Reimagining Exploration Theoretical Insights and Practical Advancements in Policy Gradient Methods

Adrien Bolland

April 16, 2025

Sequential decision-making

Taking actions based on states in an environment where time evolves.

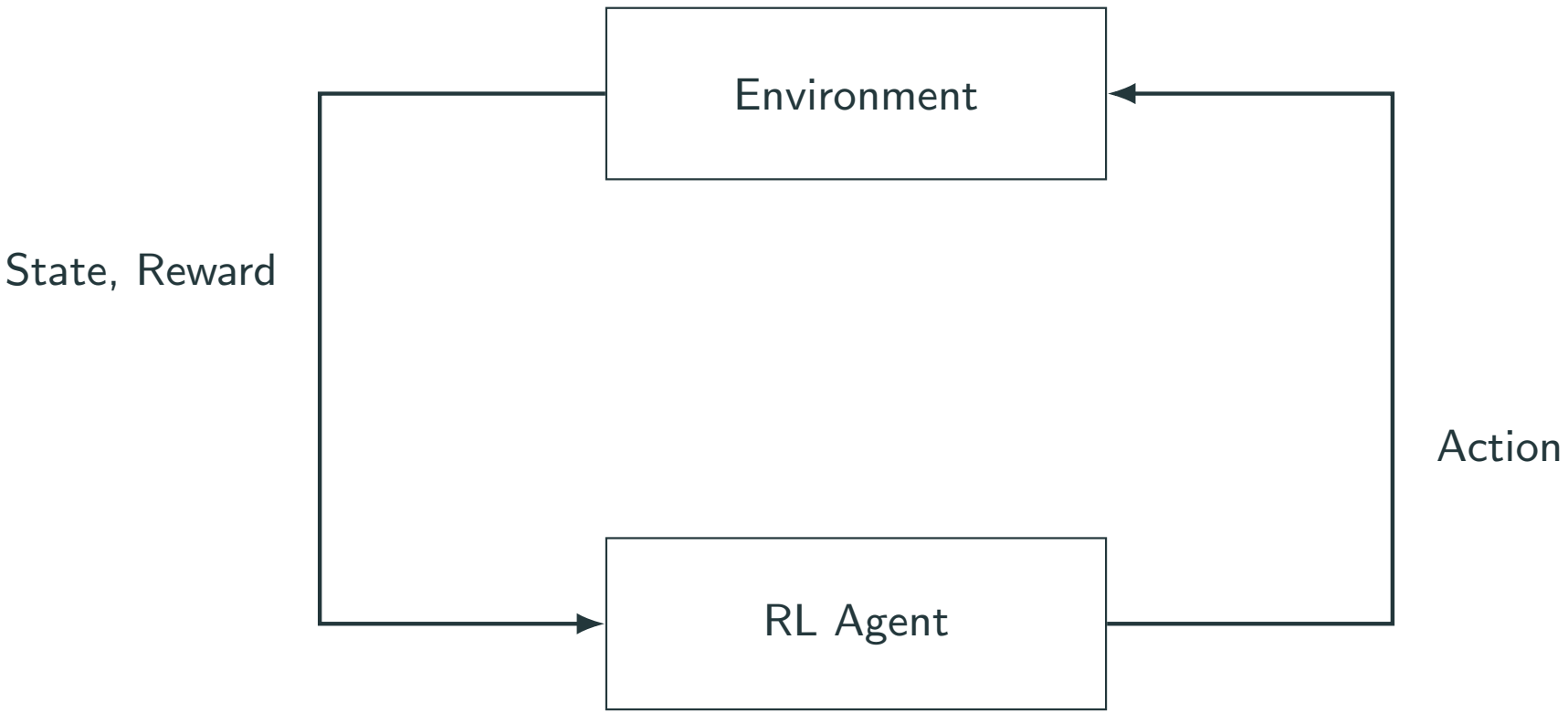
Sequential decision-making

Taking actions based on states in an environment where time evolves.

The quality of states and actions is measured using rewards.

Reinforcement learning

Agents learn from experience to act to maximize the expected sum of rewards.



Notations

Some reinforcement learning notations:

- $s_t \in \mathcal{S}$ for the states,
- $a_t \in \mathcal{A}$ for the actions,
- $p_0(s_0)$ for the initial state distribution,
- $p(s_{t+1}|s_t, a_t)$ for the transition distribution,
- $R(s_t, a_t)$ for the reward function,
- γ for the discount factor,
- $\pi(a_t|s_t)$ for stochastic policies,
- $\mu(s_t)$ for deterministic policies.

Policy gradient algorithms – Optimization problem

Definition (Optimal Policies)

Agents should act according to a policy π^* maximizing the expected return

$$J(\pi) = \mathbb{E}^{\pi}\left[\sum_{t=0}^{\infty} \gamma^t R(s_t, a_t)\right] .$$

Policy gradient algorithms – Recipe

In policy gradient algorithms:

1. The agent has a direct parameterized representation of the policy π_θ.

$$a_t \sim \mathcal{N}(\cdot|\mu_\theta(s_t), \Sigma_\theta(s_t)) \ .$$

Policy gradient algorithms – Recipe

In policy gradient algorithms:

1. The agent has a direct parameterized representation of the policy π_θ.

$$a_t \sim \mathcal{N}(\cdot|\mu_\theta(s_t), \Sigma_\theta(s_t)) \; .$$

2. The agent learns by stochastic gradient ascent.

$$\theta \leftarrow \theta + \alpha \hat{d} \qquad \hat{d} \approx \nabla_\theta J(\pi_\theta) \; .$$

Policy gradient algorithms – Exploration

The policy should remain sufficiently stochastic during the learning period to ensure sufficient exploration of the environment.

Policy gradient algorithms – Exploration

The policy should remain sufficiently stochastic during the learning period to ensure sufficient exploration of the environment.

Exploration-exploitation dilemma.

- An RL agent collects information and exploits information.
- Some actions are deliberately suboptimal.
- Necessary for some algorithms and efficiency criterion.

Numerical optimization perspective

The exploration-exploitation perspective falls short for studying policy gradients.

Numerical optimization perspective

The exploration-exploitation perspective falls short for studying policy gradients.

Forgetting about RL.

- The problem consists in maximizing an objective function.
- The optimization is performed by SGA.
- Algorithms converge at a certain rate.
- No exploration argument involved.

Numerical optimization perspective

The exploration-exploitation perspective falls short for studying policy gradients.

Forgetting about RL.

- The problem consists in maximizing an objective function.
- The optimization is performed by SGA.
- Algorithms converge at a certain rate.
- No exploration argument involved.

Understand exploration using numerical optimization arguments.

Abstract

During learning, the policy should remain sufficiently stochastic.

1. What is the influence of policy stochasticity on the optimization problem?

Exploration can be implemented using intrinsic exploration bonuses.

