Policy Improvement and KL Divergences

Alan Chan

Treaty 6 Land Acknowledgement

More Acknowledgements

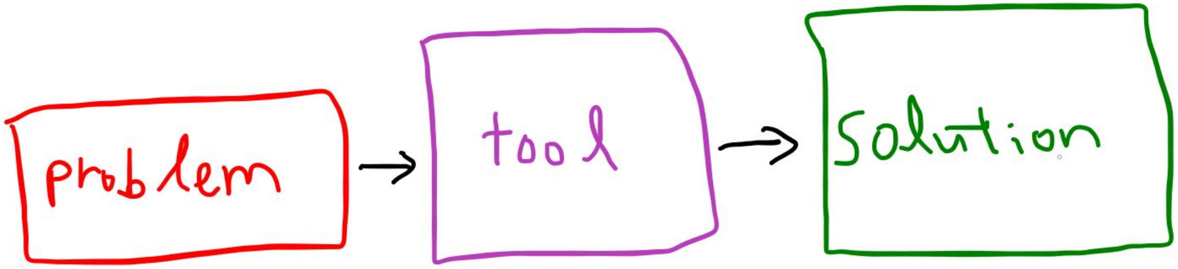
Outline

1. Overview
2. Background
3. Theory
4. Small Experiments
5. Large Experiments
6. Concluding Thoughts

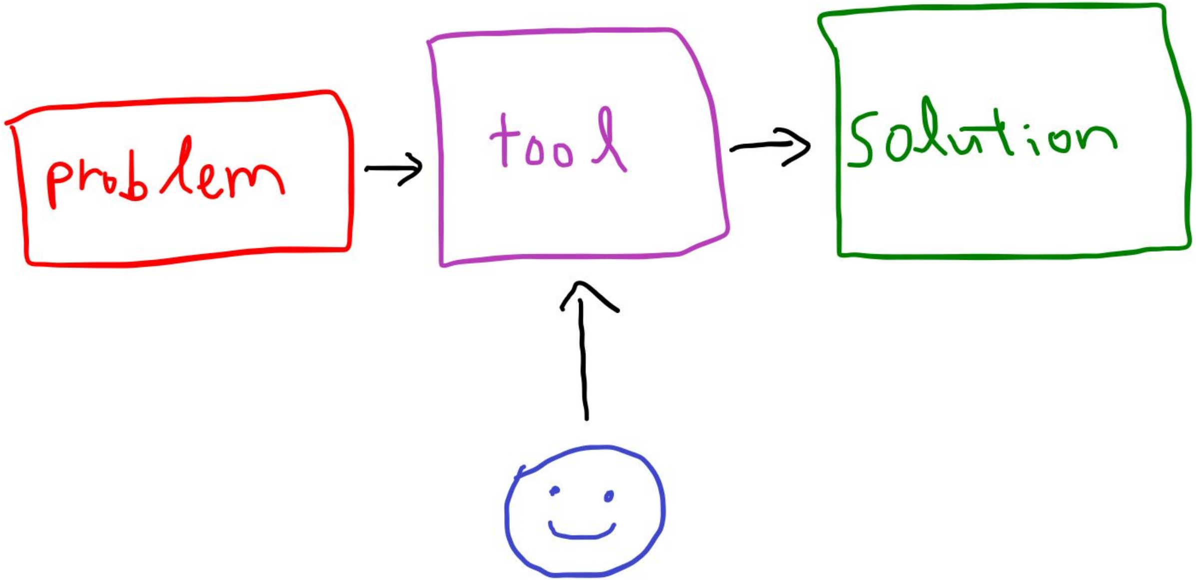
Overview

Artificial Intelligence

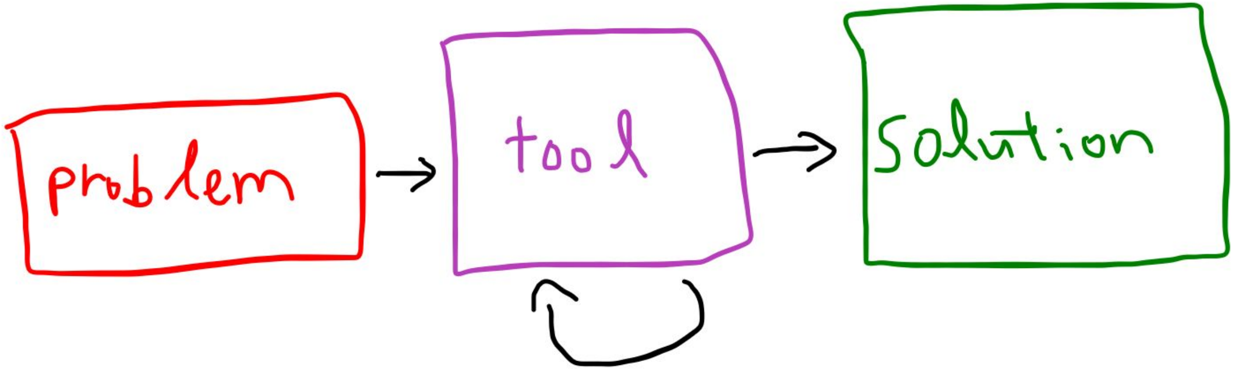
Artificial Intelligence



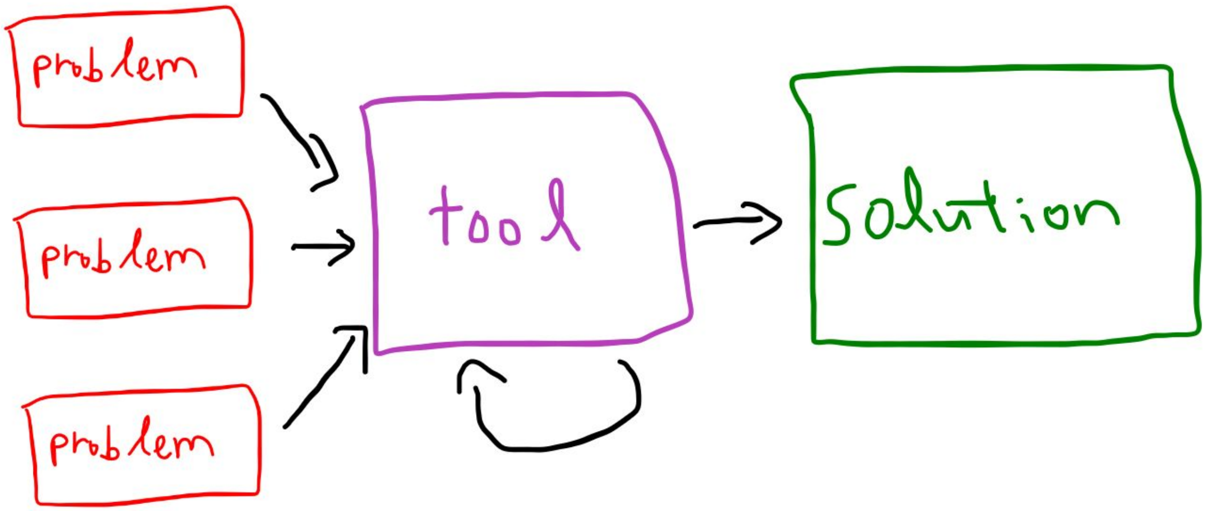
Human-designed Tools



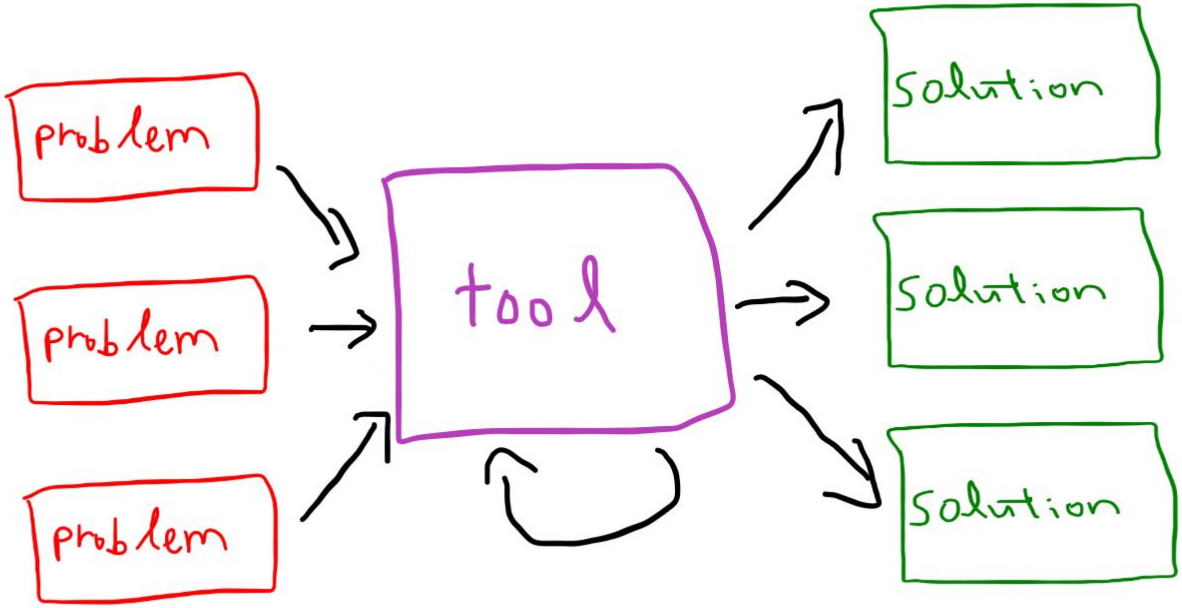
Tools that design themselves



Different Problems



Learning

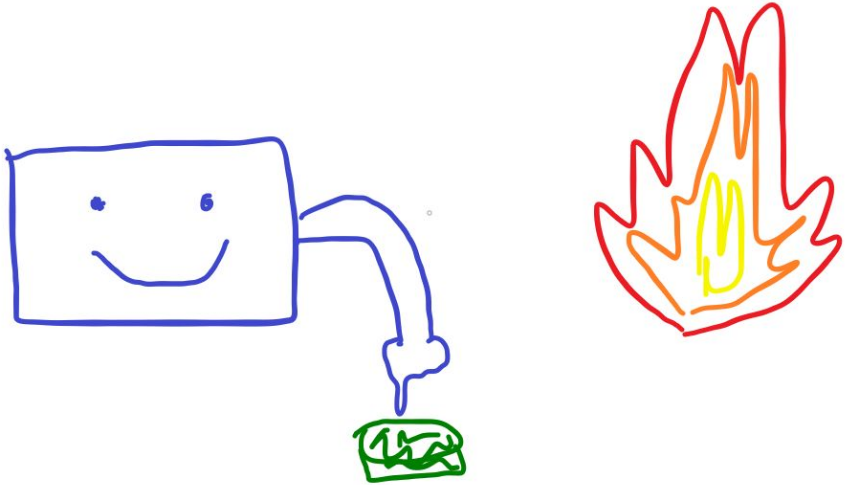


Reinforcement Learning

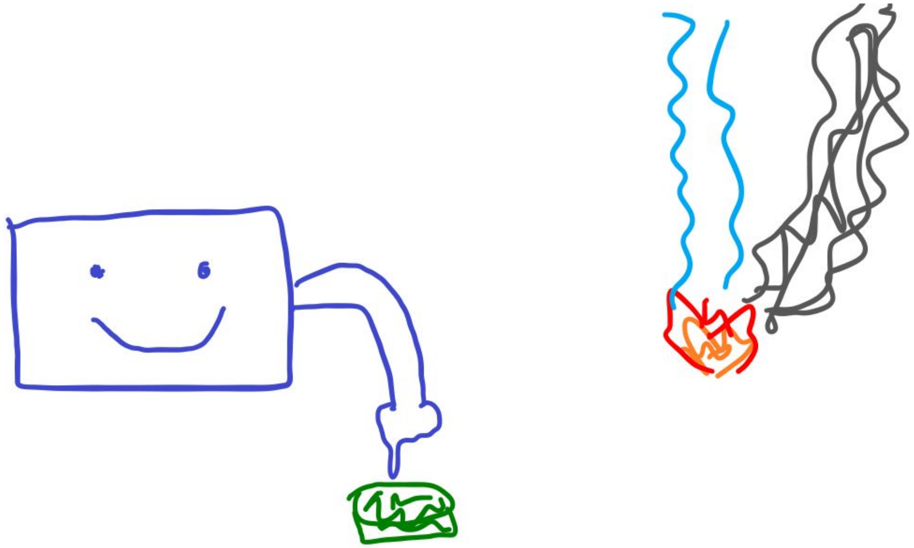
Agent

Agent observes a state

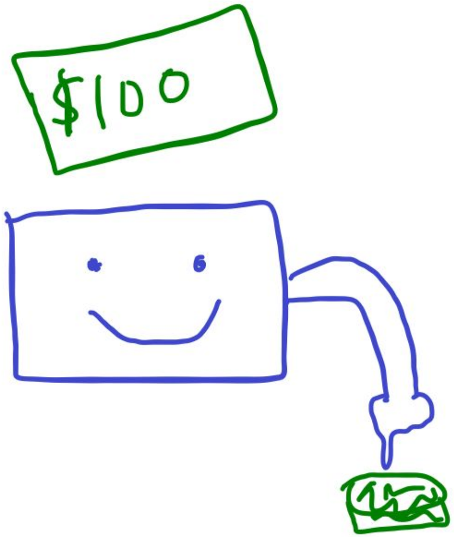
