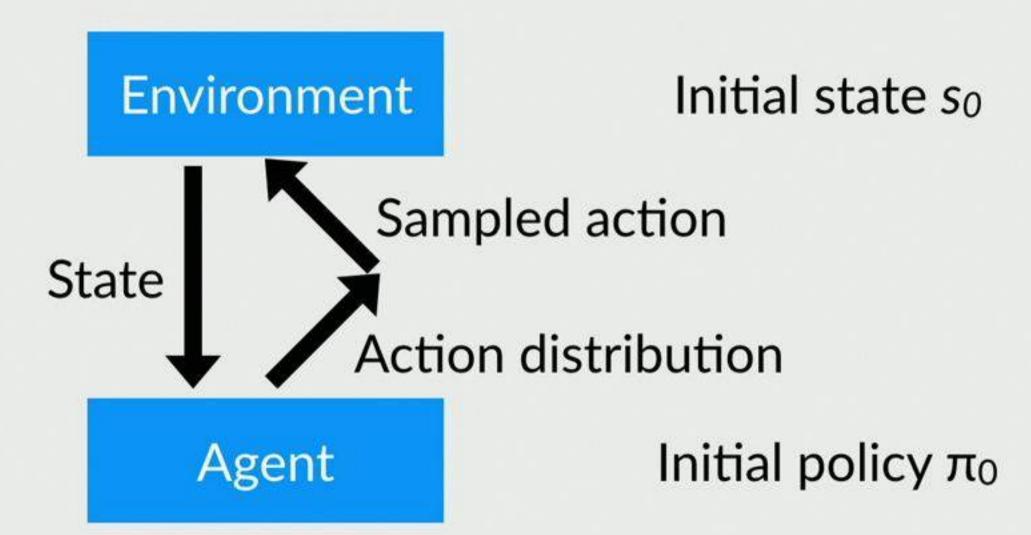
An algorithmic understanding of modern RL methods

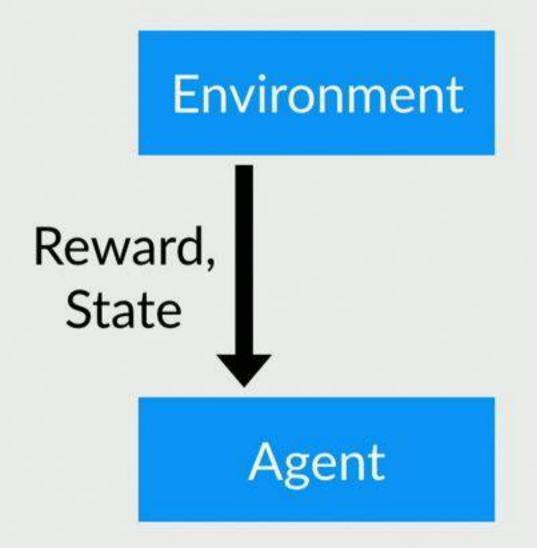
Environment

Initial state so

Agent

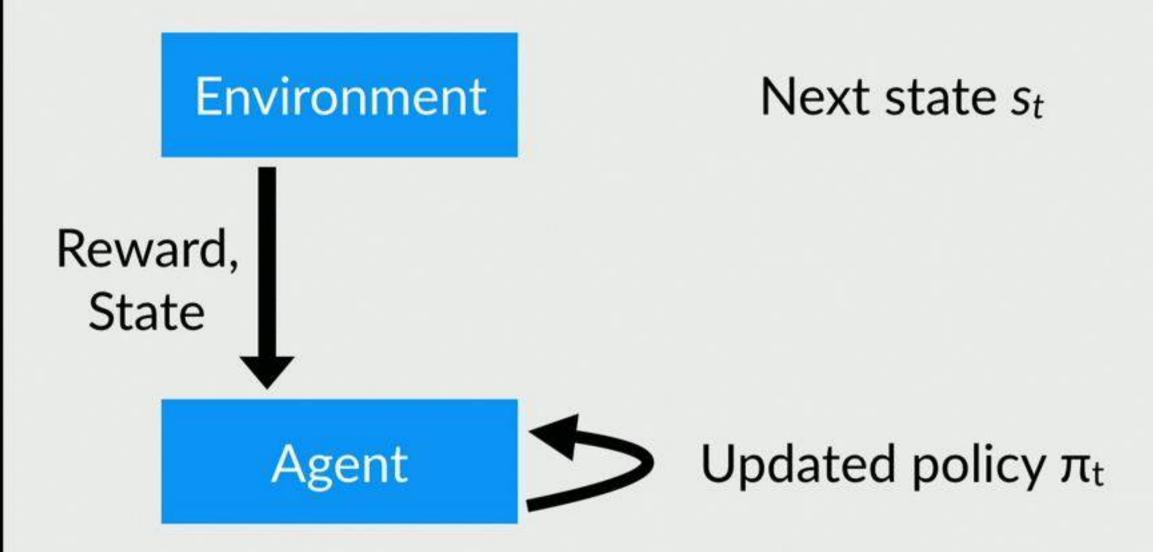
Initial policy π_0

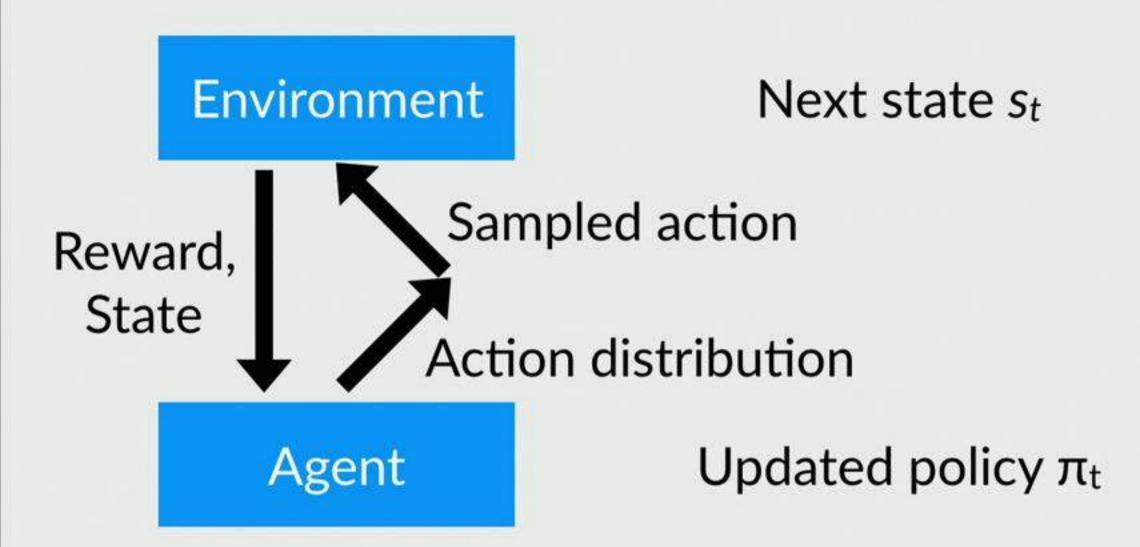




Next state s_t

Initial policy π_0





Goal: Maximize expected total reward

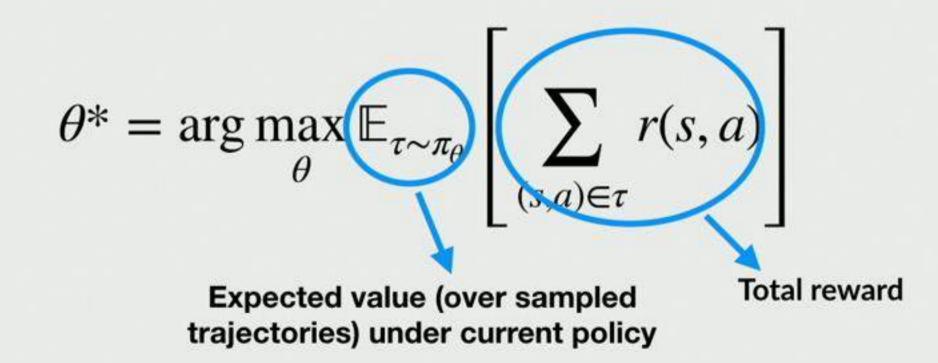
(over trajectories)

Policy Gradient Algorithms

Key Principle: View our goal as an optimization problem

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{(s,a) \in \tau} r(s,a) \right]$$

Key Principle: View our goal as an optimization problem



Key Principle: View our goal as an optimization problem

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{(s,a) \in \tau} r(s,a) \right]$$

Method of choice: gradient descent

Key Principle: View our goal as an optimization problem

$$\theta^* = \arg\max_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{(s,a) \in \tau} r(s,a) \right]$$

No gradient access

Method of choice: gradient descent

Can we instead get a good estimate of the gradient?

$$\nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left| \sum_{(s,a) \in \tau} r(s,a) \right| = ???$$

Can we instead get a good estimate of the gradient?

$$\nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{(s,a) \in \tau} r(s,a) \right] = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[g(\tau) \right]$$

The Policy Gradient

Can we instead get a good estimate of the gradient?

$$\nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{(s,a) \in \tau} r(s,a) \right] = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[g(\tau) \right]$$

$$\bigotimes_{N}^{1} \sum_{\tau \sim \pi_{\theta}} \left[g(\tau) \right]$$

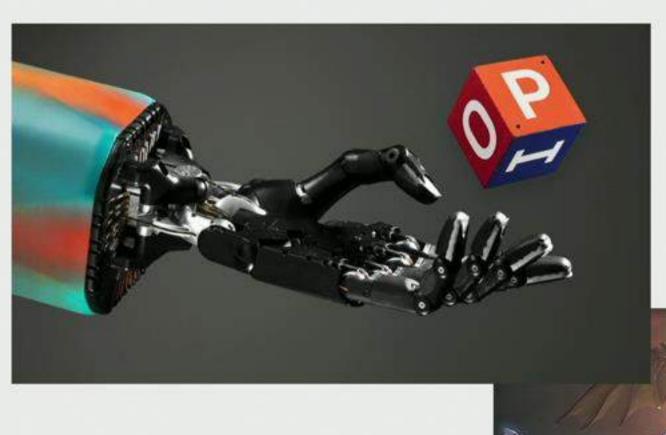
Can we instead get a good estimate of the gradient?

$$\nabla_{\theta} \mathbb{E}_{\tau \sim \pi_{\theta}} \left[\sum_{(s,a) \in \tau} r(s,a) \right] = \mathbb{E}_{\tau \sim \pi_{\theta}} \left[g(\tau) \right]$$

$$\bigotimes_{N}^{1} \sum_{\tau \sim \pi_{\theta}} \left[g(\tau) \right]$$

Then: use estimate in gradient descent!

Policy Gradient Successes

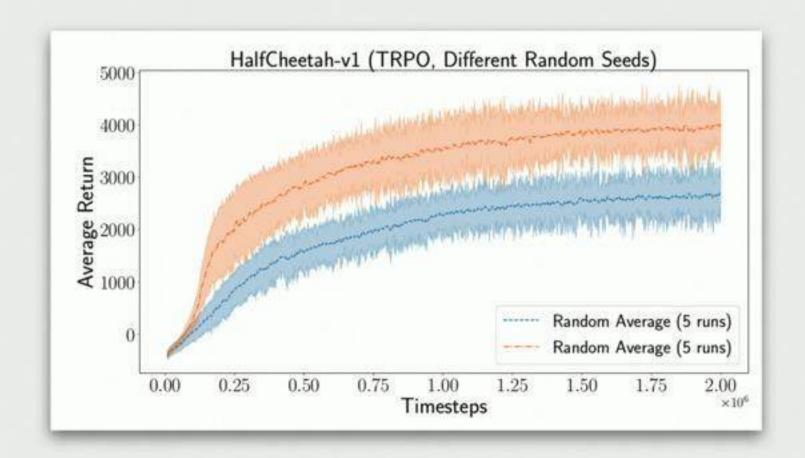






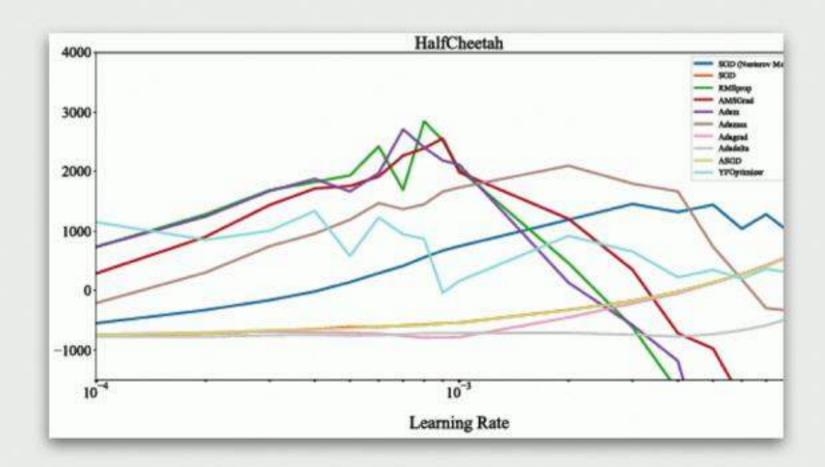
Deep RL can successfully solve tasks, but has...