2. How good is this new learning objective?
3. How to design a good exploration strategy?

The role of the stochasticity of the policy

Question

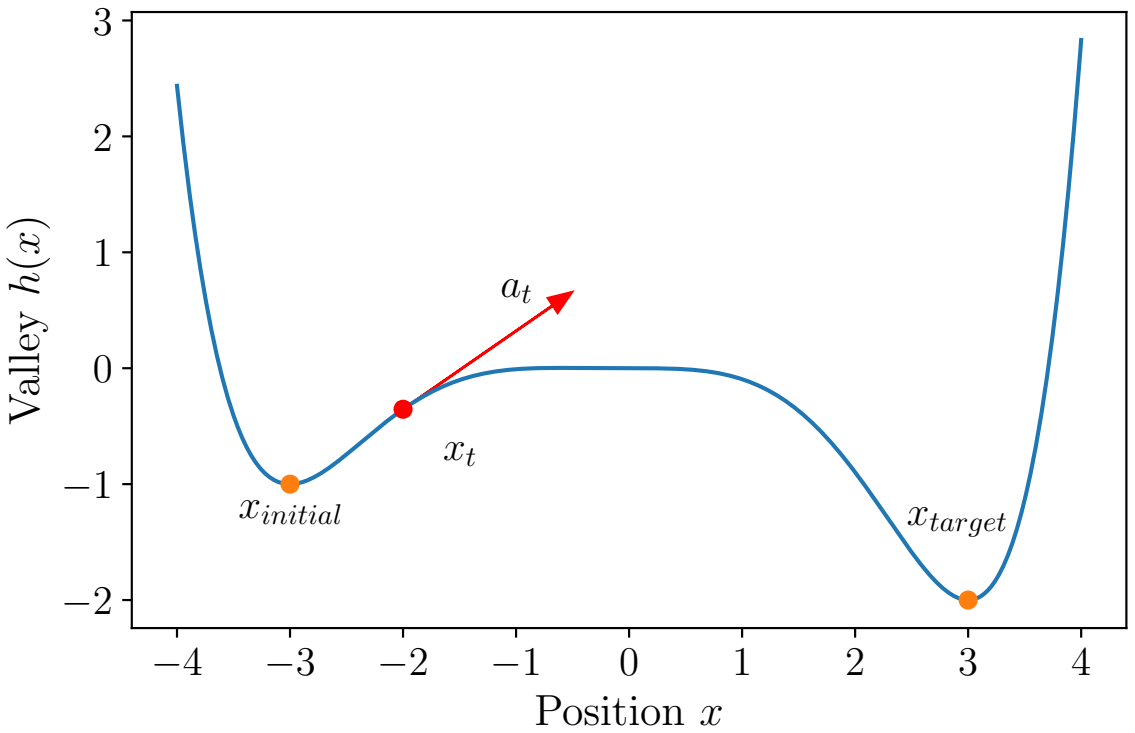
Research question

What is the influence of policy stochasticity on the optimization problem?

Bolland, A., Louppe, G., & Ernst, D. (2023). Policy Gradient Algorithms Implicitly Optimize by Continuation. Transactions on Machine Learning Research.

Illustration

We consider a car moving in a valley, and denote by x its position and by v its speed. The car starts at $x_{initial}$ and perceives rewards proportional to the depth in the valley, an optimal sequence of actions moves the car to x_{target}.

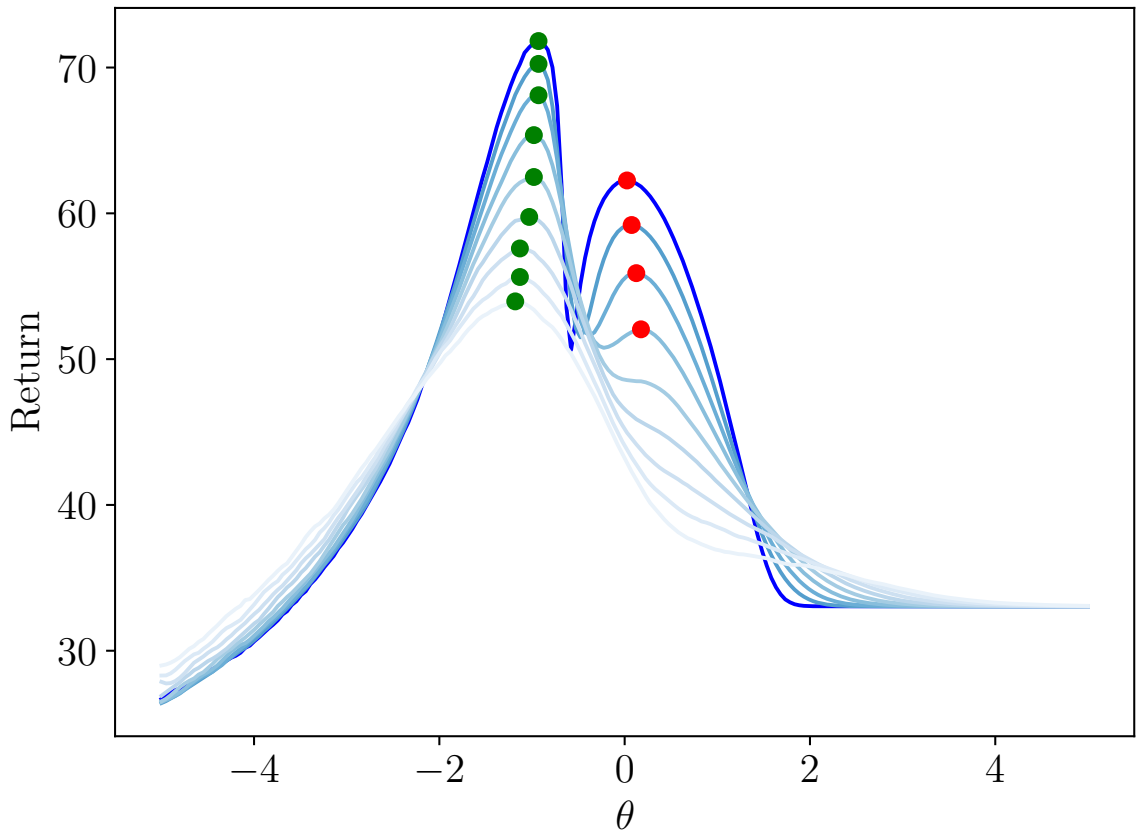


Illustration

We optimize $\pi_\theta(a|s) = \mathcal{N}(a|\mu_\theta(s), \sigma)$, where $\mu_\theta(s) = \theta \times (x - x_{target})$.