Agent acts (according to a policy)

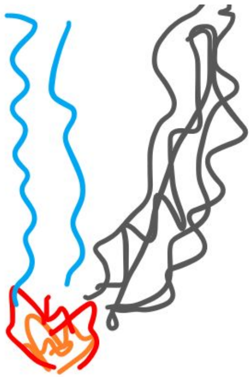


Agent observes a new state

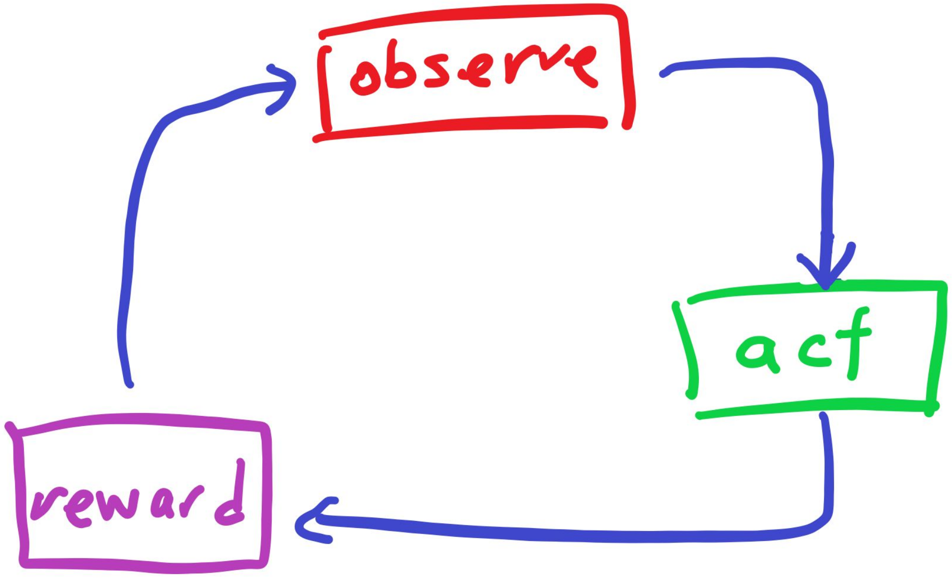


Agent gets reward and learns/improves policy





Rinse and repeat

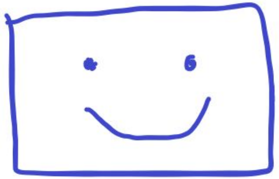


Goal of RL agents

Maximize the return

Maximum-Entropy RL

One button is easy



Many buttons is hard

In a nutshell

Maximum-entropy RL =

“goal” reward

+

“trying different things” reward

So what is this thesis about?

An analysis of the **policy improvement** properties of **two objective functions** in **MERL**

Which objective functions?

Forward KL Divergence

Reverse KL Divergence

Related Work

- Entropy regularisation (**Ziebart**, 2010; **Levine**, 2018; **Haarnoja et al.**, 2018; **Ahmed et al.**, 2019; **Mei et al.**, 2020)
- API (**Kakade and Langford**, 2002; **Perkins and Pendrith**, 2002; **Perkins and Precup**, 2003; **Bertsekas**, 2011; **Scherrer and Geist**, 2014)
- Actor-critic + policy gradient (**Sutton** 1984; **Williams** 1992; **Konda and Tsitsiklis**, 2000; **Sutton et al.**, 2000; **Silver et al.**, 2014; **Mnih et al.**, 2016; **Schulman et al.**, 2016; **Fellows et al.**, 2019; **Ryu et al.**, 2020)
- KL Divergence (**Peters et al.**, 2010; **Neumann**, 2011; **Levine**, 2018)

Contributions

1. Average policy improvement for **RKL**
2. FKL **counterexample**
3. FKL improvement with **additional conditions**
4. **Empirical** comparisons