Poor reliability over repeated runs



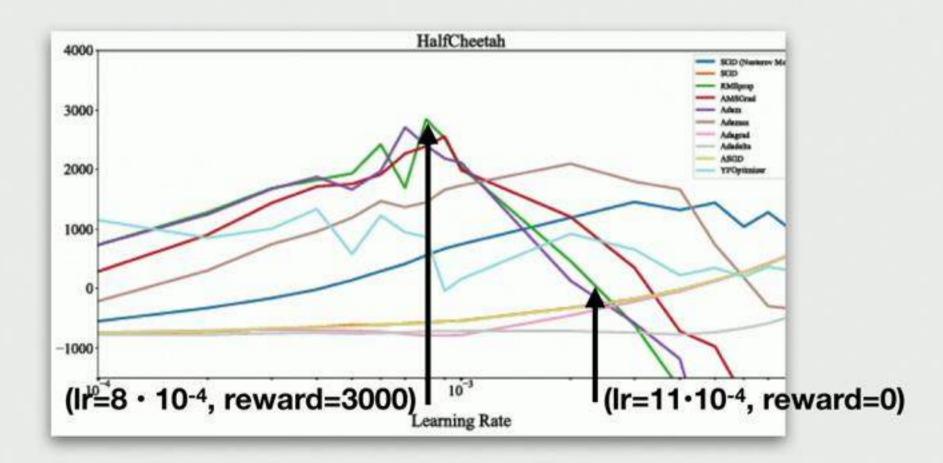
Deep RL can successfully solve tasks, but has...

- Poor reliability over repeated runs
- High sensitivity to hyperparameters



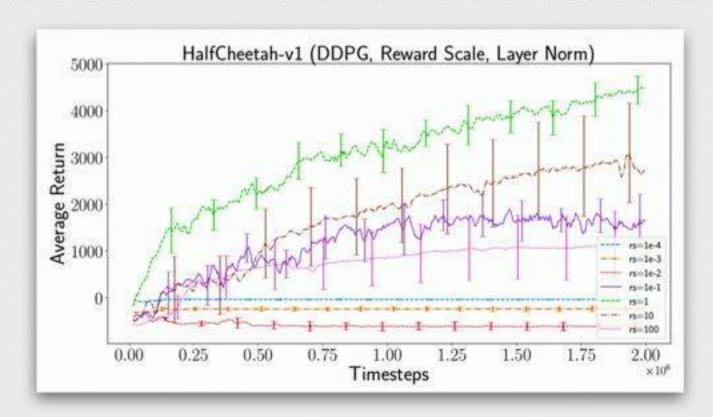
Deep RL can successfully solve tasks, but has...

- Poor reliability over repeated runs
- High sensitivity to hyperparameters



Deep RL can successfully solve tasks, but has...

- Poor reliability over repeated runs
- High sensitivity to hyperparameters
- Poor robustness to environmental artifacts



Deep RL can successfully solve tasks, but has...

- Poor reliability over repeated runs
- High sensitivity to hyperparameters
- Poor robustness to environmental artifacts

Notably, benchmarks don't reveal these problems

Deep RL can successfully solve tasks, but has...

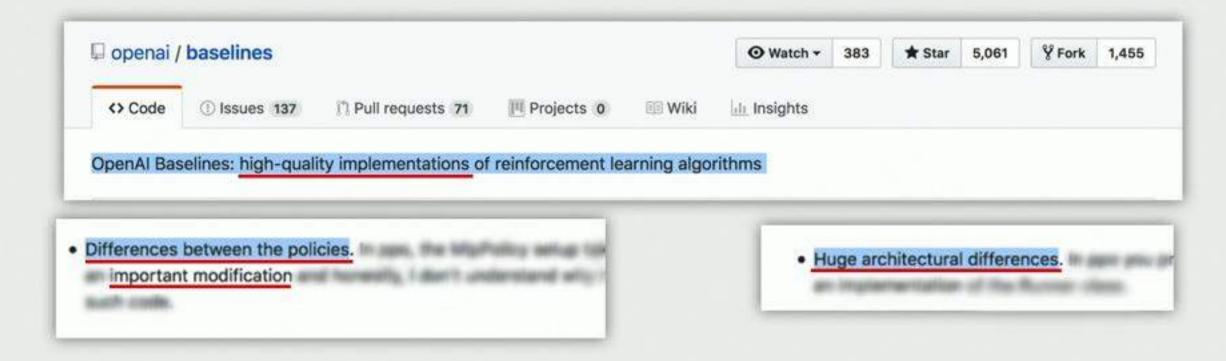
- Poor reliability over repeated runs
- High sensitivity to hyperparameters
- Poor robustness to environmental artifacts

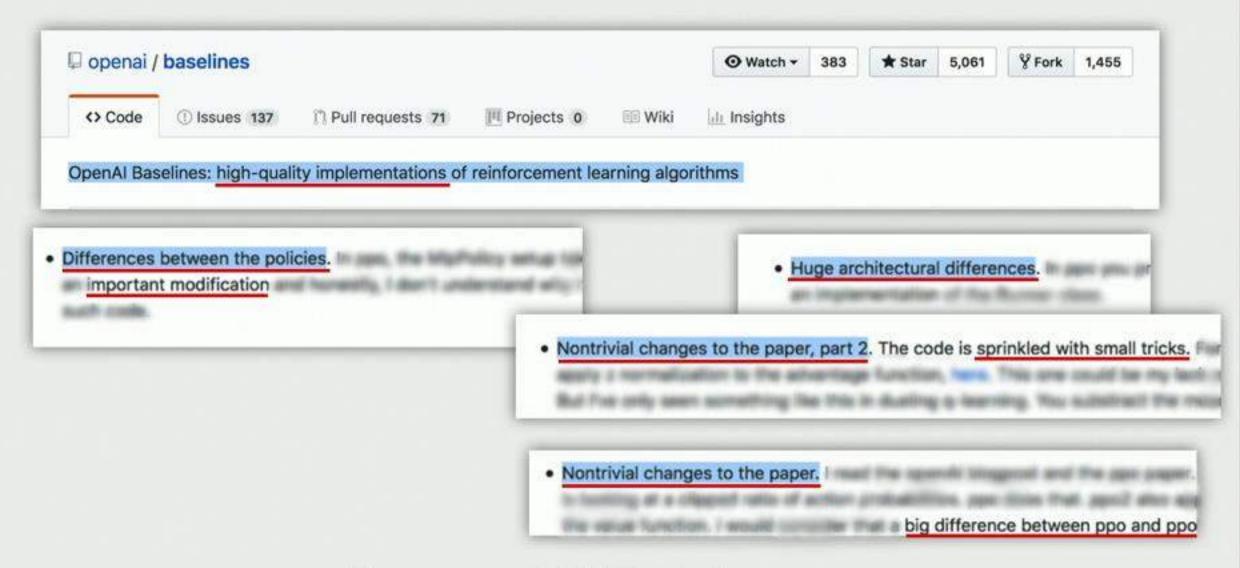
Notably, benchmarks don't reveal these problems

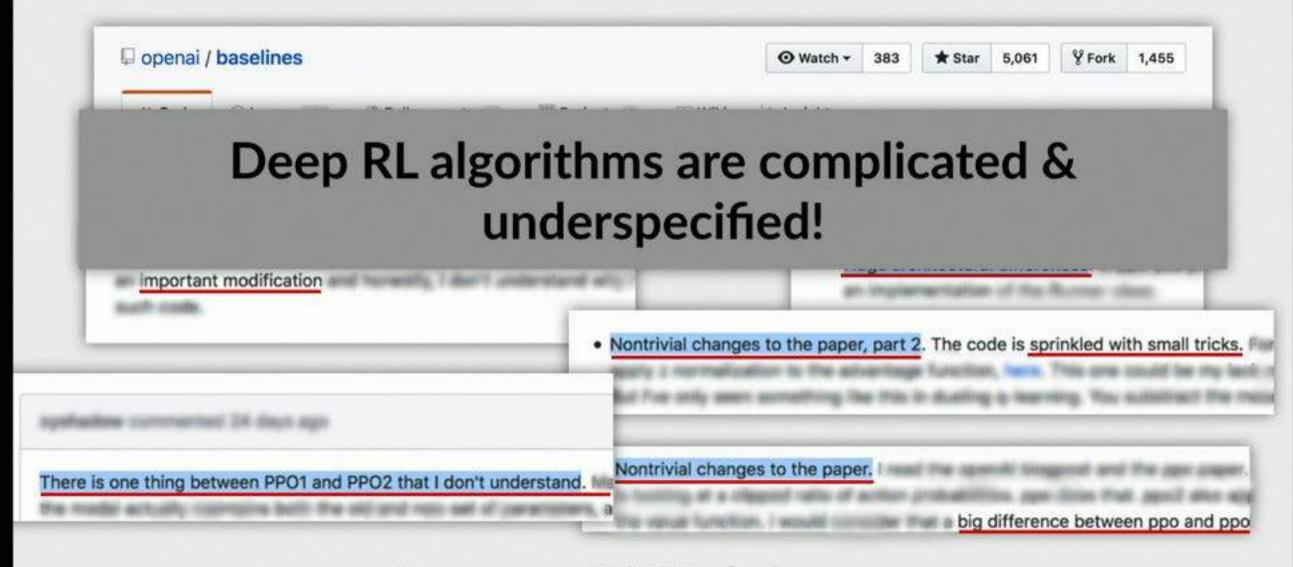
Where do such issues come from?

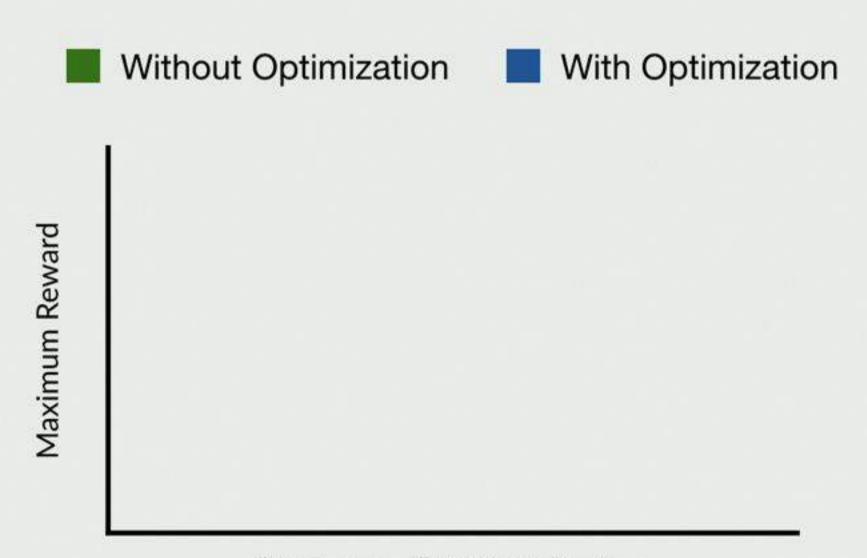
Hard to know: deep RL algorithms have many moving parts!



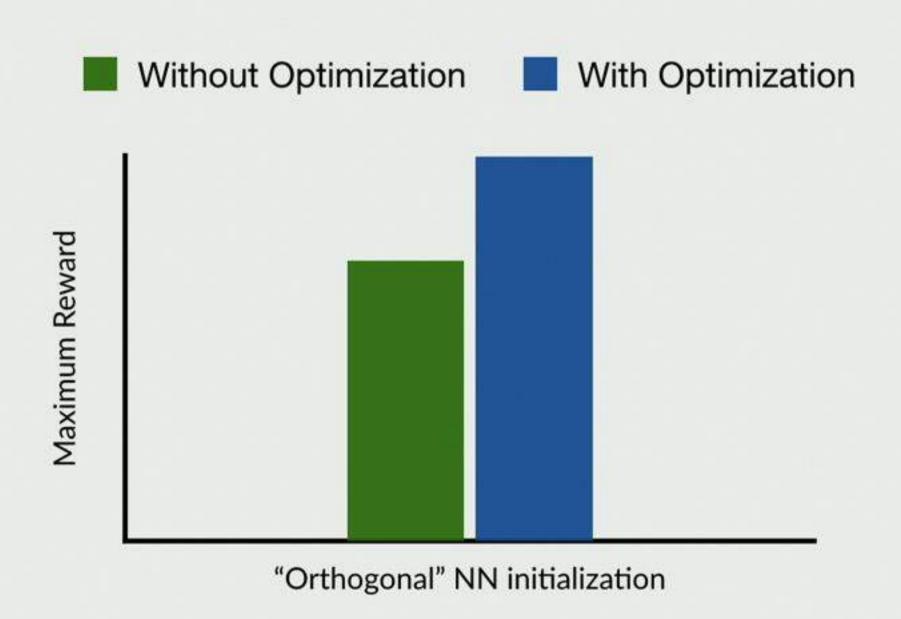


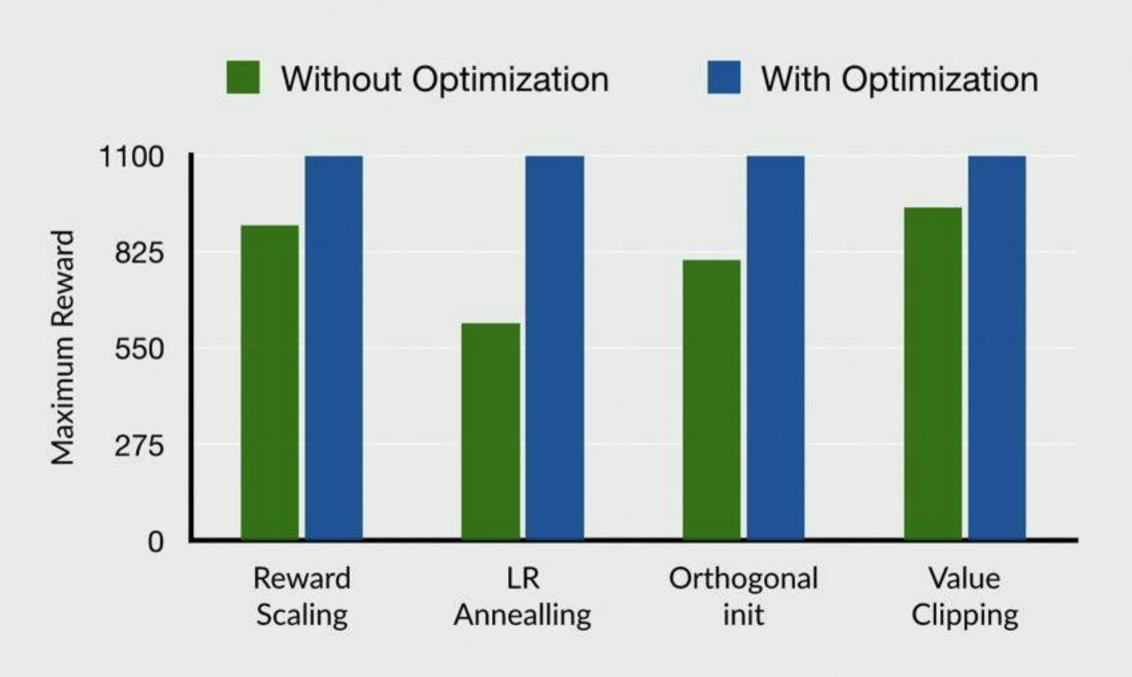






"Orthogonal" NN initialization





Implementation Obscures



Implementation Obscures Deep RL Algorithms

- Deep RL methods are complicated & underspecified
- Reasons for unreliability, performance are unclear
- Deep RL methods are poorly understood!