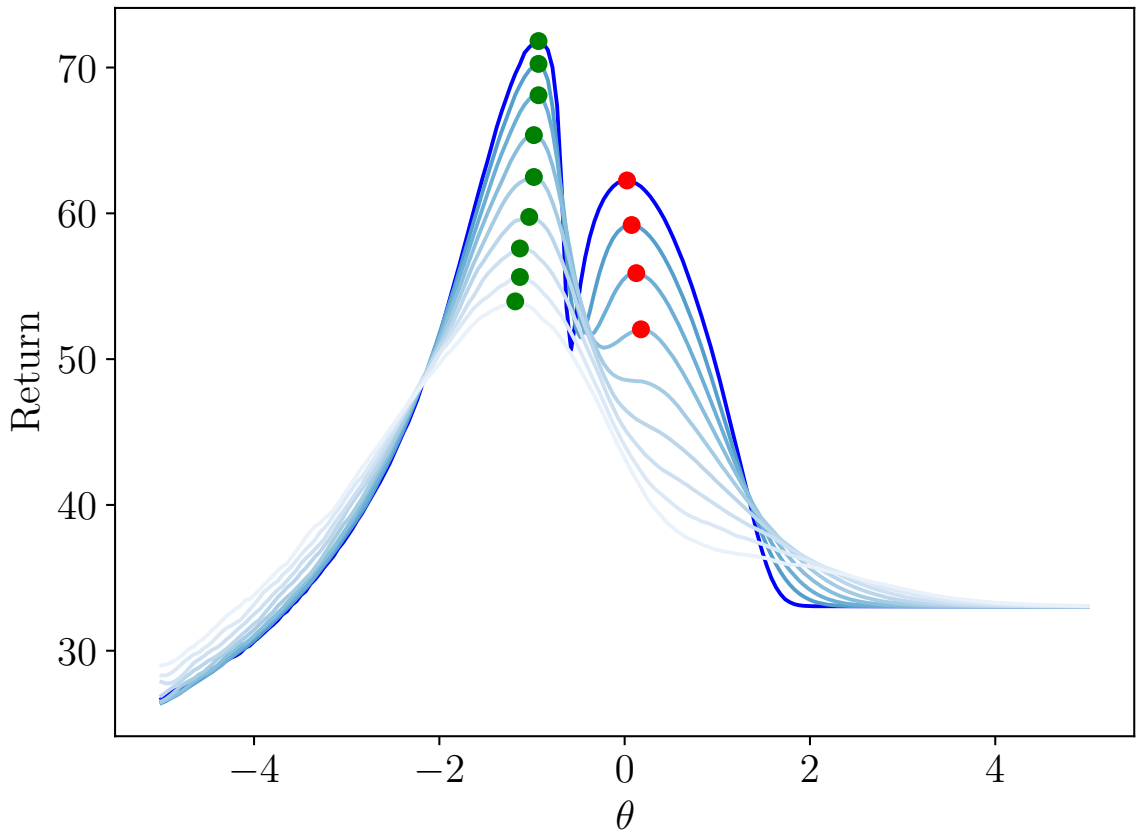
Illustration

We optimize $\pi_\theta(a|s) = \mathcal{N}(a|\mu_\theta(s), \sigma)$, where $\mu_\theta(s) = \theta \times (x - x_{target})$.



Illustration

We optimize $\pi_\theta(a|s) = \mathcal{N}(a|\mu_\theta(s), \sigma)$, where $\mu_\theta(s) = \theta \times (x - x_{target})$.



The return is concave as a function of θ for a sufficiently large σ.

The role of stochasticity

Mirror Policies

For an affine policy μ_θ, there exists an operator Φ_σ such that

$$J(\pi_\theta) = \Phi_\sigma \left(J(\mu_\theta) \right) \ ,$$

where π_θ is an affine Gaussian policy of mean μ_θ and variance σ.

The operator Φ_σ is a filter applied on the expected return of μ_θ.

The role of stochasticity

Mirror Policies

For an affine policy μ_θ, there exists an operator Φ_σ such that

$$J(\pi_\theta) = \Phi_\sigma \left(J(\mu_\theta) \right) \ ,$$

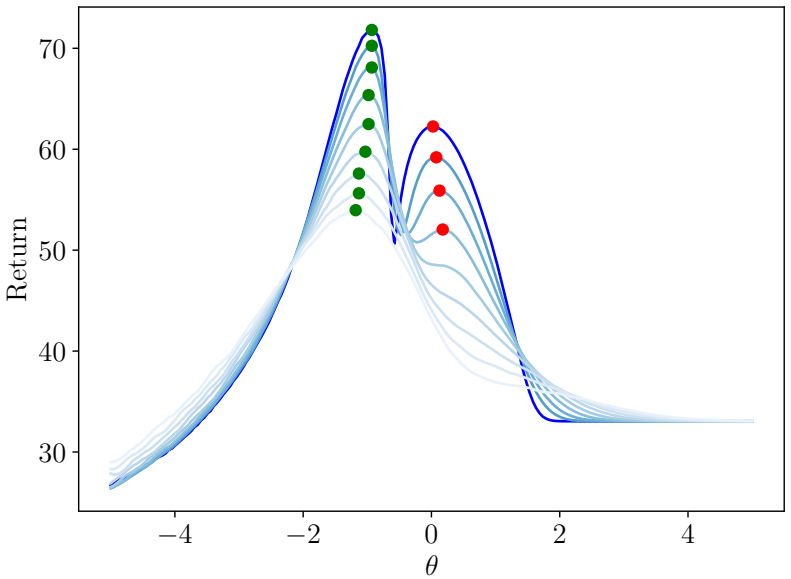
where π_θ is an affine Gaussian policy of mean μ_θ and variance σ.

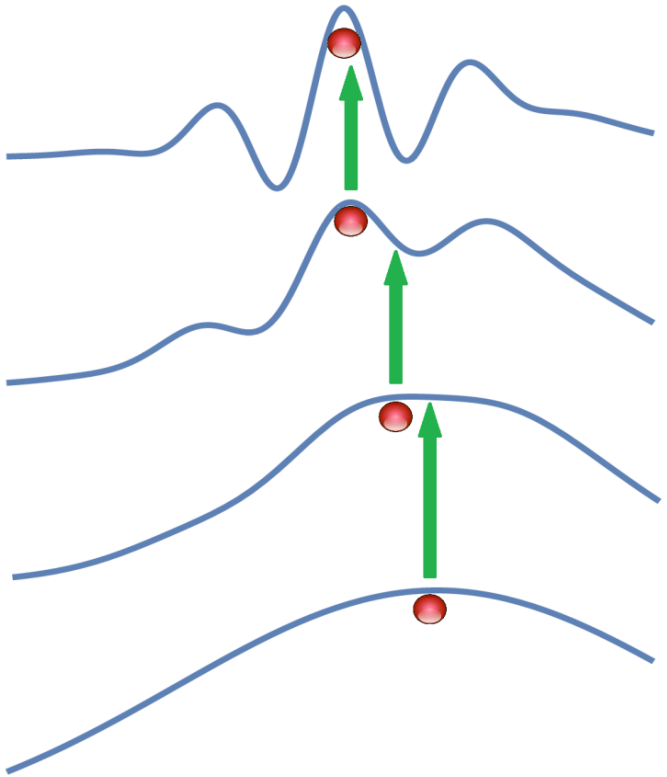
The operator Φ_σ is a filter applied on the expected return of μ_θ.