Background

RL and MDPs

MDP = 5 things

$$\left(\_,\ \_,\ \_,\ \_,\ \_\right)$$

State space

$$(\mathcal{S}, \_, \_, \_, \_)$$

Action space

$$(\mathcal{S}, \mathcal{A}, \_, \_, \_)$$

Transition kernel

$$(\mathcal{S}, \mathcal{A}, \_, \_, p)$$

$$p(s' \mid s, a)$$

Reward function

$$(\mathcal{S}, \mathcal{A}, r, \_, p)$$

$$r(s, a)$$

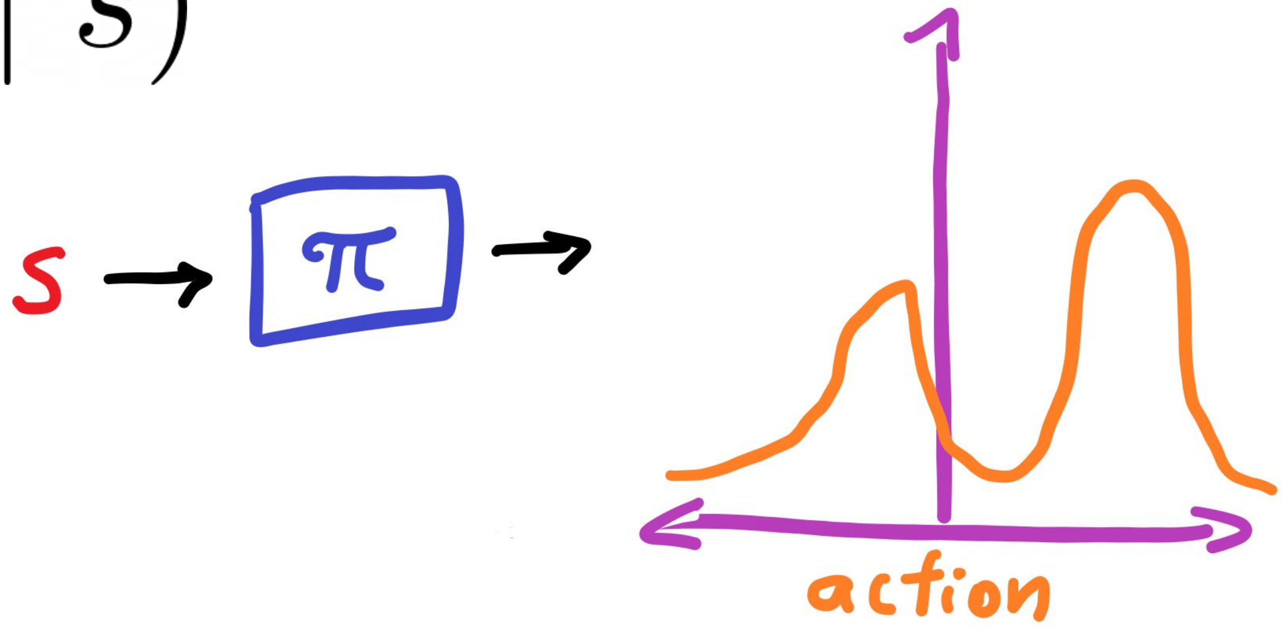
Discount factor

$$(\mathcal{S}, \mathcal{A}, r, \gamma, p)$$

$$G_t := \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

Policies

$$\pi(a \mid s)$$



Value functions

$$G := \sum_{t=0}^{\infty} \gamma^t r(s_t, a_t)$$

$$V^\pi(s) := \mathbb{E}_\pi[G \mid S_0 = s]$$

$$Q^\pi(s, a) := \mathbb{E}_\pi[G \mid S_0 = s, A_0 = a]$$

The Goal

$$\rho(s)$$

$$\max_{\pi} \sum_{s} \rho(s) V^{\pi}(s)$$

There are too many policies!

S states, A actions in each state

at least A^S possible policies

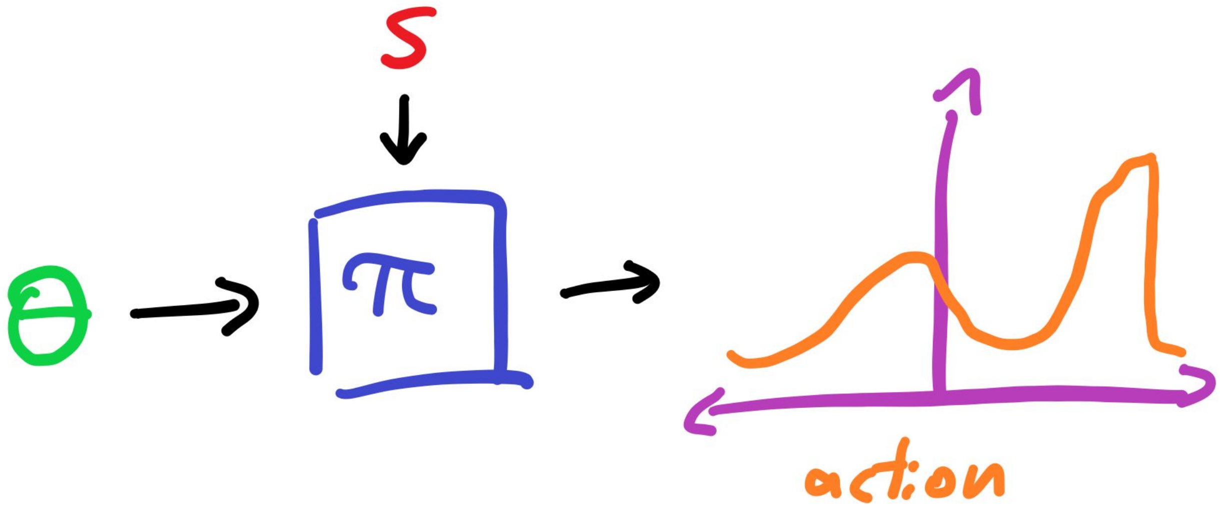
Parameterize the policy

$$\pi_\theta(a \mid s)$$

$$\theta \longrightarrow \boxed{\pi}$$

Parameterize the policy

$$\pi_\theta(a \mid s)$$



Policy Optimization

$$\eta(\theta) := \sum_s \rho(s) V^{\pi_\theta}(s)$$

$$\max_\theta \eta(\theta)$$

Policy Gradient Theorem (Sutton et al., 2000)

$$\nabla \eta(\theta) = \sum_{s,a} d^{\pi_\theta}(s) Q^{\pi_\theta}(s,a) \underbrace{\pi_\theta(a \mid s) \nabla \log \pi_\theta(a \mid s)}_{\nabla \pi_\theta(a|s)}$$

Policy Gradient Theorem (Sutton et al., 2000)

$$\nabla \eta(\theta) = \sum_{s,a} d^{\pi_\theta}(s) Q^{\pi_\theta}(s,a) \underbrace{\pi_\theta(a \mid s) \nabla \log \pi_\theta(a \mid s)}_{\nabla \pi_\theta(a|s)}$$

Future state visitation distribution

Policy Gradient Theorem (Sutton et al., 2000)

$$\nabla \eta(\theta) = \sum_{s,a} d^{\pi_\theta}(s) Q^{\pi_\theta}(s,a) \underbrace{\pi_\theta(a \mid s) \nabla \log \pi_\theta(a \mid s)}_{\nabla \pi_\theta(a|s)}$$

Action-value function

How to use it?

$$s \sim d^{\pi_\theta}(s) \qquad a \sim \pi_\theta(\cdot \mid s)$$

$$G \sim Q^{\pi_\theta}(s, a)$$

$$\theta_{t+1} \leftarrow \theta_t + \beta G \nabla \log \pi_{\theta_t}(a \mid s)$$

Learn the action-value

$$\hat{Q}(s, a) \approx Q^{\pi_\theta}(s, a)$$

Use some TD-like method

Improve the policy

$$(s, a, r, s')$$

$$\theta_{t+1} \leftarrow \theta_t + \beta \hat{Q}(s, a) \nabla \log \pi_{\theta_t}(a \mid s)$$