Back to First Principles

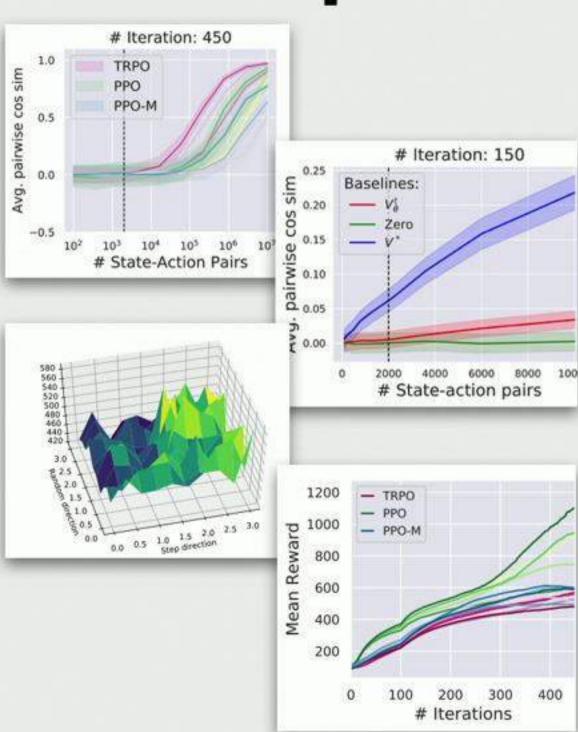
Back to First Principles

Gradient Estimates

Value Prediction

Optimization Landscapes

Trust Regions



Key assumption of policy gradient framework:

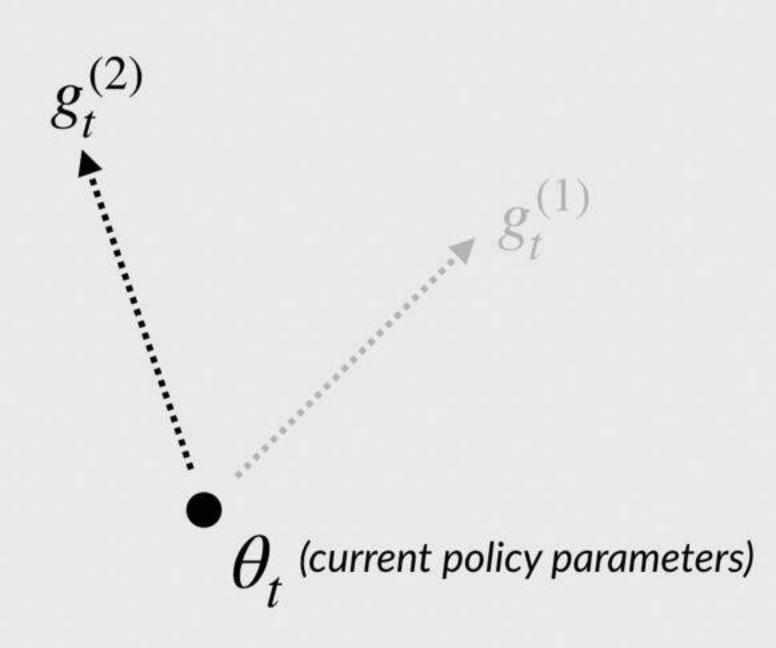
$$\nabla_{\theta} \mathbb{E}_{\tau \sim \theta} [R(\theta)] \approx \frac{1}{N} \sum_{\tau \sim \theta} g(\tau)$$

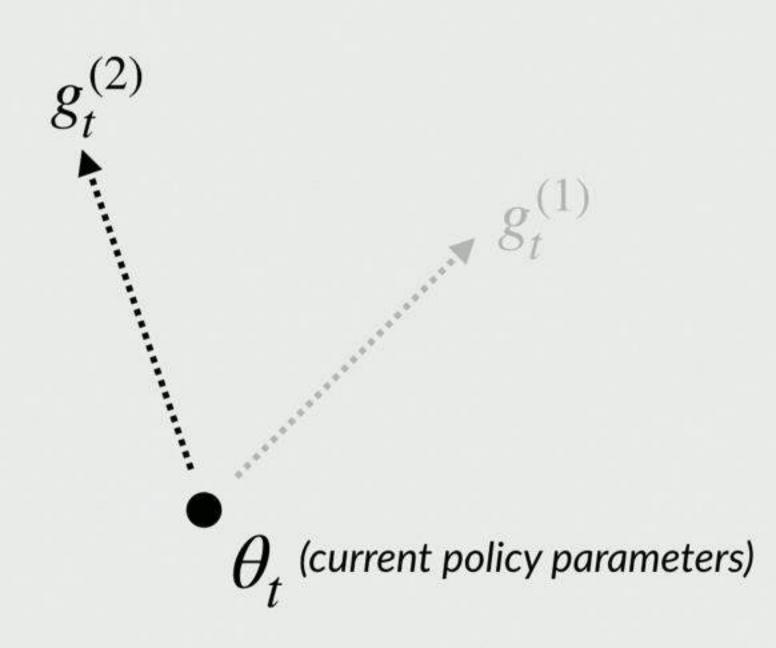
Key assumption of policy gradient framework:

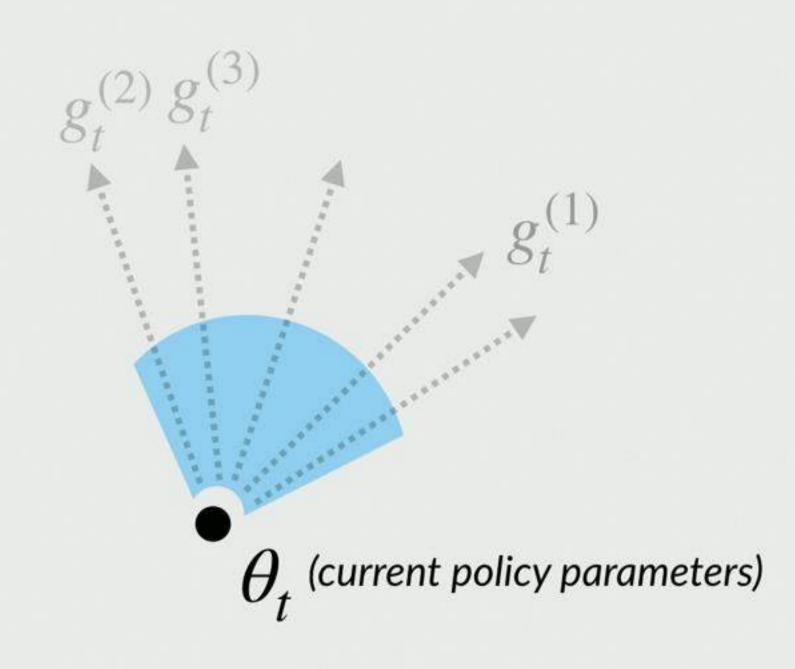
$$\nabla_{\theta} \mathbb{E}_{\tau \sim \theta}[R(\theta)] \approx \frac{1}{N} \sum_{\tau \sim \theta} g(\tau)$$

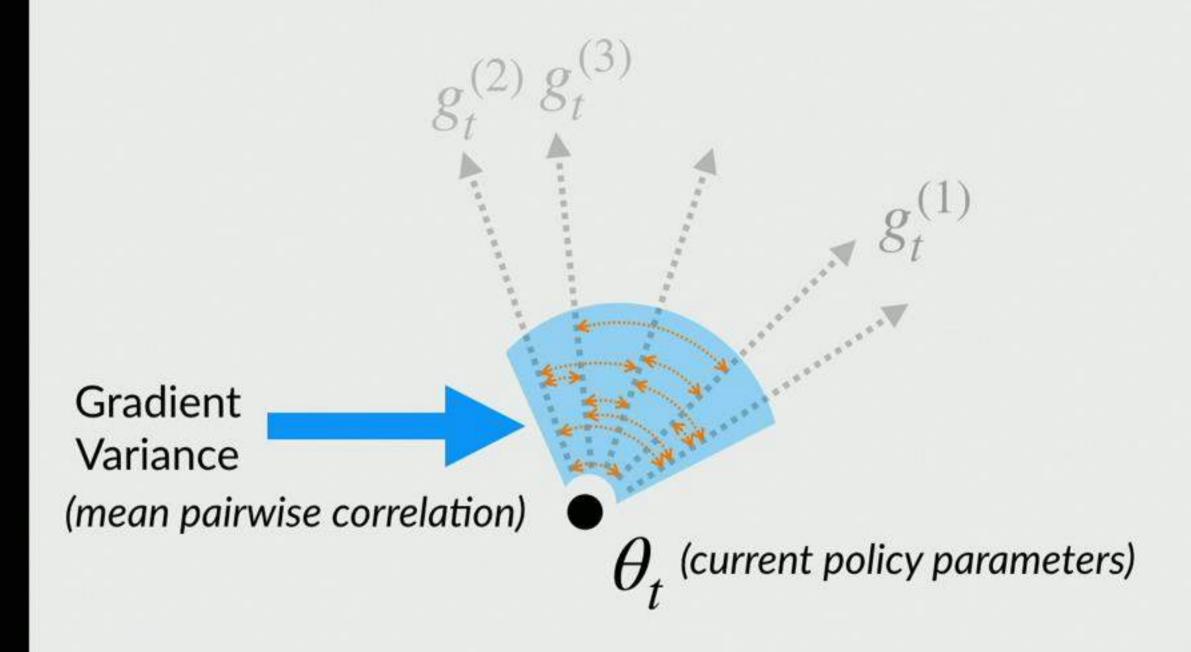
How valid is this?

ullet $heta_t$ (current policy parameters)

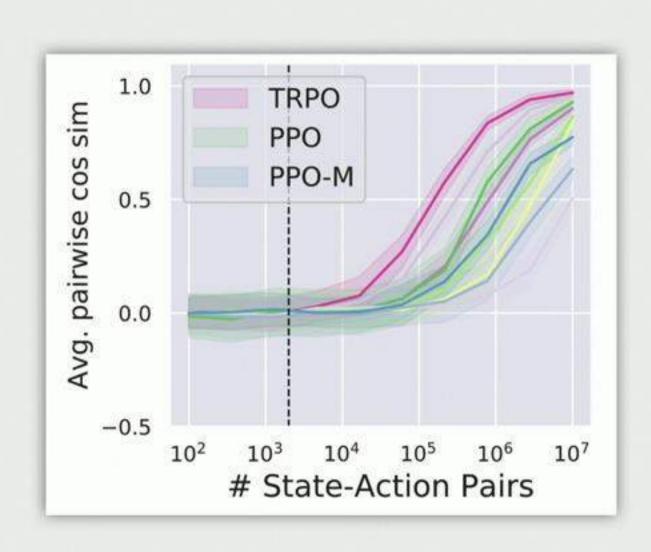




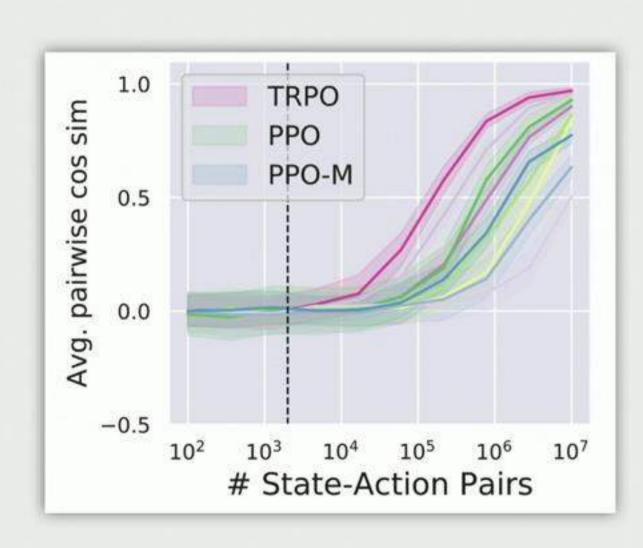




Gradient Variance



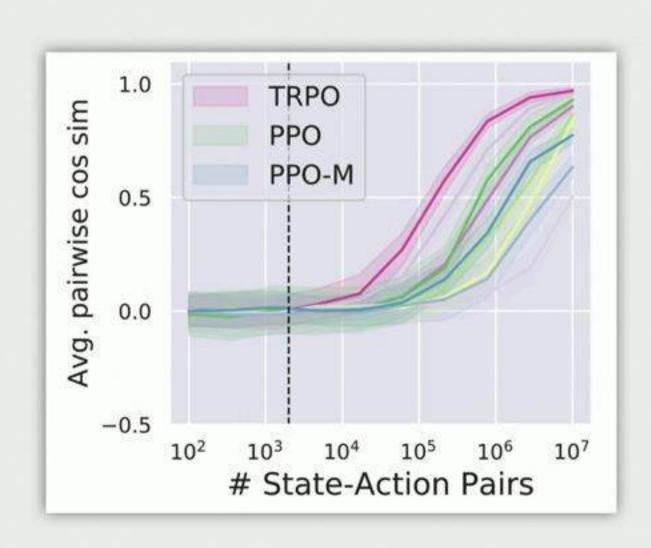
Gradient Variance



- Black line: relevant sample regime
- Gradients are less concentrated than they could be
- Less correlated for "harder" tasks, later iterations