Increasing the variance of policies can be seen as filtering the expected return.

Optimization by continuation

Optimize a sequence of surrogate objective functions.





Mobahi, H., & Fisher III, J. (2015, February). A theoretical analysis of optimization by Gaussian continuation. In Proceedings of the AAAI Conference on Artificial Intelligence.

Optimization by continuation

Optimization by continuation:

- Directly optimizing the expected return of a deterministic policy $\mu_\theta(s) = \theta \times (x - x_{target})$ would result in a locally optimal policy.
- Optimize the expected return $J(\pi_\theta) = \Phi_\sigma \left(J(\mu_\theta)\right)$ instead.
- We control σ to avoid local extrema.

Optimization by continuation

Optimization by continuation:

- Directly optimizing the expected return of a deterministic policy $\mu_\theta(s) = \theta \times (x - x_{target})$ would result in a locally optimal policy.
- Optimize the expected return $J(\pi_\theta) = \Phi_\sigma\left(J(\mu_\theta)\right)$ instead.
- We control σ to avoid local extrema.

Conversely, when we optimize $J(\pi_\theta)$ by policy gradient and schedule the variance (e.g., adding regularization), we implicitly optimize the policy μ_θ by continuation.

Discussion

Research question

What is the influence of policy stochasticity on the optimization problem?

Discussion

Research question

What is the influence of policy stochasticity on the optimization problem?

- Stochastic policies have a smoother expected return.
- The variance should be adjusted to avoid local extrema.
- There may be advantages to optimizing history-dependent policies.

Learning objective with intrinsic exploration

Intrinsic exploration

Learning objective

Policy gradient algorithms optimize the learning objective $L(\theta)$ by SGA:

$$L(\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda R^{int}(s_t, a_t)\right)\right] = J(\pi_\theta) + J^{int}(\pi_\theta)\,.$$

Intrinsic exploration

Learning objective

Policy gradient algorithms optimize the learning objective $L(\theta)$ by SGA:

$$L(\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda R^{int}(s_t, a_t)\right)\right] = J(\pi_\theta) + J^{int}(\pi_\theta)\,.$$

- Uncertainty-based motivation using model prediction errors.
- Entropy-based motivation using state-action probability.

Intrinsic exploration

Learning objective

Policy gradient algorithms optimize the learning objective $L(\theta)$ by SGA:

$$L(\theta) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda R^{int}(s_t, a_t)\right)\right] = J(\pi_\theta) + J^{int}(\pi_\theta)\,.$$

- Uncertainty-based motivation using model prediction errors.
- Entropy-based motivation using state-action probability.

$$R^s(s, a) = -\log d^{\pi_\theta, \gamma}(s)$$
$$R^a(s, a) = -\log \pi_\theta(a|s)\,.$$

Research question

Research question

What are the required conditions to compute an optimal policy π_θ^* by (stochastic) gradient ascent on a learning objective?

Bolland, A., Lambrechts, G., & Ernst, D. (2024). Behind the myth of exploration in policy gradients. arXiv preprint arXiv:2402.00162.

Convergence of SGA

Let us assume unbiased gradient estimates of the learning objective.

$$\theta \leftarrow \theta + \alpha \hat{d} \qquad \mathbb{E}\left[\hat{d}\right] = \nabla_\theta L(\theta)\,.$$

Convergence of SGA

Let us assume unbiased gradient estimates of the learning objective.

$$\theta \leftarrow \theta + \alpha \hat{d} \qquad \mathbb{E}\left[\hat{d}\right] = \nabla_\theta L(\theta)\,.$$

- SGA converges to a local maximum.
- If the function is concave, SGA converges to the global maximum.

1. Coherence criterion

A learning objective L is ε-coherent if and only if

$$J(\pi_{\theta^*}) - J(\pi_{\theta^\dagger}) \leq \varepsilon \,,$$

where $\theta^* \in \operatorname{argmax}_\theta J(\pi_\theta)$ and where $\theta^\dagger \in \operatorname{argmax}_\theta L(\theta)$.

The optimal parameter $\theta^\dagger$ corresponds to a policy at most suboptimal by ε.

2. Pseudoconcavity criterion

A learning objective L is pseudoconcave if and only if

$$\exists!\, \theta^\dagger : \nabla L(\theta^\dagger) = 0 \ \wedge \ L(\theta^\dagger) = \max_\theta L(\theta)\,.$$

If the criterion is respected, there is a single optimum, and it is possible to globally optimize the learning objective function by (stochastic) gradient ascent.

Learning objective in the hill environment

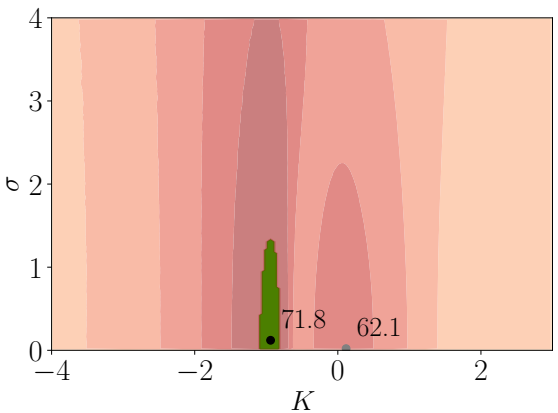
We optimize the policy $\pi_{K,\sigma}(a|s) = \mathcal{N}(a|K \times (x - x_{target}), \sigma)$ with the objective

$$L(K, \sigma) = \mathbb{E}^{\pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda_s R^s(s_t, a_t) \right) \right] .$$

Learning objective in the hill environment

We optimize the policy $\pi_{K,\sigma}(a|s) = \mathcal{N}(a|K \times (x - x_{target}), \sigma)$ with the objective

$$L(K, \sigma) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda_s R^s(s_t, a_t)\right)\right] .$$

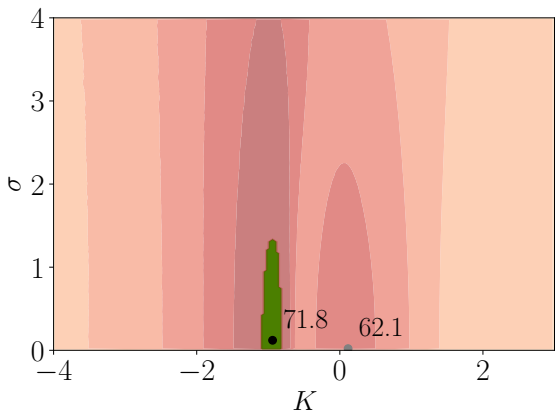


(a) $\lambda_s = 0.05$

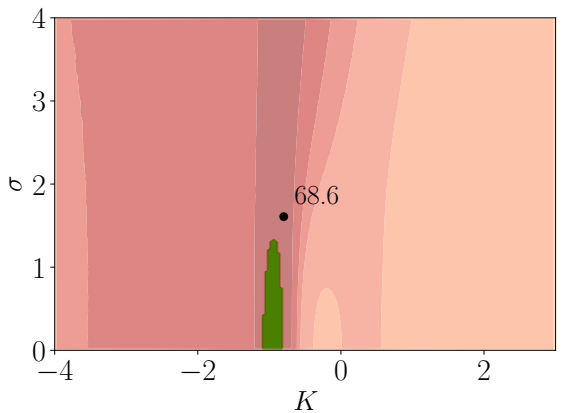
Learning objective in the hill environment