Improve the policy

$$(s, a, r, s')$$

$$\theta_{t+1} \leftarrow \theta_t + \beta \hat{Q}(s, a) \nabla \log \pi_{\theta_t}(a \mid s)$$

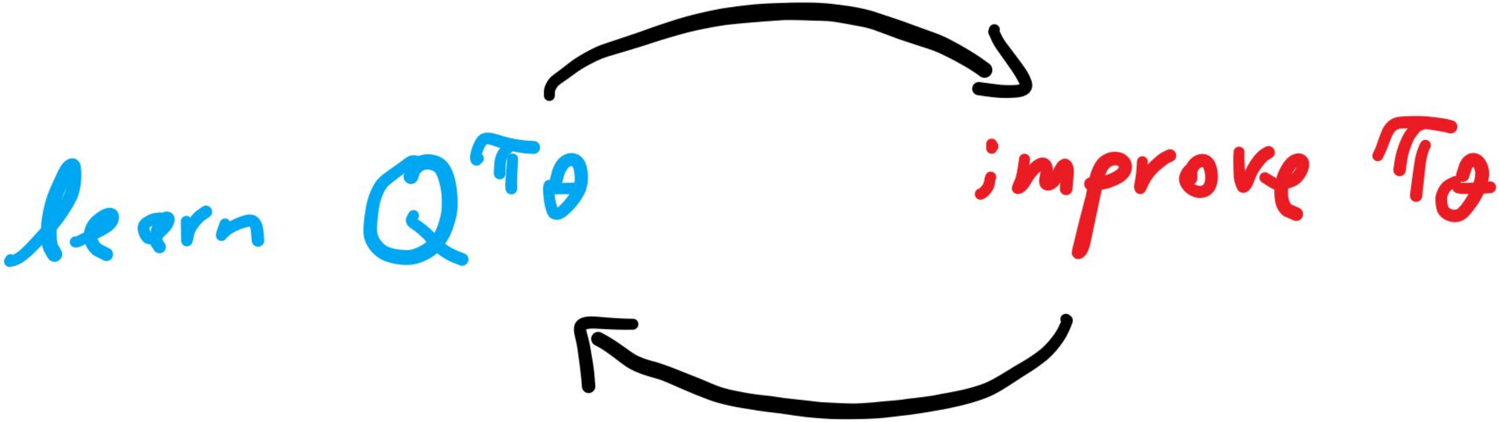
$$\theta_{t+1} \leftarrow \theta_t + \beta \gamma^t \hat{Q}(s, a) \nabla \log \pi_{\theta_t}(a \mid s)$$

Remark on bias

1. Not including γ^t

2. Incompatible $\hat{Q}(s, a)$

Approximate Policy Iteration



How else could we improve the policy?

$$\nabla \eta(\theta) = \sum_{s,a} d^{\pi_\theta}(s) Q^{\pi_\theta}(s, a) \nabla \pi_\theta(a \mid s)$$

How else could we improve the policy?

$$\nabla \eta(\theta) = \nabla \left(\sum_{s,a} d^{\mu}(s) Q^{\mu}(s,a) \pi_\theta(a \mid s) \right)_{\mu=\pi_\theta}$$

“Approximate” objective

$$\eta(\theta) \approx \sum_{s,a} d^{\mu}(s) Q^{\mu}(s, a) \pi_{\theta}(a \mid s)$$

Be greedy with respect to action-values

$$\eta(\theta_{t+1}) \approx \sum_{s,a} d^{\theta_t}(s) Q^{\theta_t}(s,a) \pi_{\theta_{t+1}}(a \mid s)$$

$$\pi_{\theta_{t+1}}(\cdot \mid s) = \mathrm{argmax}_\pi \sum_a Q^{\pi_{\theta_t}}(s,a) \pi(a \mid s)$$

$$\pi_{\theta_{t+1}}(a \mid s) = 1_{\mathrm{argmax}_b Q^{\pi_{\theta_t}}(s,b)}$$

Other target policy distributions

$$\pi_{\theta_{t+1}}(a \mid s) \approx \underbrace{\text{some better policy}}_{\text{based on } Q^{\pi_{\theta_t}}(s,\cdot)}$$

One choice of distribution

$$\pi_{\theta_{t+1}}(a \mid s) \propto \exp(\underbrace{\tau^{-1}}_{\tau > 0} Q^{\pi_{\theta_t}}(s, a))$$

Entropy

$$\mathcal{H}(\pi(\cdot \mid s)) := -\sum_{a} \pi(a \mid s) \log \pi(a \mid s)$$

Soft greedification

$$\pi_{\theta_{t+1}}(a \mid s) \propto \exp(\underbrace{\tau^{-1}}_{\tau>0} Q^{\pi_{\theta_t}}(s, a))$$

$$\pi_{\theta_{t+1}}(\cdot \mid s) = \operatorname{argmax}_{\pi} \sum_{a} Q^{\pi_{\theta_t}}(s, a)\pi(a \mid s) + \tau\mathcal{H}(\pi(\cdot \mid s))$$

Objective function aside

$$\eta(\theta) := \sum_{s} \rho(s) V^{\pi_\theta}(s)$$

Soft value functions

$$G_\tau := \sum_{t=0}^{\infty} \gamma^t \left(r(s_t, a_t) - \tau \log \pi(a_t \mid s_t) \right)$$

$$V_\tau^\pi(s) := \mathbb{E}_\pi [G_\tau \mid S_0 = s]$$

Notational aside

$$\mathcal{B}_\tau(Q)(s,\cdot) \leftrightarrow \exp(\tau^{-1}Q(s,\cdot))$$

Using estimated action-values

$$\cancel{\pi_{\theta_{t+1}}(a \mid s) = \mathcal{B}_\tau(\hat{Q})(s,a)}$$

KL Divergence

$$\mathrm{KL}(p \parallel q) = \sum_{a} p(a) \log \frac{p(a)}{q(a)}$$

But which KL?

RKL $$\min_\theta \mathrm{KL}(\pi_\theta(\cdot \mid s) \parallel \mathcal{B}_\tau(\hat{Q})(s, \cdot))$$

FKL $$\min_\theta \mathrm{KL}(\mathcal{B}_\tau(\hat{Q})(s, \cdot) \parallel \pi_\theta(\cdot \mid s))$$

Can also take limits of temperature

Hard RKL $$\min_\theta - \sum_a \hat{Q}(s, a)\pi_\theta(a \mid s)$$

Hard FKL $$\min_\theta - \log \pi_\theta \left(\mathrm{argmax}_a \hat{Q}(s, a) \mid s\right)$$

Main question

If I (approximately) minimize with either KL, how good is the resulting policy?