- No good understanding of training dynamics
 - How does variance influence optimization?
 - Can we use insights from stochastic opt?
- Missing a link from reliability to sample size

Gradient Variance



- No good understanding of training dynamics
 - How does variance influence optimization?
 - Can we use insights from stochastic opt?
- Missing a link from reliability to sample size

Gradient estimation is hindered by high variance!

Observation: If we can estimate the value of a state, can significantly lower variance

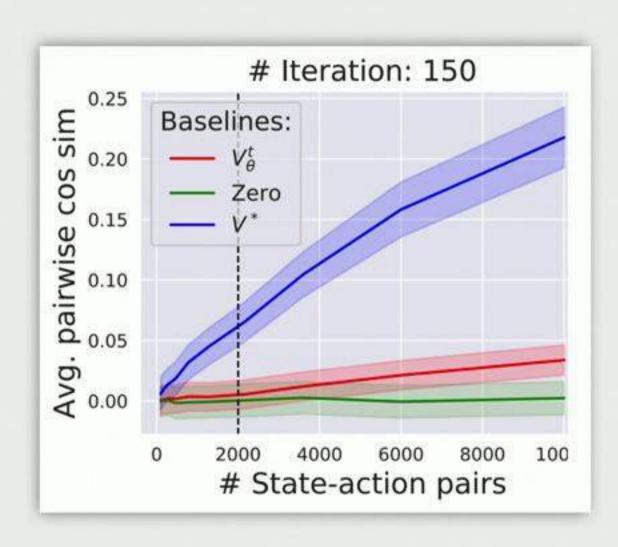
(The value of a state is the cumulative expected reward received after visiting the state)

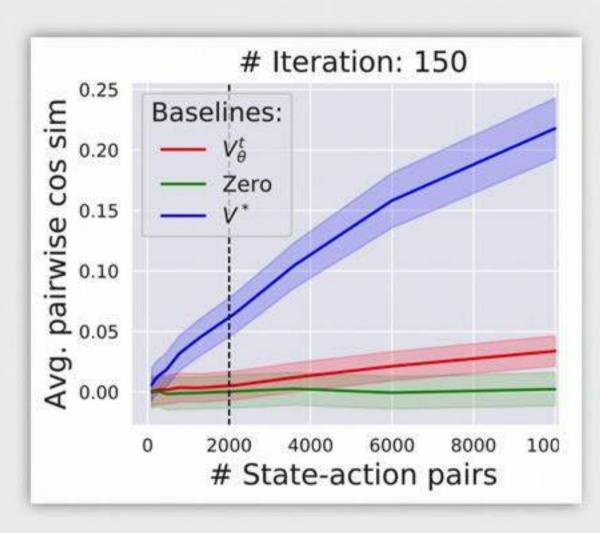
Intuition: Need to separate action quality from state quality

Variance reduction needs good value estimates

In Deep RL, values come from a neural network

To what degree do we actually reduce variance?



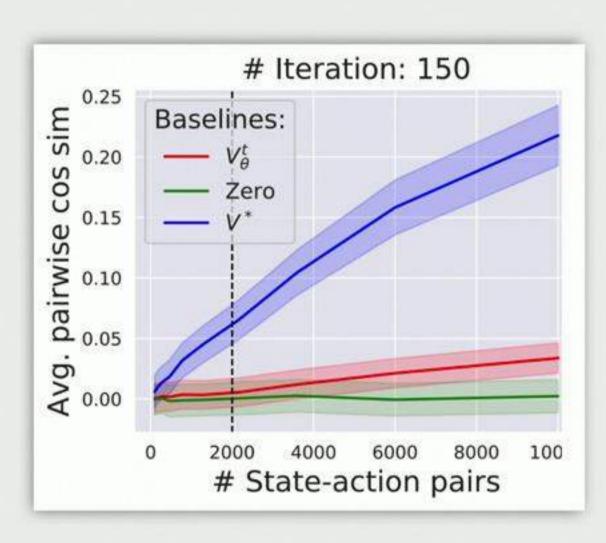


True value function

Agent's value function

No value function

Agent does significantly worse than optimal!



- Might look small, but using a value network makes big difference
- ▶ How would using the true value affect training?
- ▶ Can we get better value estimates?

True value function

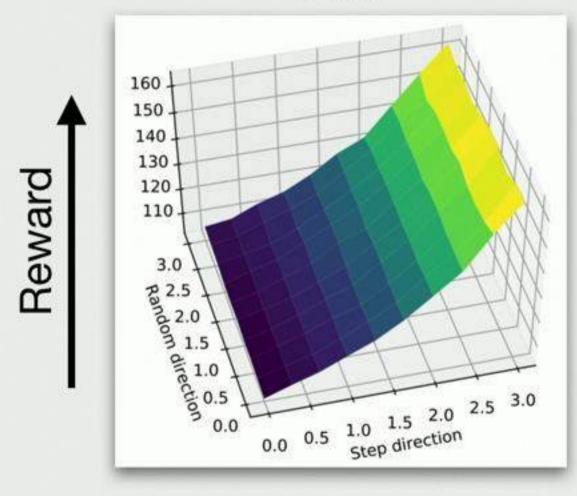
Agent's value function

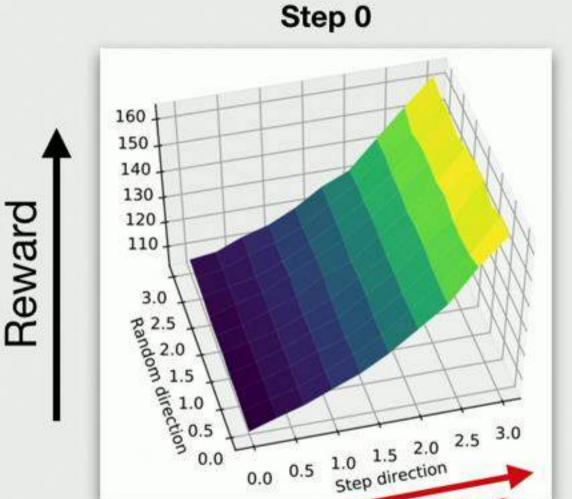
No value function

Assumption: taking gradient steps increases reward

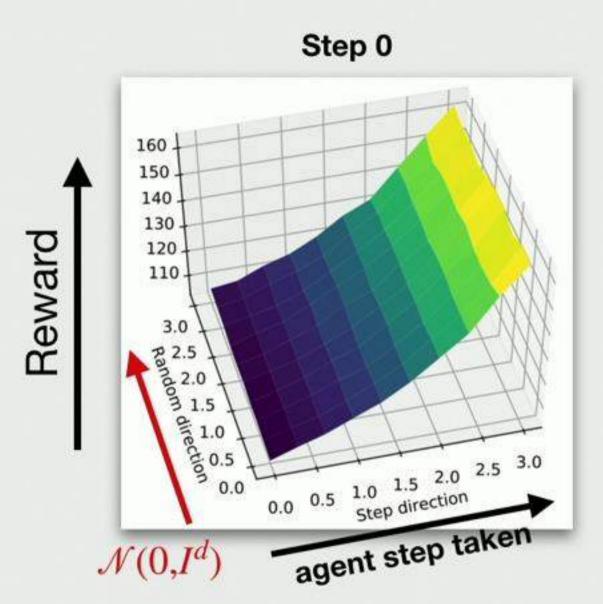
How valid is this assumption in practice?