We optimize the policy $\pi_{K,\sigma}(a|s) = \mathcal{N}(a|K \times (x - x_{target}), \sigma)$ with the objective

$$L(K, \sigma) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda_s R^s(s_t, a_t)\right)\right] .$$



(a) $\lambda_s = 0.05$

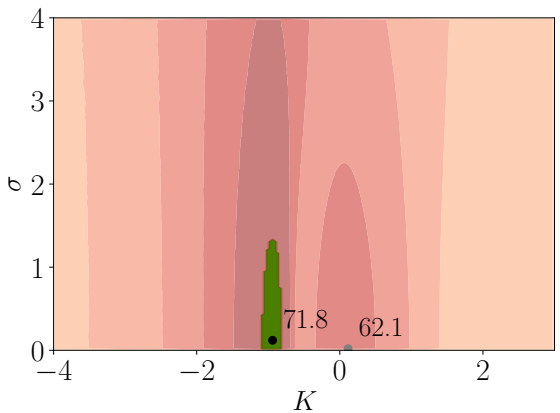


(c) $\lambda_s = 1$

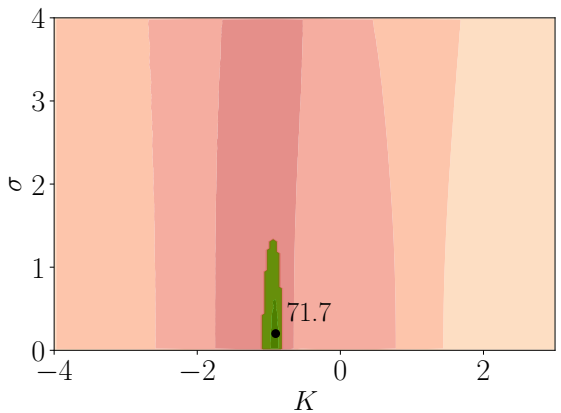
Learning objective in the hill environment

We optimize the policy $\pi_{K,\sigma}(a|s) = \mathcal{N}(a|K \times (x - x_{target}), \sigma)$ with the objective

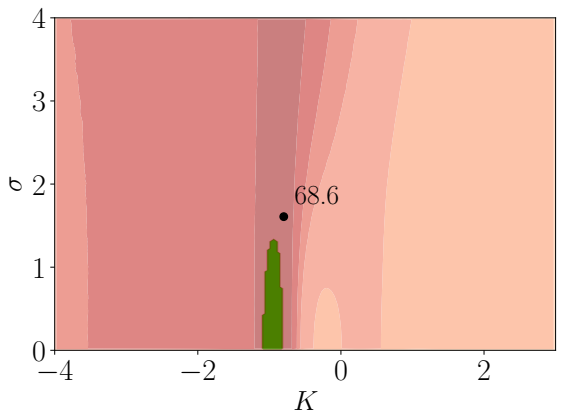
$$L(K, \sigma) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda_s R^s(s_t, a_t)\right)\right] .$$



(a) $\lambda_s = 0.05$



(b) $\lambda_s = 0.1$



(c) $\lambda_s = 1$

Learning objective in the hill environment

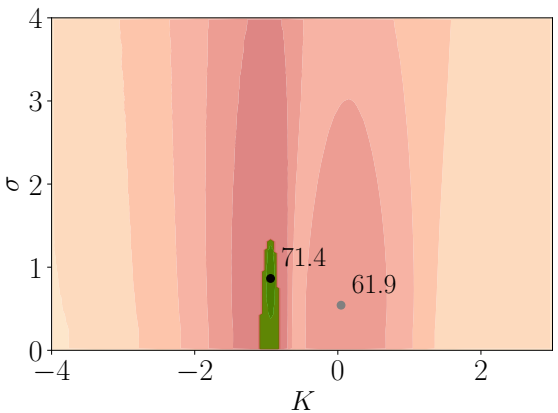
We optimize the policy $\pi_{K,\sigma}(a|s) = \mathcal{N}(a|K \times (x - x_{target}), \sigma)$ with the objective

$$L(K, \sigma) = \mathbb{E}^{\pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda_a R^a(s_t, a_t) \right) \right] .$$

Learning objective in the hill environment

We optimize the policy $\pi_{K,\sigma}(a|s) = \mathcal{N}(a|K \times (x - x_{target}), \sigma)$ with the objective

$$L(K, \sigma) = \mathbb{E}^{\pi_\theta} \left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda_a R^a(s_t, a_t) \right) \right] .$$

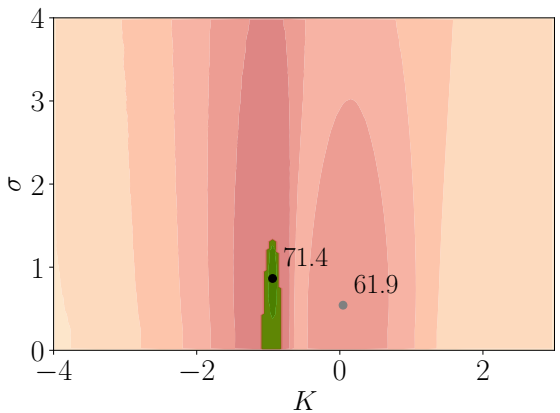


(a) $\lambda_a = 0.01$

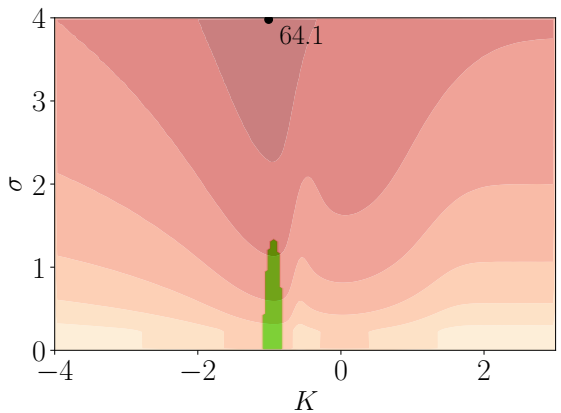
Learning objective in the hill environment

We optimize the policy $\pi_{K,\sigma}(a|s) = \mathcal{N}(a|K \times (x - x_{target}), \sigma)$ with the objective

$$L(K, \sigma) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda_a R^a(s_t, a_t)\right)\right] .$$