What's next?

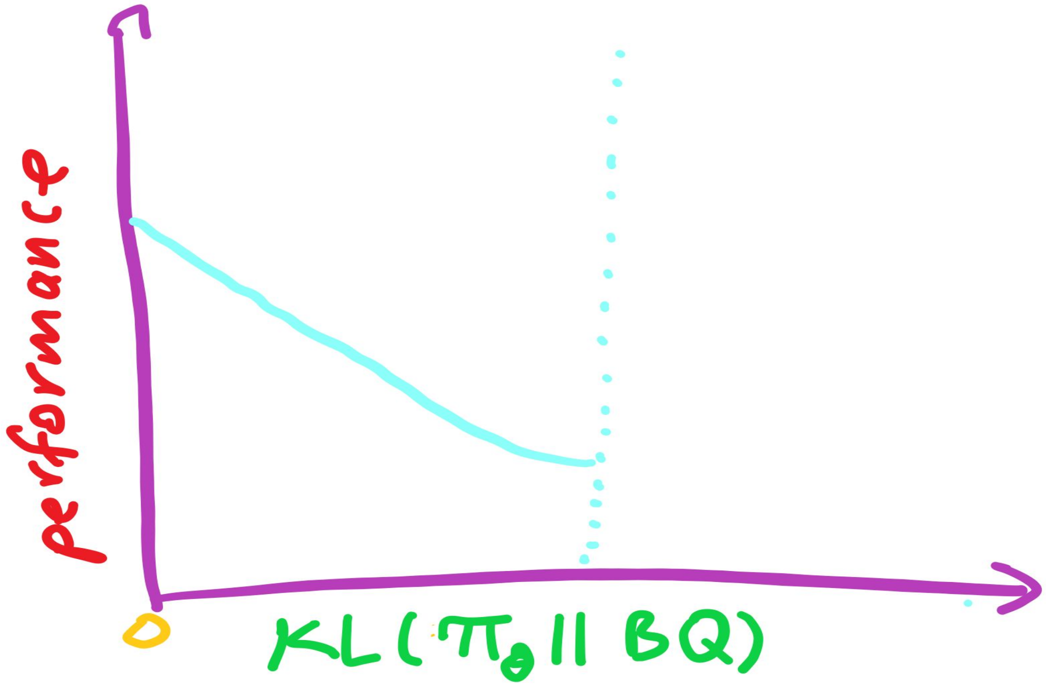
Theoretical and empirical comparison of the FKL and RKL

Theory

Previous Work

Previous Work (Haarnoja et al., 2018)

Previous Work (Haarnoja et al., 2018)



Previous Work (Haarnoja et al., 2018)

Lemma 1 (Policy Improvement under RKL Reduction, Restatement of Lemma 2 (Haarnoja, Zhou, et al., 2018)). *For $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, if for all s*

$$\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \,\|\, \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot)) \leq \mathrm{KL}(\pi_{\text{old}}(\cdot \mid s) \,\|\, \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot))$$

then $Q_\tau^{\pi_{\text{new}}}(s, a) \geq Q_\tau^{\pi_{\text{old}}}(s, a)$ for all (s, a) and $\tau > 0$.

Previous Work (Haarnoja et al., 2018)

Lemma 1 (Policy Improvement under RKL Reduction, Restatement of Lemma 2 (Haarnoja, Zhou, et al., 2018)). *For* $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, *if for all* s

$$\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \,\|\, \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot)) \leq \mathrm{KL}(\pi_{\text{old}}(\cdot \mid s) \,\|\, \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot))$$

then $Q_\tau^{\pi_{\text{new}}}(s, a) \geq Q_\tau^{\pi_{\text{old}}}(s, a)$ *for all* (s, a) *and* $\tau > 0$.

Previous Work (Haarnoja et al., 2018)

Lemma 1 (Policy Improvement under RKL Reduction, Restatement of Lemma 2 (Haarnoja, Zhou, et al., 2018)). *For* $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, *if for all* s

$$\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \parallel \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s,\cdot)) \leq \mathrm{KL}(\pi_{\text{old}}(\cdot \mid s) \parallel \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s,\cdot))$$

then $Q_\tau^{\pi_{\text{new}}}(s,a) \geq Q_\tau^{\pi_{\text{old}}}(s,a)$ *for all* (s,a) *and* $\tau > 0$.

Previous Work (Haarnoja et al., 2018)

Lemma 1 (Policy Improvement under RKL Reduction, Restatement of Lemma 2 (Haarnoja, Zhou, et al., 2018)). *For* $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, *if for all* s

$$\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \parallel \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot)) \leq \mathrm{KL}(\pi_{\text{old}}(\cdot \mid s) \parallel \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot))$$

then $Q_\tau^{\pi_{\text{new}}}(s, a) \geq Q_\tau^{\pi_{\text{old}}}(s, a)$ *for all* (s, a) *and* $\tau > 0$.

Problem

Requires RKL reduction or maintenance in every state

An average RKL reduction is sufficient for policy improvement

Some definitions

$$\eta_\tau(\pi) := \mathbb{E}_\rho\left[V_\tau^\pi(s)\right]$$

$$A_\tau^\pi(s, a) := Q_\tau^\pi(s, a) - \tau \log \pi(a \mid s) - V_\tau^\pi(s)$$