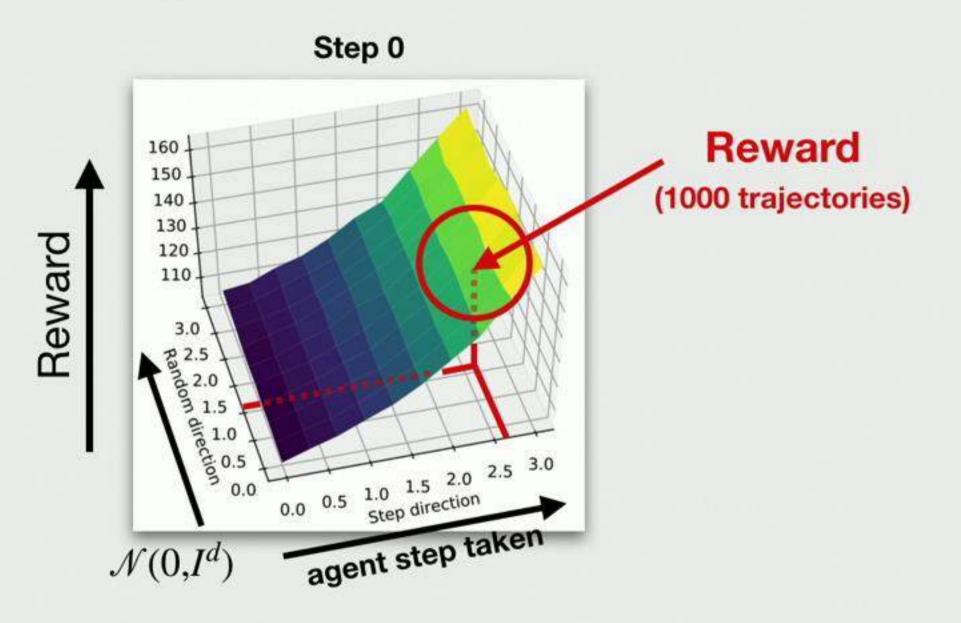


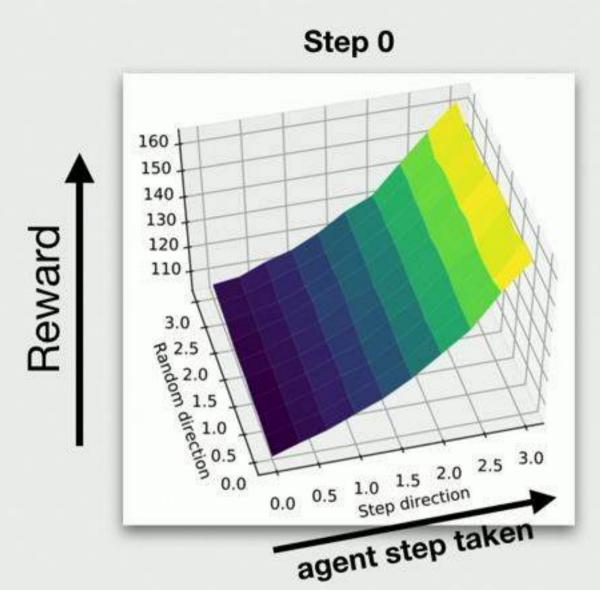


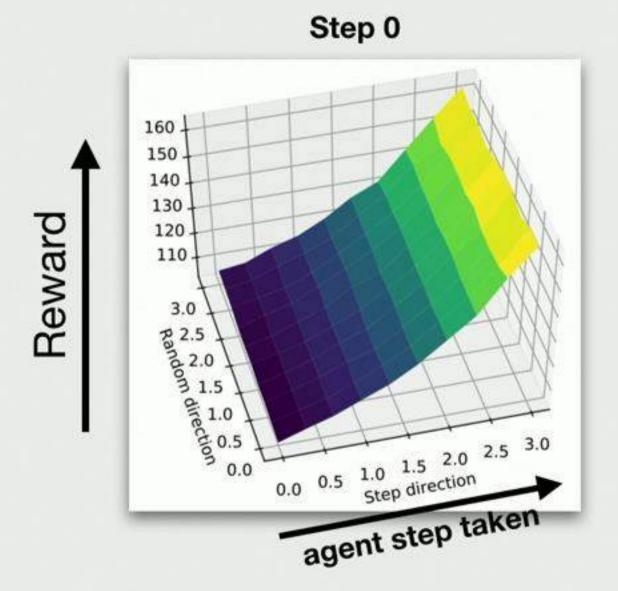


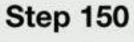
agent step taken

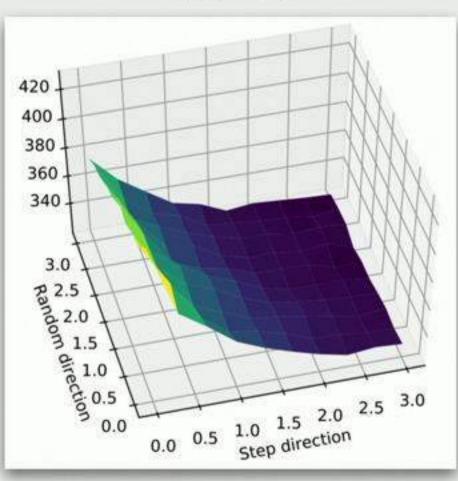


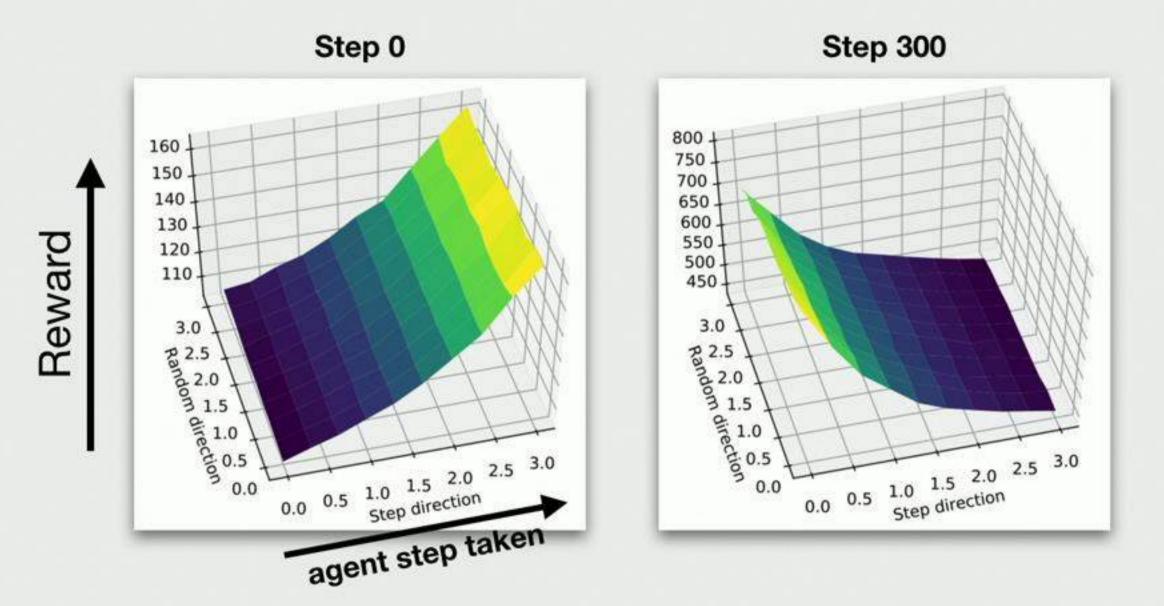




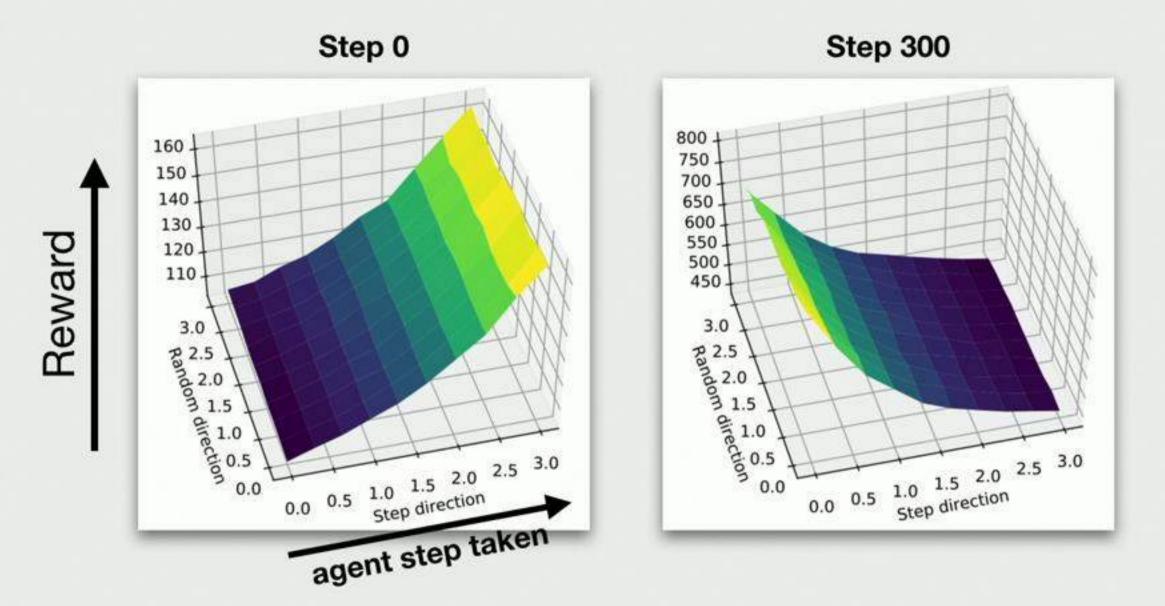








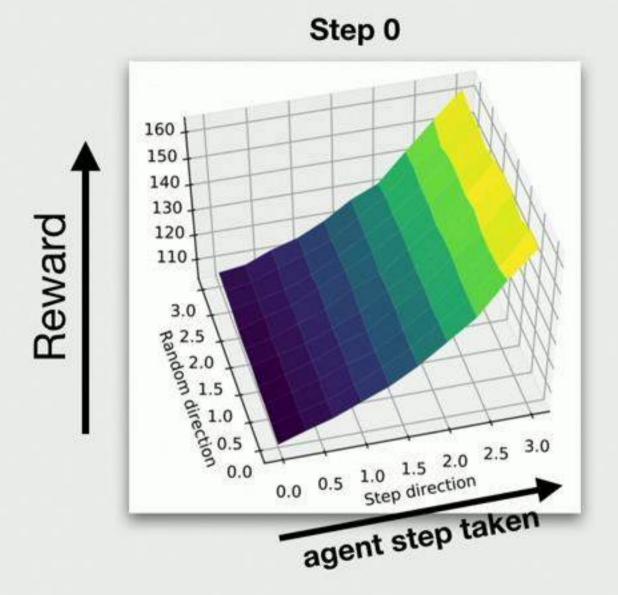
Steps are often not predictive

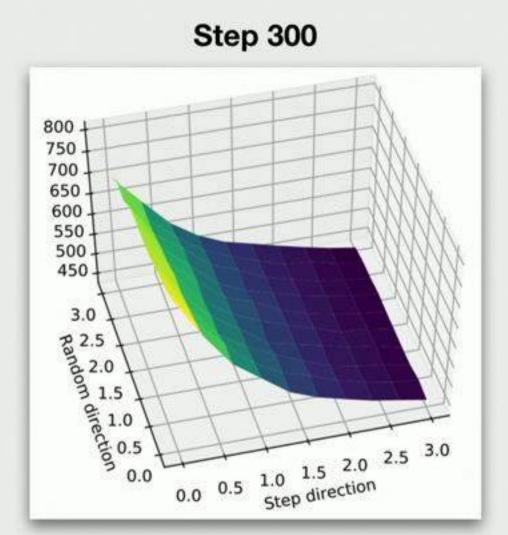


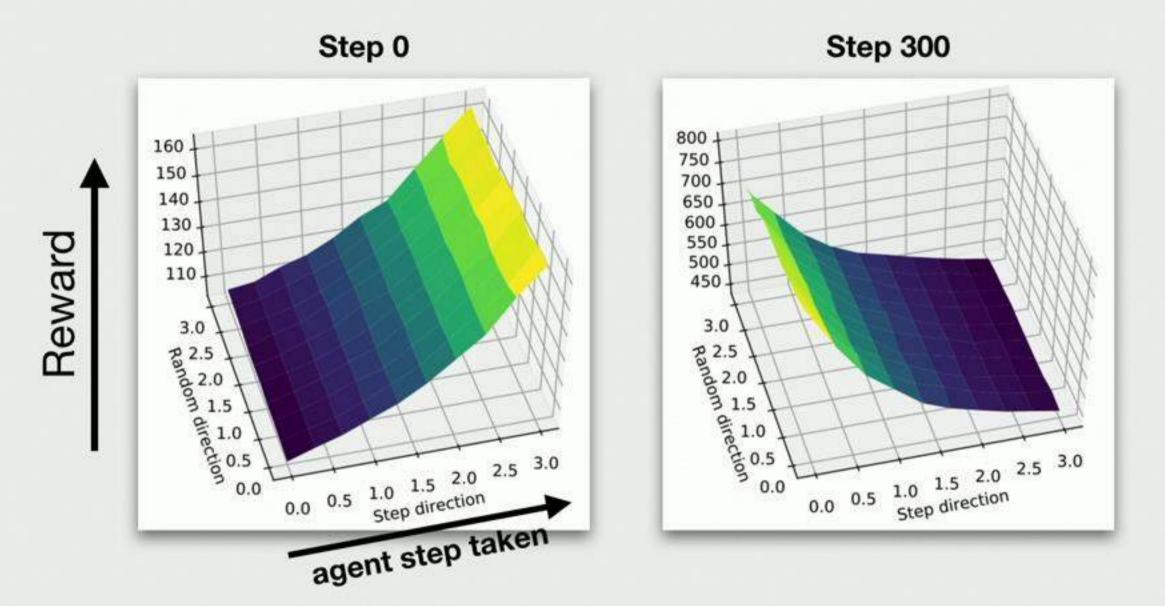
Steps are often not predictive

What's going on here?

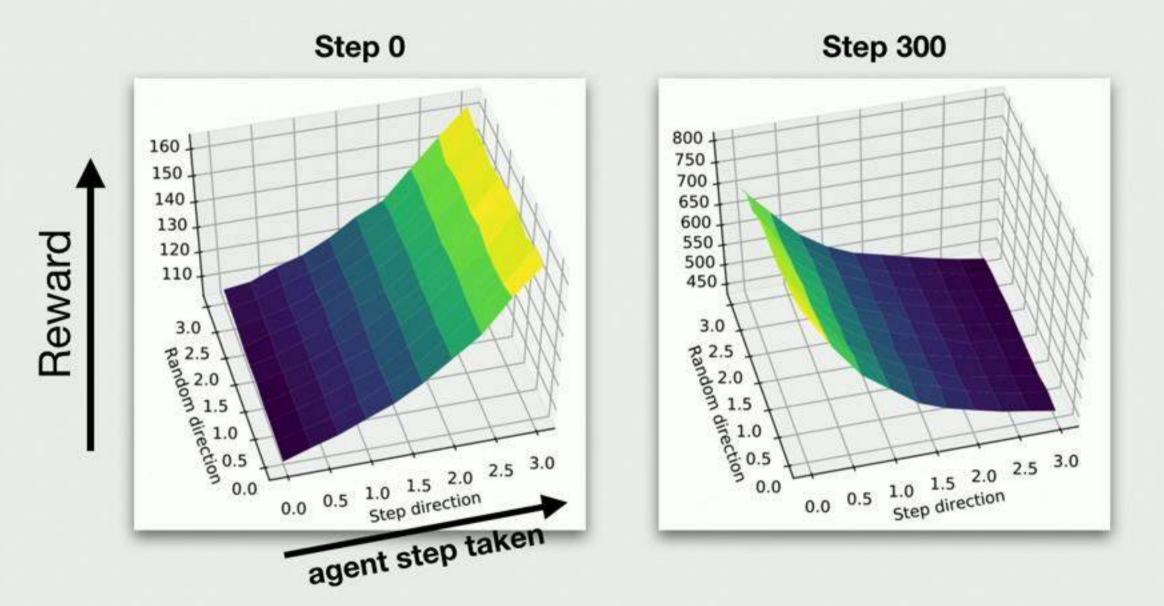
Methods iteratively maximize a "surrogate reward" (not the true reward!)







Steps are often not predictive



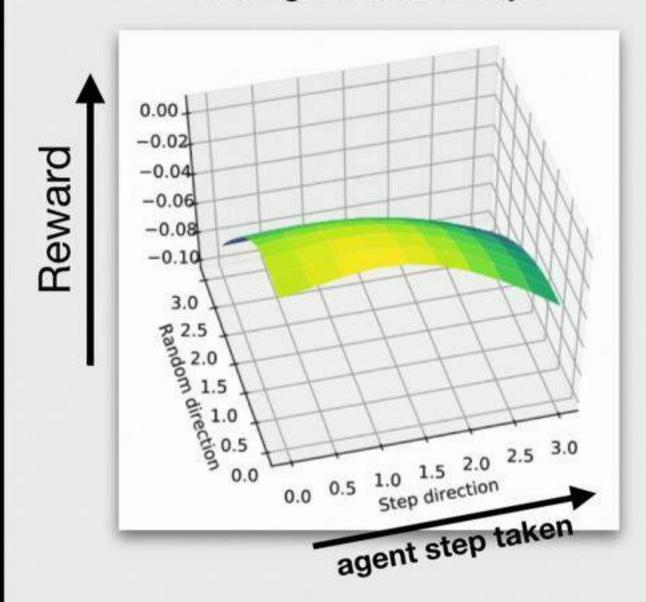
Steps are often not predictive

What's going on here?