(a) $\lambda_a = 0.01$

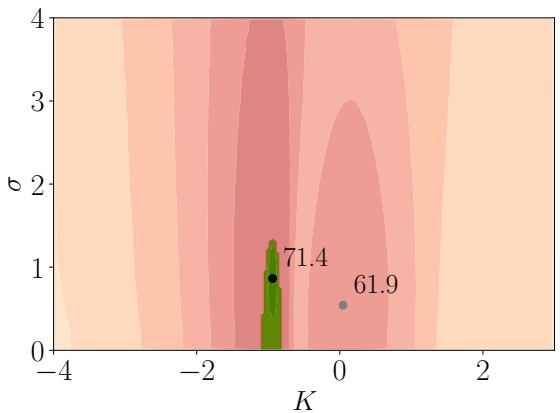


(c) $\lambda_a = 0.5$

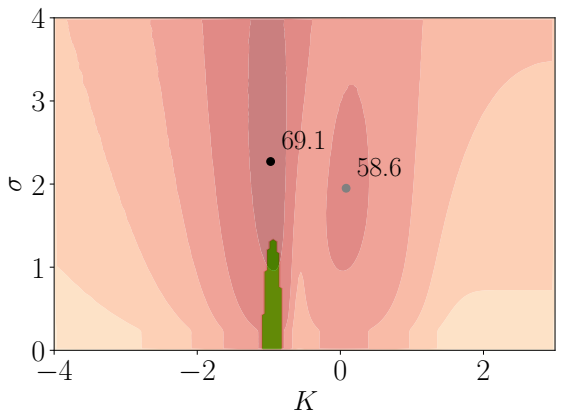
Learning objective in the hill environment

We optimize the policy $\pi_{K,\sigma}(a|s) = \mathcal{N}(a|K \times (x - x_{target}), \sigma)$ with the objective

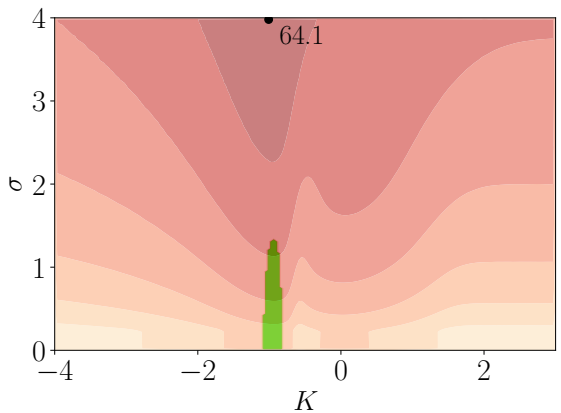
$$L(K, \sigma) = \mathbb{E}^{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \left(R(s_t, a_t) + \lambda_a R^a(s_t, a_t)\right)\right] .$$



(a) $\lambda_a = 0.01$



(b) $\lambda_a = 0.1$



(c) $\lambda_a = 0.5$

Discussion

Research question

What are the required conditions to compute an optimal policy π_θ^* by (stochastic) gradient ascent on a learning objective?

Discussion

Research question

What are the required conditions to compute an optimal policy π_θ^* by (stochastic) gradient ascent on a learning objective?

- Learning objective functions should be coherent and pseudoconcave.
- There is a tradeoff between both criteria.
- Balancing the criteria can be achieved by scheduling the weights.
- Entropy bonuses have no exploration role.

Research question

Pseudoconcave and coherent learning objective functions can be challenging to optimize with stochastic approximations.

Bolland, A., Lambrechts, G., & Ernst, D. (2024). Behind the myth of exploration in policy gradients. arXiv preprint arXiv:2402.00162.

Research question

Pseudoconcave and coherent learning objective functions can be challenging to optimize with stochastic approximations.

Research question

What are the required conditions for exploration to accelerate the convergence speed of the SGA?

Bolland, A., Lambrechts, G., & Ernst, D. (2024). Behind the myth of exploration in policy gradients. arXiv preprint arXiv:2402.00162.

Probability of improvement with SGA

The improvement of f after a SGA step in the direction $\hat{d}$ is

$$X = f(\theta + \alpha\hat{d}) - f(\theta) \approx \alpha \left\langle \hat{d}, \nabla_\theta f(\theta) \right\rangle ,$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product.

Probability of improvement with SGA

The improvement of f after a SGA step in the direction $\hat{d}$ is

$$X = f(\theta + \alpha\hat{d}) - f(\theta) \approx \alpha \left\langle \hat{d}, \nabla_\theta f(\theta) \right\rangle ,$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product.

Asymptotic convergence is deduced from $\mathbb{E}[X]$.

Probability of improvement with SGA

The improvement of f after a SGA step in the direction $\hat{d}$ is

$$X = f(\theta + \alpha\hat{d}) - f(\theta) \approx \alpha \left\langle \hat{d}, \nabla_\theta f(\theta) \right\rangle ,$$

where $\langle \cdot, \cdot \rangle$ is the Euclidean scalar product.

Asymptotic convergence is deduced from $\mathbb{E}[X]$.

Assuming $\mathbb{E}[X] > 0$, is $X > 0$ a rare event?

3. Efficiency criterion

An exploration strategy is δ-efficient if and only if

$$\forall^{\infty}\theta : \mathbb{P}(D > 0) \geq \delta \,,$$

where $D = \langle \hat{d}, \nabla_\theta L(\theta) \rangle$.

Following the ascent direction $\hat{d} \approx \nabla_\theta L(\theta)$ has a probability of increasing the learning objective larger than δ.

4. Attraction criterion

An exploration strategy is δ-attractive if and only if

$$\exists B(\theta^\dagger) : \theta^{int} \in B(\theta^\dagger) \wedge \forall^\infty \theta \in B(\theta^\dagger) : \mathbb{P}(G > 0) \geq \delta \,,$$

where $\theta^{int} = \arg\max_\theta J^{int}(\pi_\theta)$, $B(\theta^\dagger)$ is a ball centered in $\theta^\dagger$, and $G = \langle \hat{d}, \nabla_\theta J(\pi_\theta) \rangle$.

If the criterion is respected for large δ, policy gradients will tend to improve the expected return of the policy if it approaches θ^{int} and enters the ball $B(\theta^\dagger)$.

Sparse reward environment – Simple maze

Let us consider a maze environment consisting of a horizontal corridor.

Sparse reward environment – Simple maze

Let us consider a maze environment consisting of a horizontal corridor.