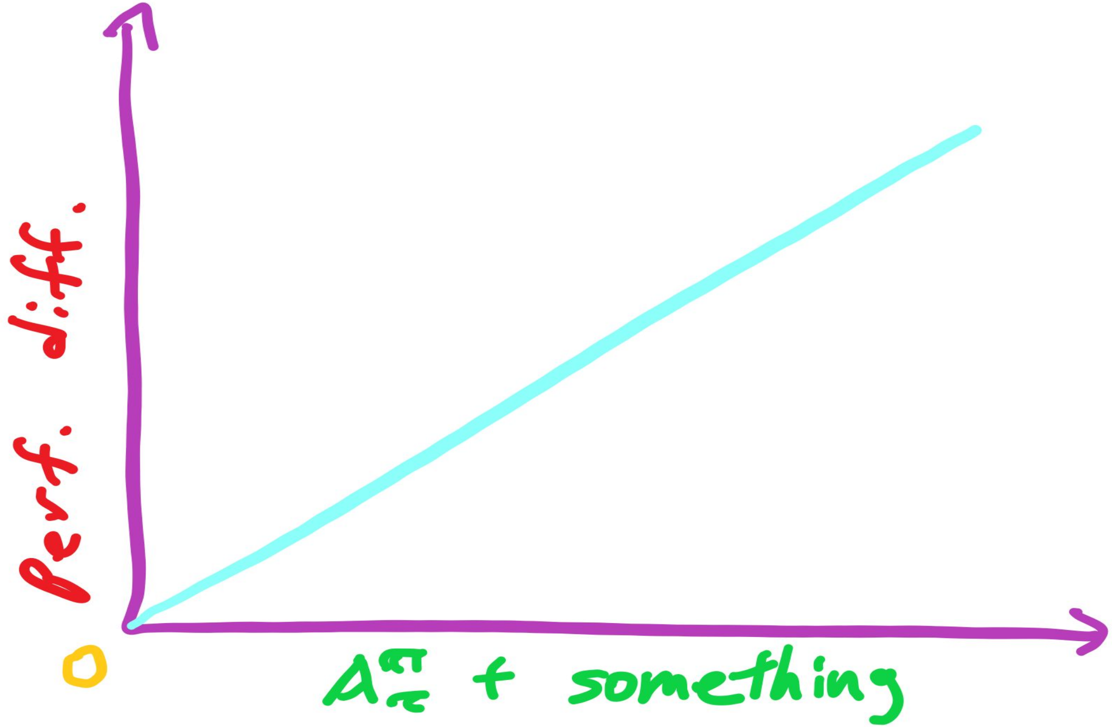
Goal

Compare two policies

$$\pi_{new}, \pi_{old}$$

Idea of the lemma

Idea of the lemma



Soft Performance Difference

Lemma 3 (Soft Performance Difference). *For any policies* $\pi_{\text{old}}, \pi_{\text{new}}$, *the following is true for any* $\tau \geq 0$.

$$\eta_\tau(\pi_{\text{new}}) - \eta_\tau(\pi_{\text{old}}) = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_{\text{new}}},\pi_{\text{new}}}[A_\tau^{\pi_{\text{old}}}(s,a)] + \frac{\tau}{1-\gamma}\mathbb{E}_{d^{\pi_{\text{new}}}}[\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \parallel \pi_{\text{old}}(\cdot \mid s))].$$

Soft Performance Difference

Lemma 3 (Soft Performance Difference). *For any policies $\pi_{\text{old}}, \pi_{\text{new}}$, the following is true for any $\tau \geq 0$.*

$$\eta_\tau(\pi_{\text{new}}) - \eta_\tau(\pi_{\text{old}}) = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_{\text{new}}},\pi_{\text{new}}}[A_\tau^{\pi_{\text{old}}}(s,a)] + \frac{\tau}{1-\gamma}\mathbb{E}_{d^{\pi_{\text{new}}}}[\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \,\|\, \pi_{\text{old}}(\cdot \mid s))].$$

Soft Performance Difference

Lemma 3 (Soft Performance Difference). *For any policies $\pi_{\text{old}}, \pi_{\text{new}}$, the following is true for any $\tau \geq 0$.*

$$\eta_\tau(\pi_{\text{new}}) - \eta_\tau(\pi_{\text{old}}) = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_{\text{new}}},\pi_{\text{new}}}[A_\tau^{\pi_{\text{old}}}(s,a)] + \frac{\tau}{1-\gamma}\mathbb{E}_{d^{\pi_{\text{new}}}}[\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \,\|\, \pi_{\text{old}}(\cdot \mid s))].$$

Soft Performance Difference

Lemma 3 (Soft Performance Difference). *For any policies $\pi_{\text{old}}, \pi_{\text{new}}$, the following is true for any $\tau \geq 0$.*

$$\eta_\tau(\pi_{\text{new}}) - \eta_\tau(\pi_{\text{old}}) = \frac{1}{1-\gamma}\mathbb{E}_{d^{\pi_{\text{new}}},\pi_{\text{new}}}[A_\tau^{\pi_{\text{old}}}(s,a)] + \frac{\tau}{1-\gamma}\mathbb{E}_{d^{\pi_{\text{new}}}}[\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \parallel \pi_{\text{old}}(\cdot \mid s))].$$

Proof idea

Just write it out

RKL Average Guarantee

Proposition 1 (Policy Improvement under Average RKL Reduction). *For* $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, *if*

$$\mathbb{E}_{d^{\pi_{\text{new}}}}\left[\mathrm{KL}(\pi_{\text{old}}(\cdot \mid s) \,\|\, \mathcal{B}Q_{\tau}^{\pi_{\text{old}}}(s, \cdot))\right] \geq \mathbb{E}_{d^{\pi_{\text{new}}}}\left[\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \,\|\, \mathcal{B}Q_{\tau}^{\pi_{\text{old}}}(s, \cdot))\right],$$

then $\eta_\tau(\pi_{\text{new}}) \geq \eta_\tau(\pi_{\text{old}})$.

RKL Average Guarantee

Proposition 1 (Policy Improvement under Average RKL Reduction). *For* $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, *if*

$$\mathbb{E}_{d^{\pi_{\text{new}}}}\left[\mathrm{KL}(\pi_{\text{old}}(\cdot \mid s) \,\|\, \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot))\right] \geq \mathbb{E}_{d^{\pi_{\text{new}}}}\left[\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \,\|\, \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot))\right],$$

then $\eta_\tau(\pi_{\text{new}}) \geq \eta_\tau(\pi_{\text{old}})$.

RKL Average Guarantee

Proposition 1 (Policy Improvement under Average RKL Reduction). *For* $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, *if*

$$\mathbb{E}_{d^{\pi_{\text{new}}}}[\mathrm{KL}(\pi_{\text{old}}(\cdot \mid s) \parallel \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot))] \geq \mathbb{E}_{d^{\pi_{\text{new}}}}[\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \parallel \mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot))],$$

then $\eta_\tau(\pi_{\text{new}}) \geq \eta_\tau(\pi_{\text{old}})$.