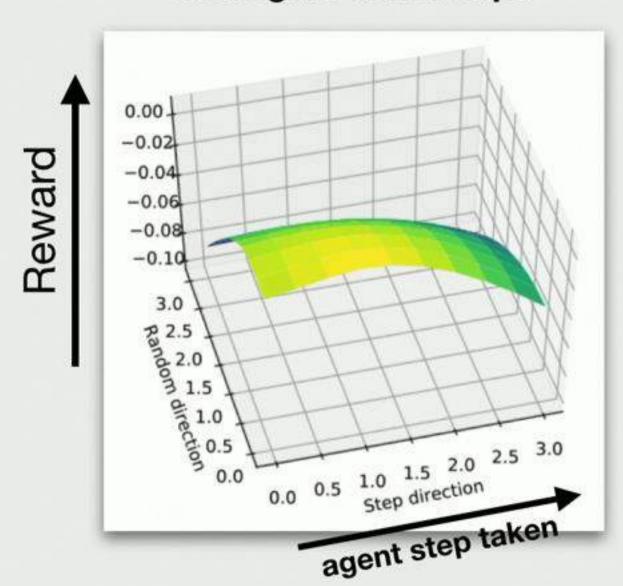
Methods iteratively maximize a "surrogate reward" (not the true reward!)

How do surrogate rewards compare with true rewards?

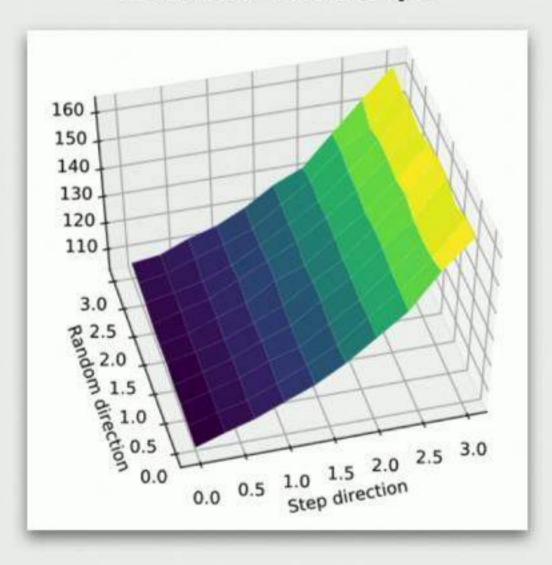
Surrogate Landscape



Surrogate Landscape



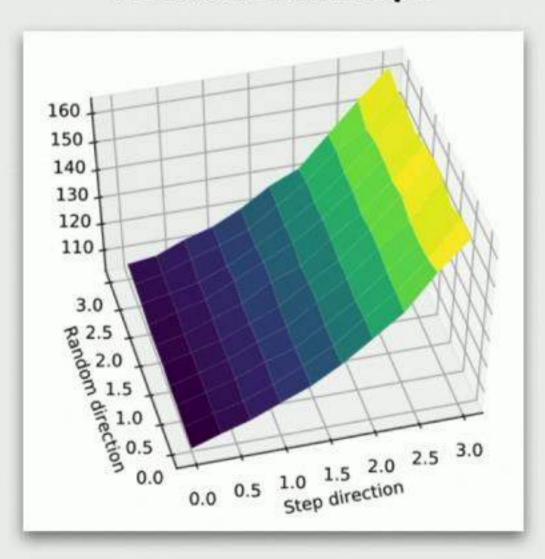
Reward Landscape



Surrogate Landscape

0.00 -0.02Reward -0.04-0.06-0.08-0.103.0 Random direction 0.5 1.0 1.5 2.0 2.5 3.0 0.0 Step direction agent step taken

Reward Landscape

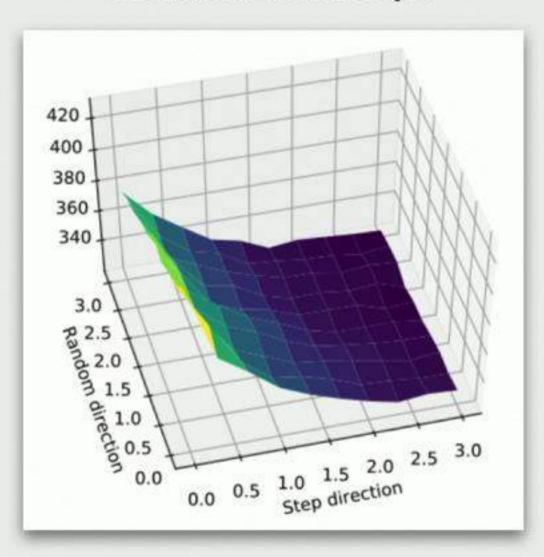


Step 0

Surrogate Landscape

0.0025 0.0000 Reward -0.002-0.005-0.0125-0.01503.0 Random direction 0.0 0.5 1.0 1.5 2.0 2.5 3.0 Step direction agent step taken

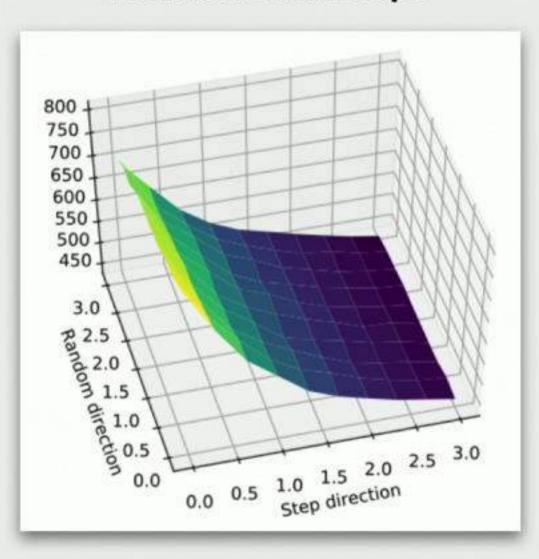
Reward Landscape



Surrogate Landscape

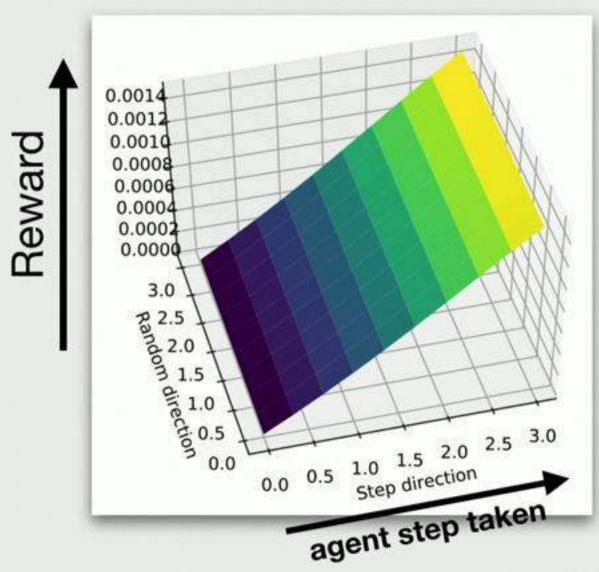
0.0024 0.001 0.000 Reward -0.003-0.002-0.001-0.004-0.005-0.0043.0 Random direction 0.5 1.0 1.5 2.0 2.5 3.0 0.0 Step direction agent step taken

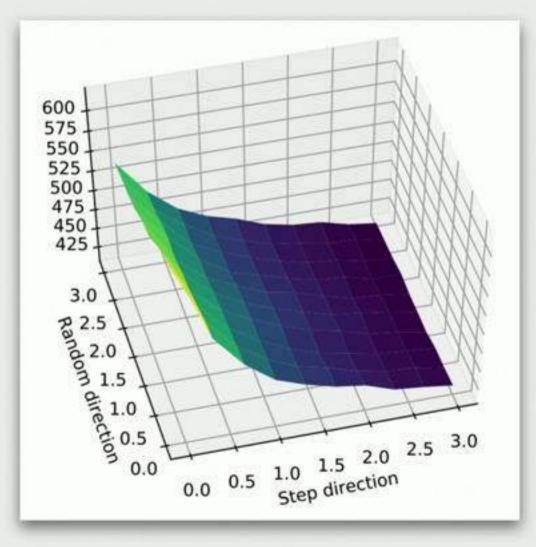
Reward Landscape

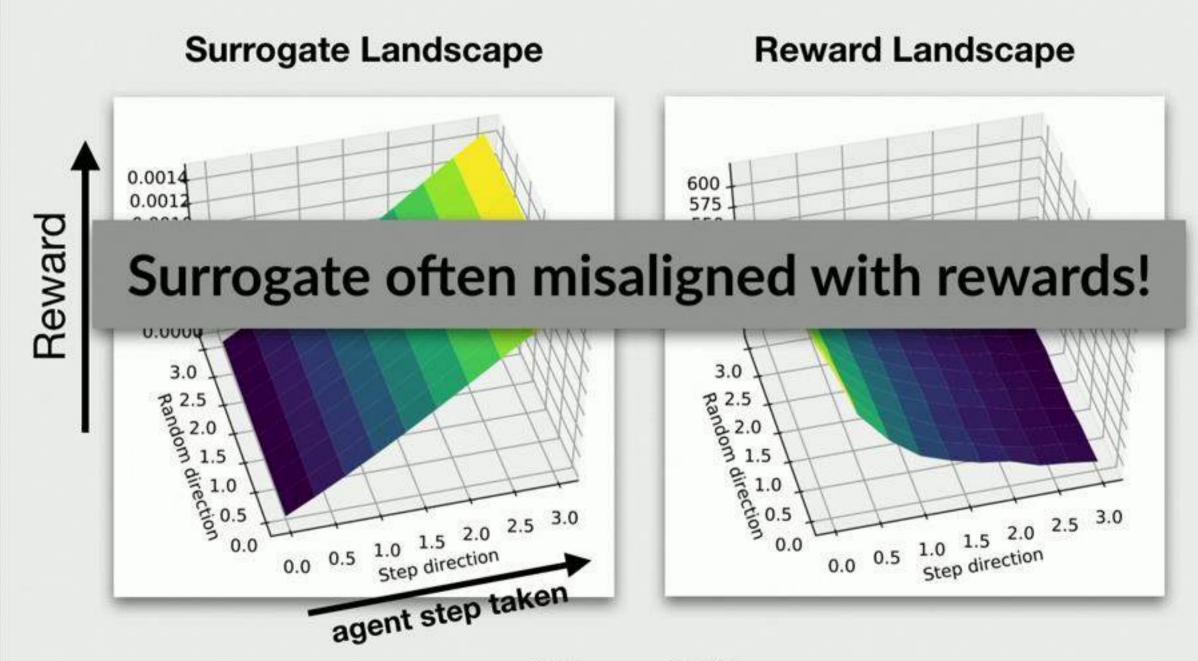


Surrogate Landscape

Reward Landscape







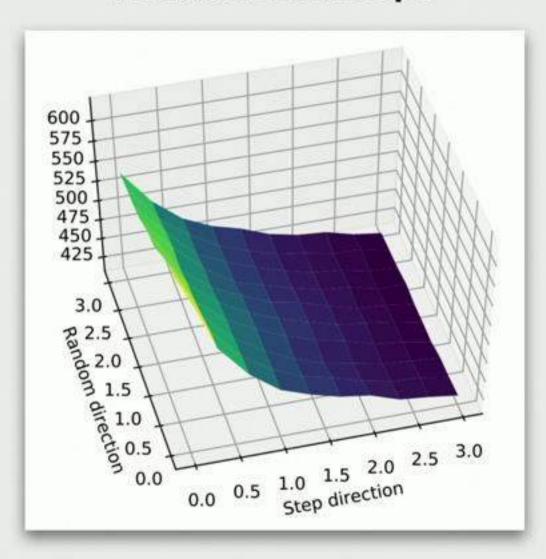
Surrogate Landscape

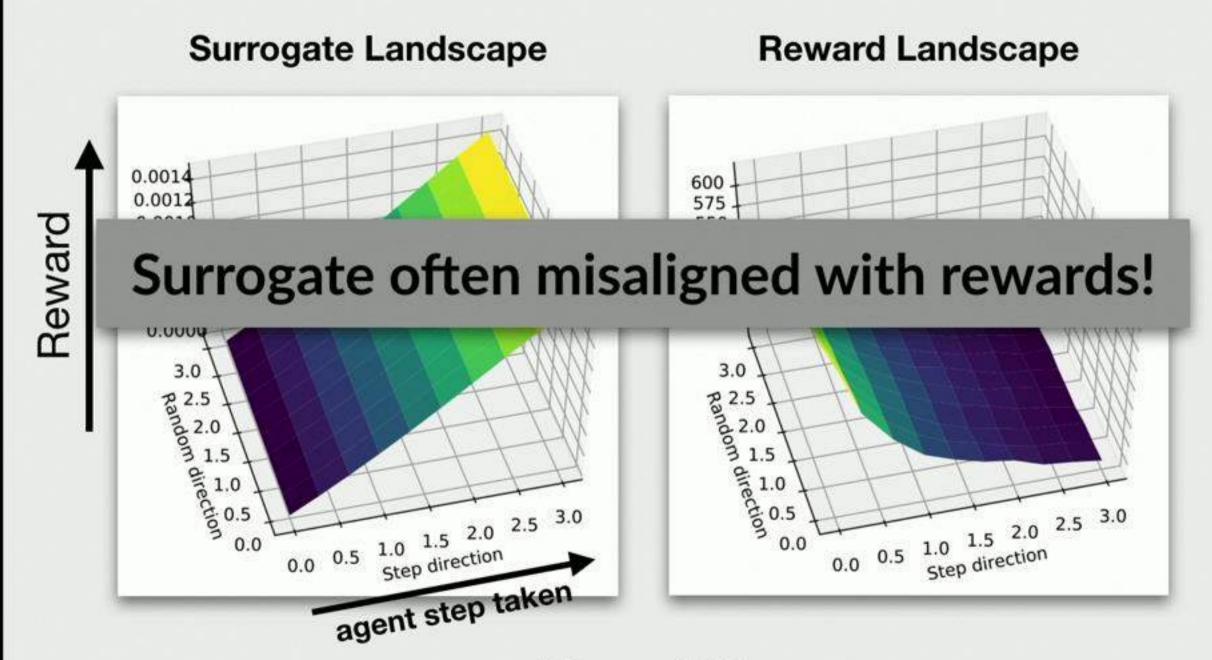
0.0014 0.0012 Reward 0.0010 0.0008 0.0006 0.0004 0.0002 0.0000 Random direction 0.0 0.0 0.5 1.0 1.5 2.0 2.5 3.0