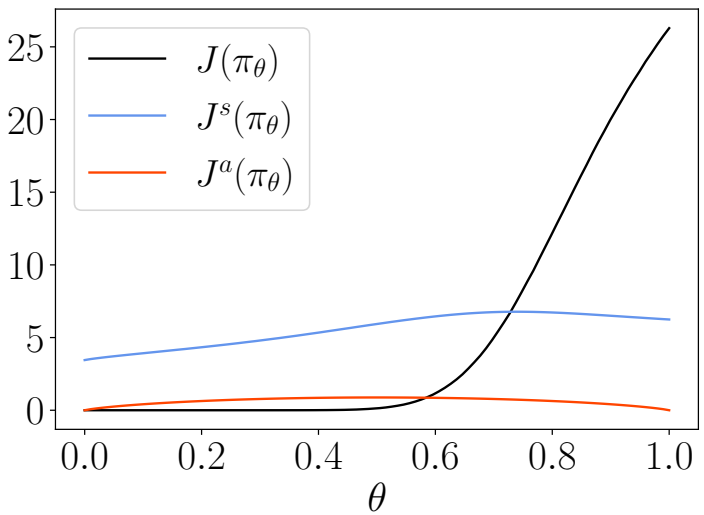
We optimize a one-parameter policy:

$$\pi_\theta(a|s) = \begin{cases} \theta & \text{if } a = \text{RIGHT} \\ 1-\theta & \text{if } a = \text{LEFT}\,. \end{cases}$$

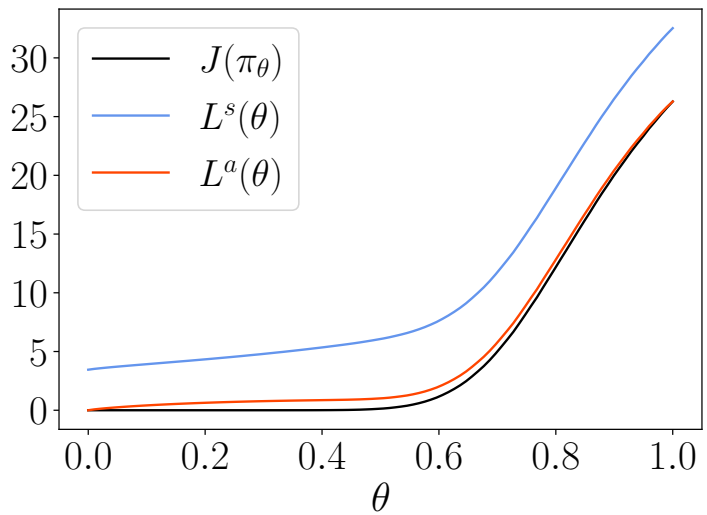
Learning objective functions in the maze

We consider two intrinsic reward bonuses:

$$R^s(s,a) = -\log d^{\pi_\theta,\gamma}(s)$$
$$R^a(s,a) = -\log \pi_\theta(a|s)\,.$$



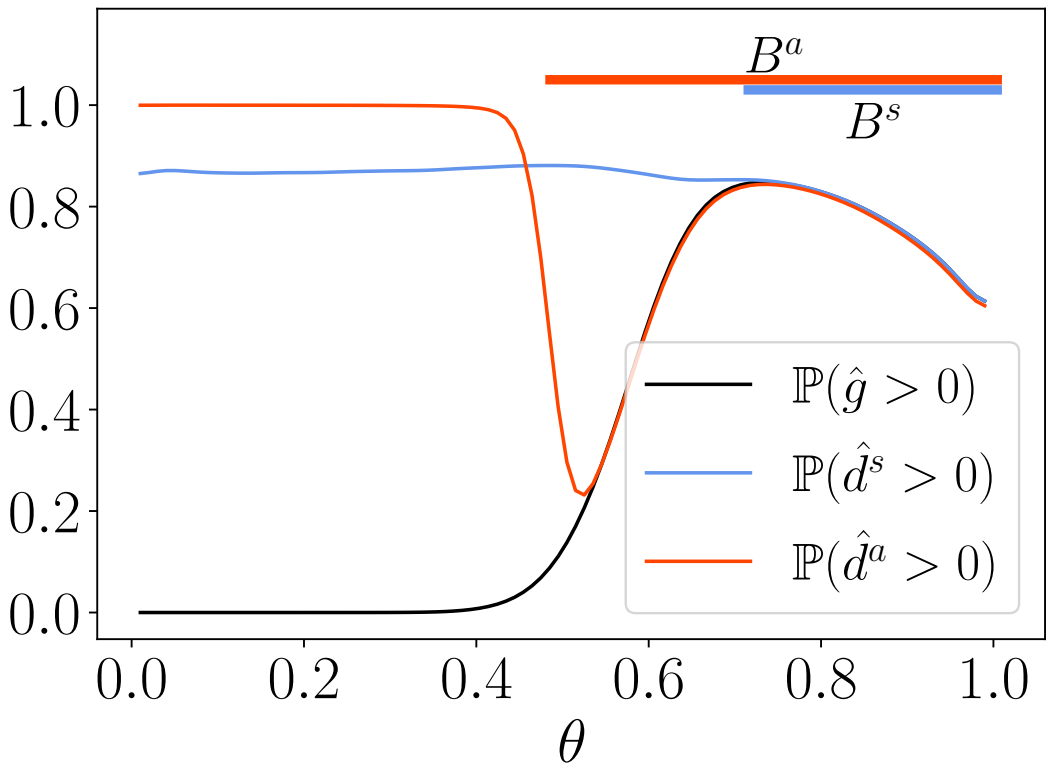
(a) Return



(b) Learning objectives

Probability of improving the return in the maze

Let us compute the probability that the gradient is in the correct direction.



Discussion

Research question

What are the required conditions for exploration to accelerate the convergence speed of the SGA?

Discussion

Research question

What are the required conditions for exploration to accelerate the convergence speed of the SGA?

- Exploration improves the stochastic ascent directions.
- The analysis is valid for any surrogate learning objective.

Exploring future states and actions

Research question

What is a good exploration strategy?

- State-action space exploration.
- Long-term exploration.
- Simple and efficient to explore.

Bolland, A., Lambrechts, G., & Ernst, D. (2024). Off-Policy Maximum Entropy RL with Future State and Action Visitation Measures. arXiv preprint arXiv:2412.06655.

Research question

What is a good exploration strategy?

- State-action space exploration.
- Long-term exploration.
- Simple and efficient to explore.

Research question

What intrinsic reward bonus allows *good exploration*?

Bolland, A., Lambrechts, G., & Ernst, D. (2024). Off-Policy Maximum Entropy RL with Future State and Action Visitation Measures. arXiv preprint arXiv:2412.06655.

Future states and actions

In a state s, followed by an action a, we want all future states and actions to be uniformly distributed.

Future states and actions

In a state s, followed by an action a, we want all future states and actions to be uniformly distributed.

How to measure the next states and actions?

$$d^{\pi,\gamma}(\bar{s}|s,a) = (1-\gamma)\sum_{\Delta=1}^{\infty}\gamma^{\Delta}p_{\Delta}^{\pi}(\bar{s}|s,a)$$

$$d^{\pi,\gamma}(\bar{s},\bar{a}|s,a) = \pi(\bar{a}|\bar{s})d^{\pi,\gamma}(\bar{s}|s,a)\ .$$

Future features

In a state s, followed by an action a, we want all future features to be uniformly distributed.

Future features

In a state s, followed by an action a, we want all future features to be uniformly distributed.

Stochastic projection of the states and actions.

$$q^{\pi}(z|s,a) = \int h(z|\bar{s},\bar{a})\pi(\bar{a}|\bar{s})d^{\pi,\gamma}(\bar{s}|s,a)\, d\bar{s}\, d\bar{a}\,.$$

It can be computed solely based on $d^{\pi,\gamma}(\bar{s}|s,a)$ ...

Exploring with maximum entropy RL

In a state s, followed by an action a, we want all future features to be distributed according to q^*.

Exploring with maximum entropy RL

In a state s, followed by an action a, we want all future features to be distributed according to q^*.

This intrinsic behavior is enforced with an intrinsic reward.

$$\begin{aligned} R^{int}(s,a) &= -KL_z\left[q^\pi(z|s,a) \| q^*(z)\right] \\ &= \mathop{\mathbb{E}}_{z\sim q^\pi(\cdot|s,a)} \left[\log q^*(z) - \log q^\pi(z|s,a)\right] . \end{aligned}$$

Efficient intrinsic reward learning

The distribution of future states is the fixed-point of a contractive operator.

$$\mathcal{T}^{\pi} d^{\pi,\gamma}(\bar{s}|s,a) = d^{\pi,\gamma}(\bar{s}|s,a) \; .$$

Efficient intrinsic reward learning

The distribution of future states is the fixed-point of a contractive operator.

$$\mathcal{T}^{\pi} d^{\pi,\gamma}(\bar{s}|s,a) = d^{\pi,\gamma}(\bar{s}|s,a) \; .$$

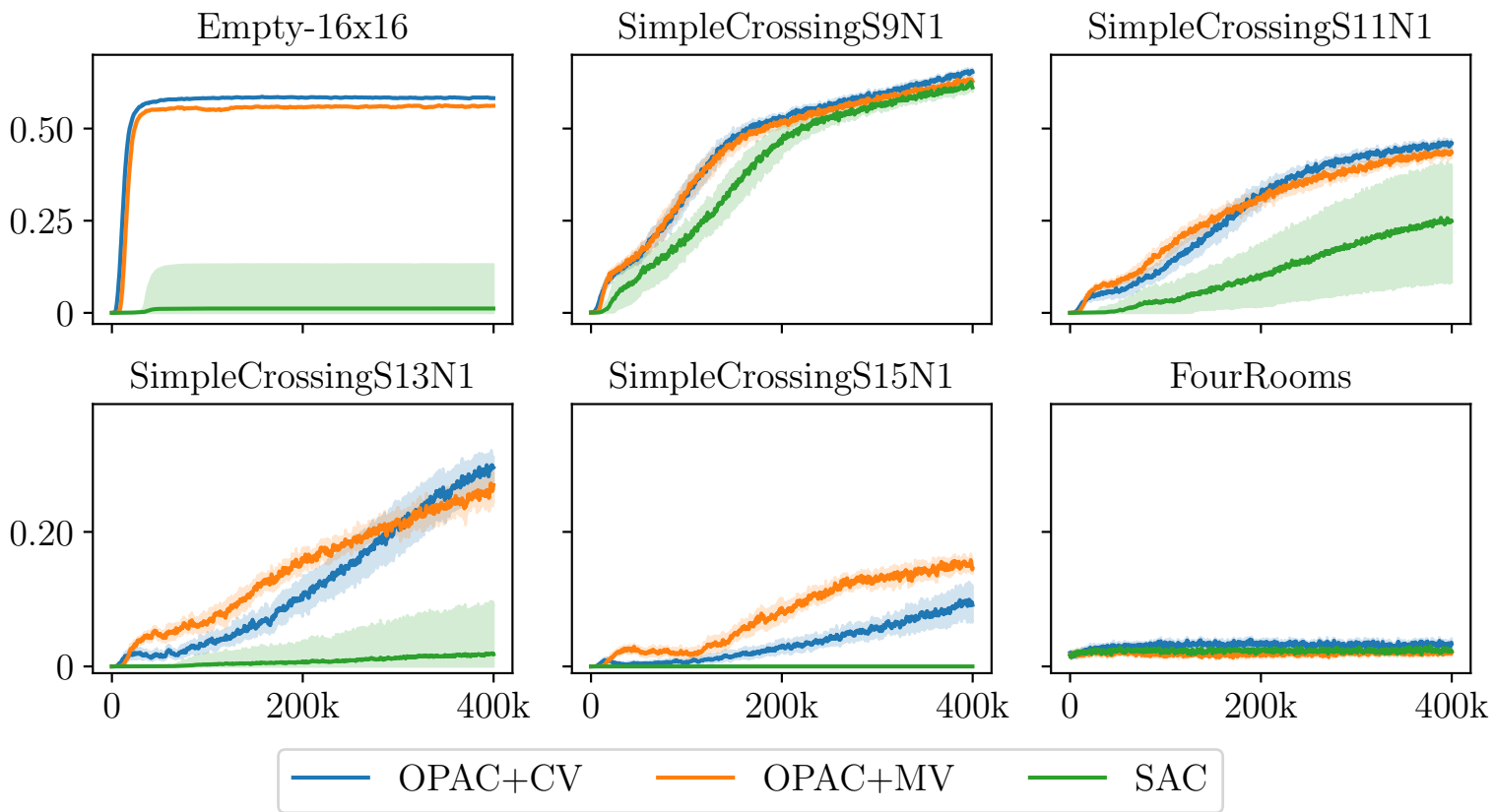
The fixed-point can be estimated with minimum cross-entropy.

$$\arg\min_{\psi} \mathop{\mathbb{E}}_{\substack{s,a\sim g(\cdot,\cdot) \\ \bar{s}\sim(\mathcal{T}^{\pi})^{N} d_{\psi}(\cdot|s,a)}} [-\log d_{\psi}(\bar{s}|s,a)] \; .$$

It can be solved off-policy as TD-like method.

Minigrid control

Expected return of policies.



Discussion

Research question

What intrinsic reward bonus allows *good exploration*?

Discussion

Research question

What intrinsic reward bonus allows *good exploration*?

- The algorithm meets intuitive exploration requirements.
- The future state distribution has many other useful applications.
- How does it influence the learning objective function?

Conclusion

Conclusion

Exploration for learning optimal policies.

- Helps improving the convergence of policy gradient methods.
- Intrinsic motivation is widely inspired by intuition.

Conclusion

Exploration for learning optimal policies.

- Helps improving the convergence of policy gradient methods.
- Intrinsic motivation is widely inspired by intuition.

Other exploration purposes.

- Unsupervised discovering of behaviors.
- Unsupervised environment discovering.
- Unsupervised feature extraction.
- Generalization across tasks.