Step direction

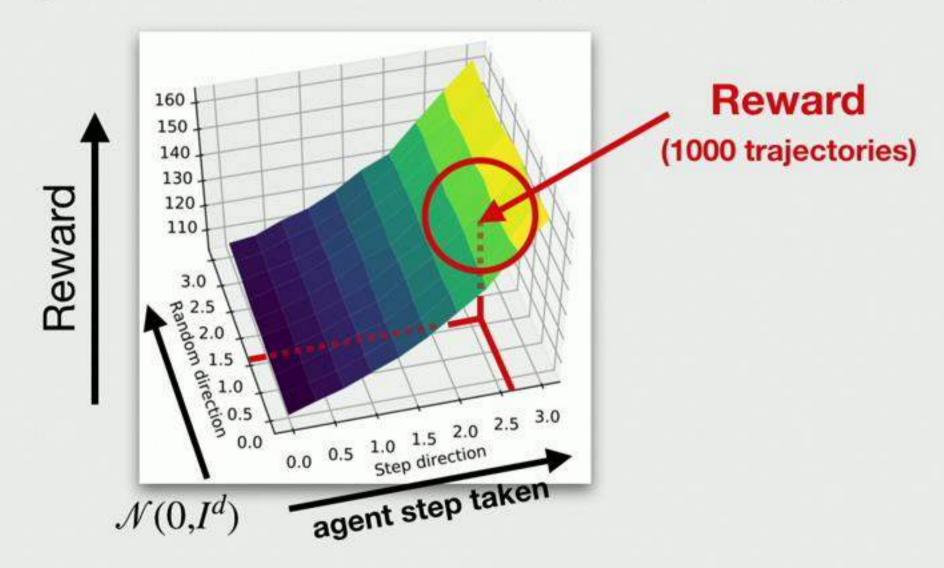
agent step taken

Reward Landscape



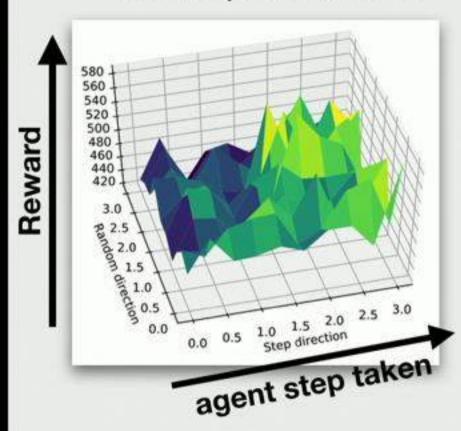


All landscapes so far are in the high sample regime

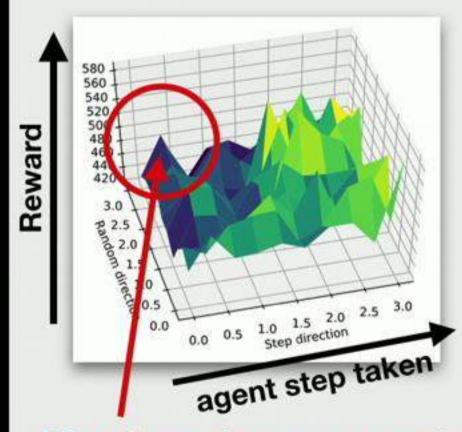


How do landscapes appear to the agent?
(~20 trajectories)

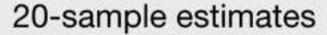
20-sample estimates



20-sample estimates

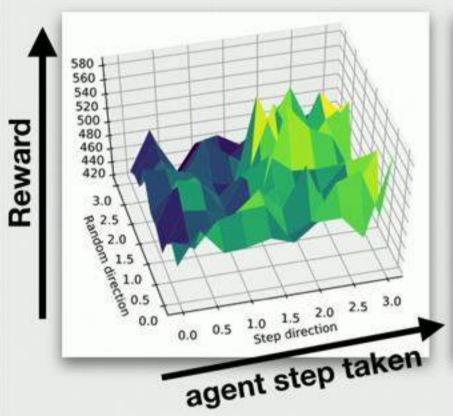


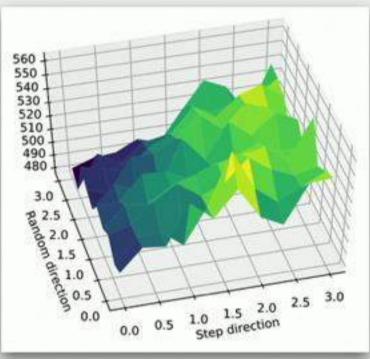
20 trajectories per reward estimate

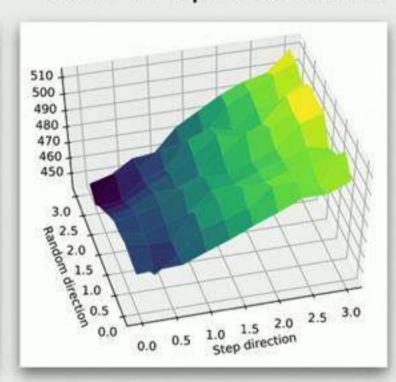


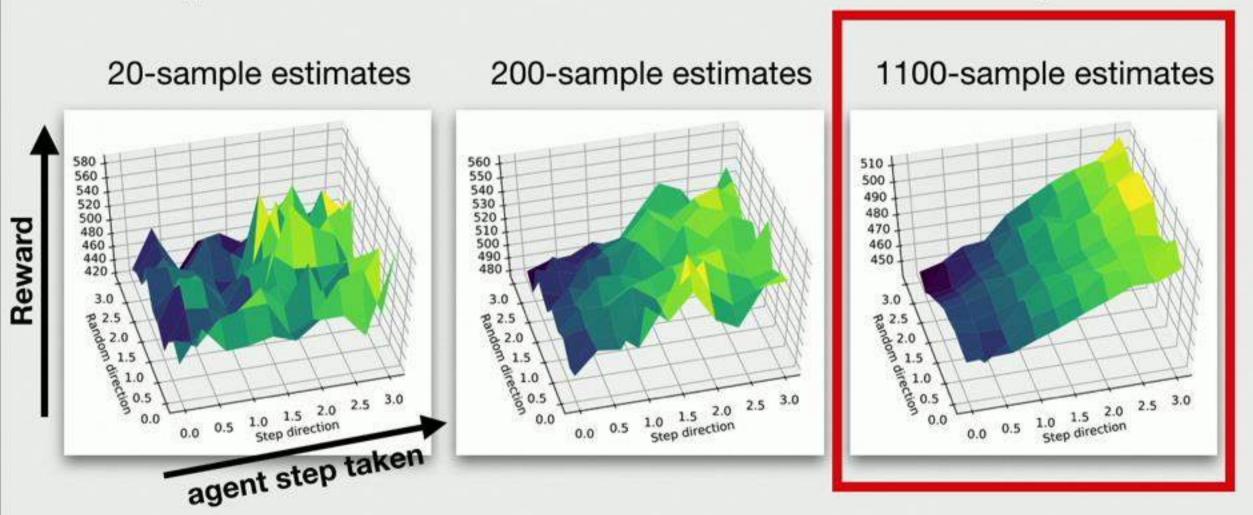


1000-sample estimates

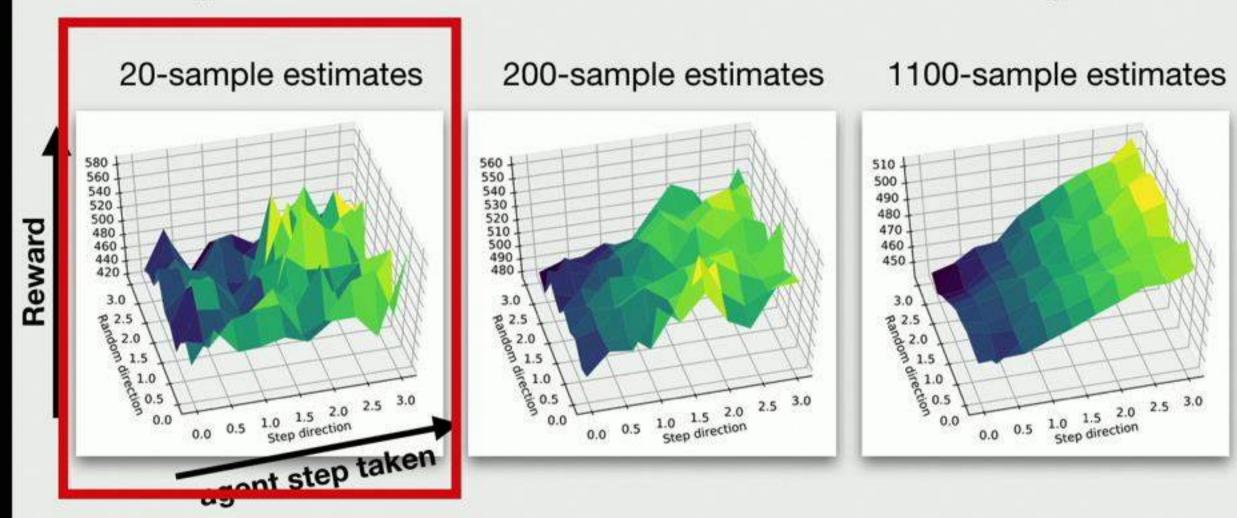








using many samples induces a smooth landscape...



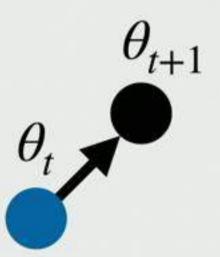
using many samples induces a smooth landscape...

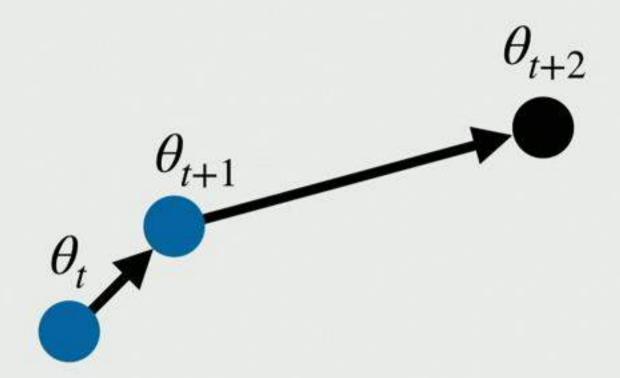
... but improvement is hard to detect in the agent's sample regime

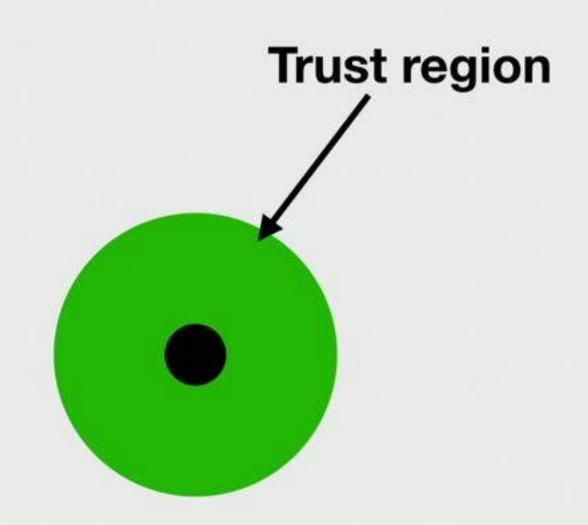
Surrogate landscapes are often not reflective of rewards

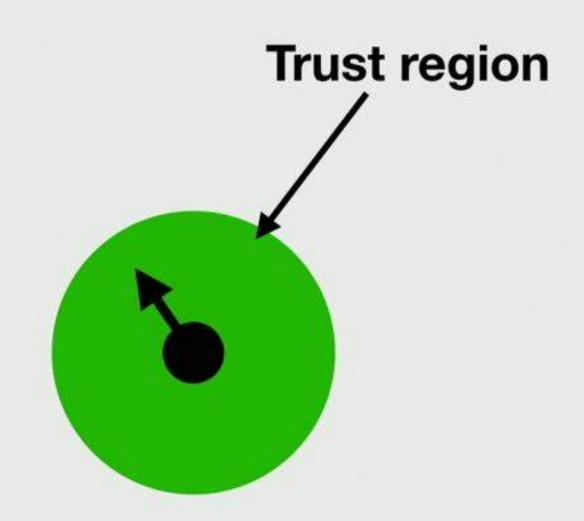
How can we better navigate the reward landscape?

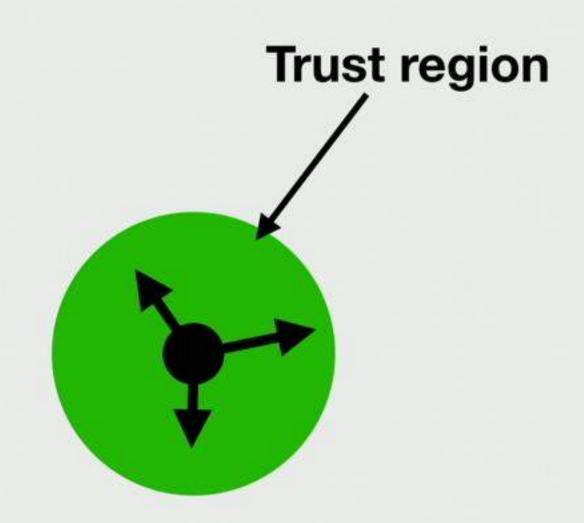
 θ_t

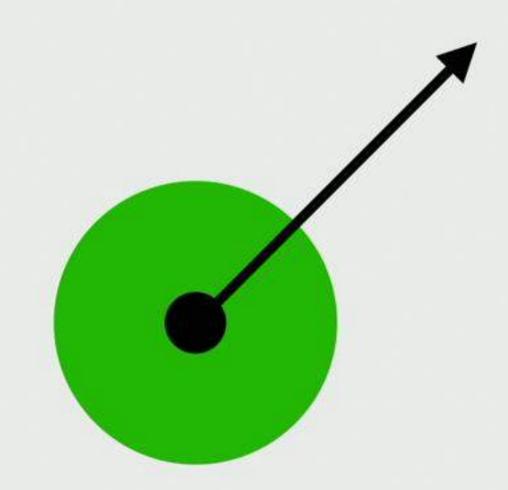


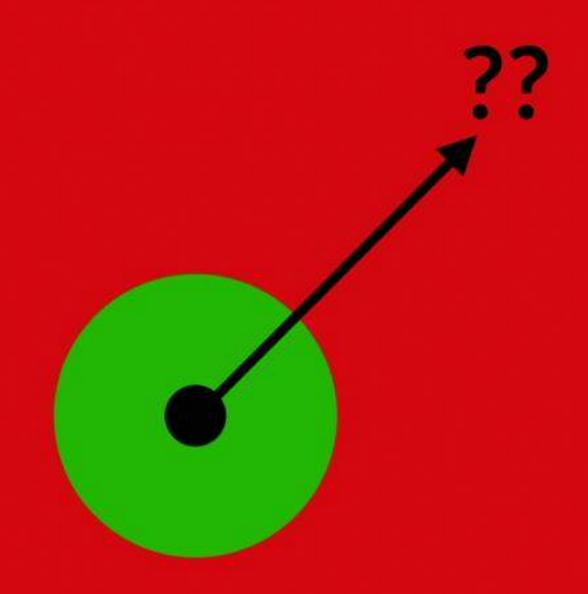












TRPO and PPO: Motivated by KL-based trust region:

$$\max_{s} D_{KL} \left(\pi_{\theta_{t+1}}(\cdot \mid s) \middle| \middle| \pi_{\theta_{t}}(\cdot \mid s) \right) \leq \delta$$

"keep the max distance between action distributions small"

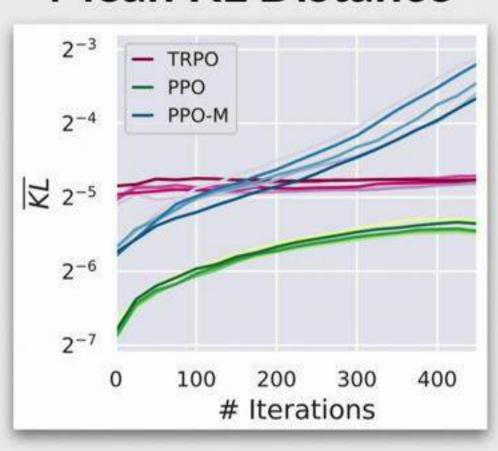
But relax to an expectation:

$$\mathbb{E}_{s \sim \theta_t} \left[D_{KL} \left(\pi_{\theta_{t+1}}(\cdot \mid s) \middle| \middle| \pi_{\theta_t}(\cdot \mid s) \right) \right] \leq \delta$$

"keep the mean distance between action distributions small"

What happens in practice?

Mean KL Distance



- ▶ TRPO maintains trust region
- PPO algorithm does not!
- ... but optimizations help

- What part of algorithms keep trust regions?
- How do we reason about algorithms when they use such loose relaxations?
- How can we capture different kinds of uncertainty in our trust regions?

Takeaways

Recap

- Deep RL methods are complicated
- Deep RL training dynamics are poorly understood
 - Steps are often uncorrelated
 - Surrogate rewards do not match true rewards
 - Trust regions do not hold

How do we proceed?

- Reconciling RL with our conceptual framework
 - How can we make algorithms better follow our conceptual framework?
- Rethinking primitives for modern settings
 - ► How do we deal with high dimensionality? Algorithm "optimizations?" Non-convex function approximators?
- Better evaluation for RL systems
 - Benchmarks don't capture reliability, safety, or robustness of RL agents

Read more

Paper

Revisiting the Primitives of Deep Policy Gradient Algorithms

We study how the behavior of deep policy gra-ducer algorithms reflects the concepted framewith materialist from development. We propose a line-grained analysis of more of the set methods based on key aspects of the haspework gradient aritmation, value production, optimization landscapes, and tree region influencement, from the perspective, the futures of deep policy gra-dense algorithms office devices from what don't medinating framework may product. Our analysis suggests first supe nemark willdrying the fram-

son bookmark-come portrains perfectively.

of the most published adversaries of another machine bearing (Salver et al., 2017; OpenA, 2018; Deputation et al., 2018). To many, this transmiss, coloradors the promise of the mini-world impact of machine training. However, the deep RI, toollish has not per attained the same Root of exponentia subside as, for example, the current days congressed housing trainments, basical super mod-ine (Roudevers et al., 2017) decements that state-of-the at-deep RL algorithms rather from memoriate by to began parameters, lack of consistency and poor reproducibility.

The same of affairs regions that it night be occurred to se-custom the concepted audosphasings of deep 60, methodology. Most precisely, the eventwhing species that makington this week in:

To what aligner along the content practice of along AC seglect — in the extreme sample regime. the principles that informat its directions of t

Parliation; with Under review by the International Continued on Martine Lyapung (ICML). Do not distribute.

Out goal is to cighter the issued to which man-of-the-art. regionsociations of these methods second at multiplied for key pressions of the greenil policy gradient framework.

method, presented policy operations (PPO) (Scholman et al., 2013). We shad that PPO's performance deposits handly on optimization outside of the ones algorithm. This argument the faction account of PPO might nor to fully explainable by in manifesting theoretical transversit.

This observation prompts as to ladar a broader levit of poli-try produced algorithms and their relation to their sudorly-ing Statements. With this perspective or exact, we purform a time potential examination of key EE, primitions as they namifor to practice. Concentrally, we study

Deep satisfactories bearing (BE) in as the core of some improving in come of proved, the gradient colonies and to spikes their personners are often provily constituted with the true gradient. We also find that photom colorate qual-ity decays with training progress and tack complexity.

Value Freedottime our experiments technic that value not works occurredaily wider the supervised brancing task they are travel on, has do not fit the lose value function. Addisorally, employing a value between as a hundrer only suggested decisions the taristics of guident extractive companed to using that value that dissembledly increases agent's performance companed to using an healther a fall.

Optionization Landwagen; we also observe that the optimoterian beads upo induced by sorders policy produced al-portions to other our suffection of the analyting tree reward landscape, and than the latter is office presily helacoid.

The specific facus of this paper is no deep policy problem. Said, as preciously produce of the paper is no deep policy problems of smellinds, a widely used close of deep KE, algorithms.

Ast, as preciously policy operations, these reductions are found a handward problem in the algorithm's design.

These that this share does not period to guidents from hing which falled, the adoption to be the acts in every mechanic up-removable falled, the control of the control of the degree or out-per members are falled from the control of the falled fall

https://bit.ly/2UQGpad

Blog Posts

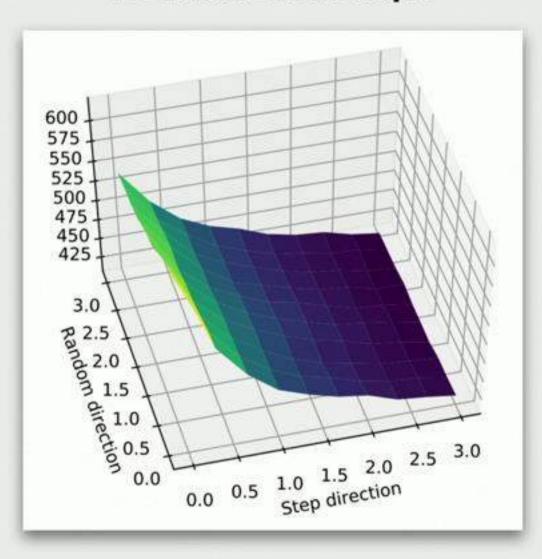


https://bit.ly/2WVPjAn

Surrogate Landscape

0.0014 0.0012 Reward 0.0010 0.0008 0.0006 0.0004 0.0002 0.0000 Random direction 0.0 0.5 1.0 1.5 2.0 2.5 3.0 Step direction agent step taken

Reward Landscape



Read more

Paper

Revisiting the Primitives of Deep Policy Gradient Algorithms

Assertance Author

Abstract

National Was study have the Spinners of deep policy gradient algorithms to Barts the underposed. Unanwork medicates from developments. We propose a time-granted study on of state-of-the-per softned branch for the appears of their intersection, gradient extensions, value production, optimization Londongers, and bear region collections, from the perspective, the believes of deep policy gratice perspective. One believe we deep policy gra-

detti arasimini, vatai produtini, apotatationi laudingon, odi triai regioi nitrivingoni. Piene diri perspettive. De fallavi era si dang pelalay graduan diproduni nilau deviane fine vibili dipet merioring francivari mag pendiri. Oler analysis saggiosi. Inti supe remarks velidiliying dia humudatimas of their algorithem, and in particular in distance of their algorithem, and in particular in distance of their algorithem and layeratific varitions book limital, annex conditions methodologi.

1. Introductio

Deep standard, most depring (BEL) in at the core of some of the most publishmed substrained of modern standard business (Show et al., 2023; Special, 2024; Separatha, et al., 2023). The many offse finderweak codes from the modern standard of the modern standard sta

The new of affairs suggests that it might be necessary to re-continue the conceptual and optionings of deep \$0, textbodelegy. More precisely, the events/tring openion that mentions this work in:

So what dispers about the content processor of about the patient of the extremal sample regime.

The principles that informated the distributions of the patients of the patie

The specific facus of this paper is on deep policy gradant healthch, a which would chost of deep NE, algorithms.

Annesses Baltima, Annesses Chy, Annesses Br ges, Annesses Court, Carepostos tr, Annesses An

Parliament wint. Under revers by the Immunication Continues on Marking Courses; If Mr. J. De not distribute.

One goal is to explore the count to which tops of the art implementation of these methods assured at medicing the key precessor of the green's policy gradient framework.

1.1. Our contribution

Our potent of mart in a presentent deep policy grades to technical, presented policy representation (IPCO) (Schulmaner al., 2001). We find the PPC's portionation depends has been expensionally a certain of the core algorithm. This originate fluid practical occurs of PPCO ought not be fully explainable by to meeting fluoristical fluoreweek.

This observation primages us to take a breakly look of policy product algorithms and their orbitals to their underlying framework. With this primpictive in mind, we justified a fine-granted reactionables of facts EL primitives as they assemble on practice. Contributy, no study.

Over minimization Mantaing (BL) is at the case of some of the most published abbriometer of modern resolution of abbriometer of modern resolution of abbriometer of modern resolution is abbriometer of modern resolution in publish their presentation of the proof or modern resolution of abbriometer of a 2001. Operation of a 2001. To make, this followerest, softwarfs for their gradient." We also finished professor columns qualitation at al. 2003. To make, this followerest, softwarfs for

Valor Feedbellines our experiences indicate that valor tearents inscreediby solve the important incoming teal, they are regard one. But do set if it be loss while function. Additionally, respituing a value network as a function callinguistly. Opinious the trainers of guidals outputs companied for using time value (the disassimely) incorrangates, a politerature companied in using an inactive at all.

Optionization Landwages: we also observe that the optineutrains hashe ups to be of by another policy gradient asgradient is often out reflection of the another land was reward landwage, and then the latter is often poorly behand in the network sample region.

Treat Regimes: we had deep policy gradiest apportions sessions violate theoretically sestimated tool regions. In San, se procedure policy systemation, these confusions seem from a handamental problem to the algorithm's design.

Then the this share does not packed guidents from hing under history the acquires on the filtrate or most machines optimized markey. However, it is notice on this degree or out the interestation from their unitraje to due; \$1, Alias, and find the guident randoms formula to complex may not other factors.

https://bit.ly/2UQGpad

Blog Posts



https://bit.ly/2WVPjAn