RKL Average Guarantee

Proposition 1 (Policy Improvement under Average RKL Reduction). *For* $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$, *if*

$$\mathbb{E}_{d^{\pi_{\text{new}}}}\left[\mathrm{KL}(\pi_{\text{old}}(\cdot \mid s) \parallel \mathcal{B}Q_{\tau}^{\pi_{\text{old}}}(s, \cdot))\right] \geq \mathbb{E}_{d^{\pi_{\text{new}}}}\left[\mathrm{KL}(\pi_{\text{new}}(\cdot \mid s) \parallel \mathcal{B}Q_{\tau}^{\pi_{\text{old}}}(s, \cdot))\right],$$

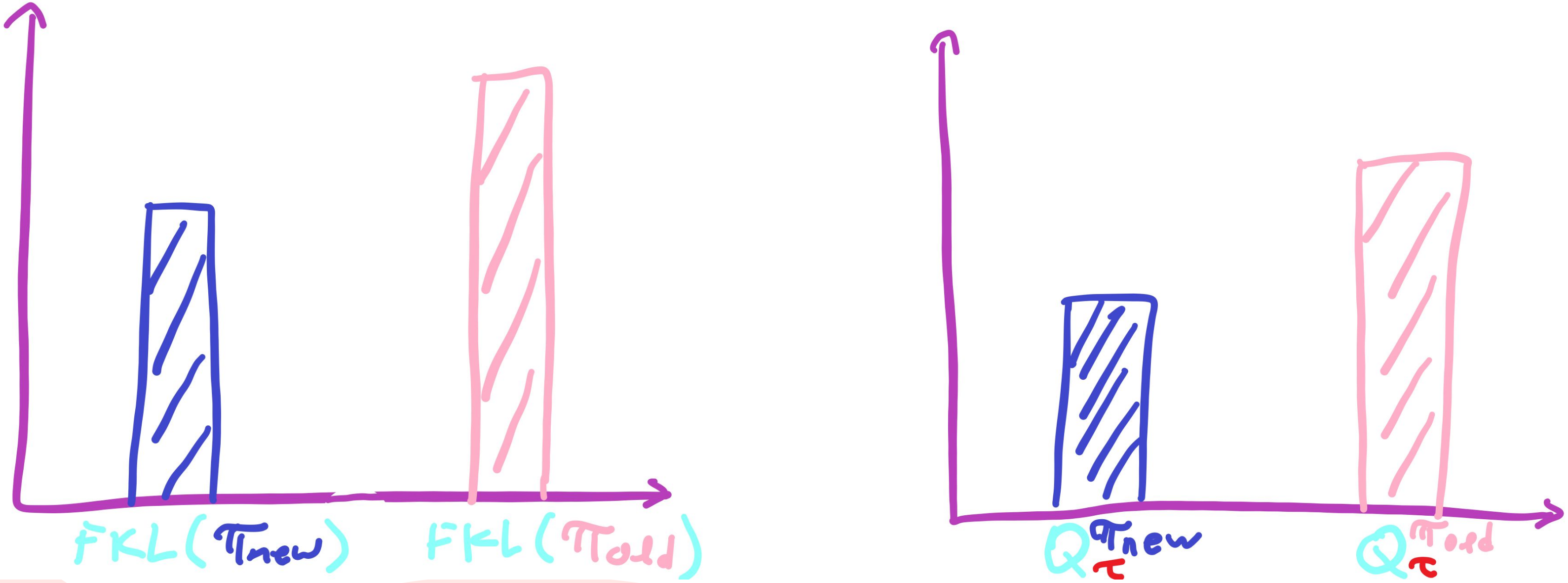
then $\eta_{\tau}(\pi_{\text{new}}) \geq \eta_{\tau}(\pi_{\text{old}})$.

Proof idea

Just write it out and use the soft performance difference lemma

The FKL can fail to induce improvement

What do we want?



FKL Counterexample

Proposition 2 (Counterexample for Policy Improvement with FKL). *There exists an MDP, a state s', an initial policy π_{old}, policy π_{new}, and temperature $\tau > 0$ such that for any $\gamma \in (0, 1]$*

$$\forall s \in S,\ \mathrm{KL}(\mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot) \parallel \pi_{\text{old}}(\cdot \mid s)) \geq \mathrm{KL}(\mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot) \parallel \pi_{\text{new}}(\cdot \mid s))$$

but $\forall a \in \mathcal{A},\ Q_\tau^{\pi_{\text{new}}}(s', a) < Q_\tau^{\pi_{\text{old}}}(s', a)$.

FKL Counterexample

Proposition 2 (Counterexample for Policy Improvement with FKL). *There exists an MDP, a state s', an initial policy π_{old}, policy π_{new}, and temperature $\tau > 0$ such that for any $\gamma \in (0, 1]$*

$$\forall s \in S, \mathrm{KL}(\mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot) \parallel \pi_{\text{old}}(\cdot \mid s)) \geq \mathrm{KL}(\mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot) \parallel \pi_{\text{new}}(\cdot \mid s))$$

but $\forall a \in \mathcal{A}, Q_\tau^{\pi_{\text{new}}}(s', a) < Q_\tau^{\pi_{\text{old}}}(s', a)$.

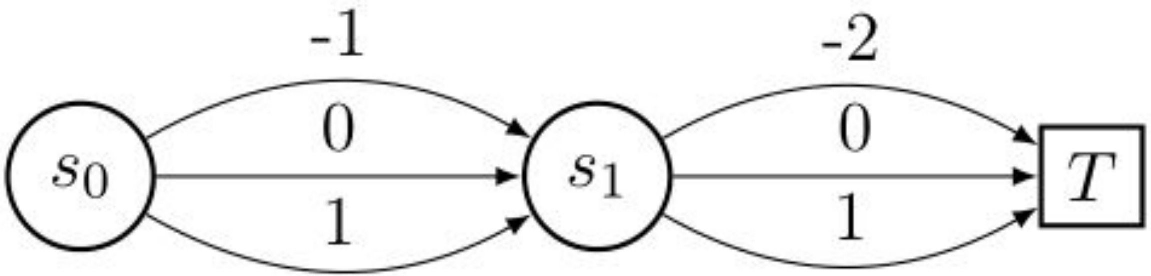
FKL Counterexample

Proposition 2 (Counterexample for Policy Improvement with FKL). *There exists an MDP, a state s', an initial policy π_{old}, policy π_{new}, and temperature $\tau > 0$ such that for any $\gamma \in (0, 1]$*

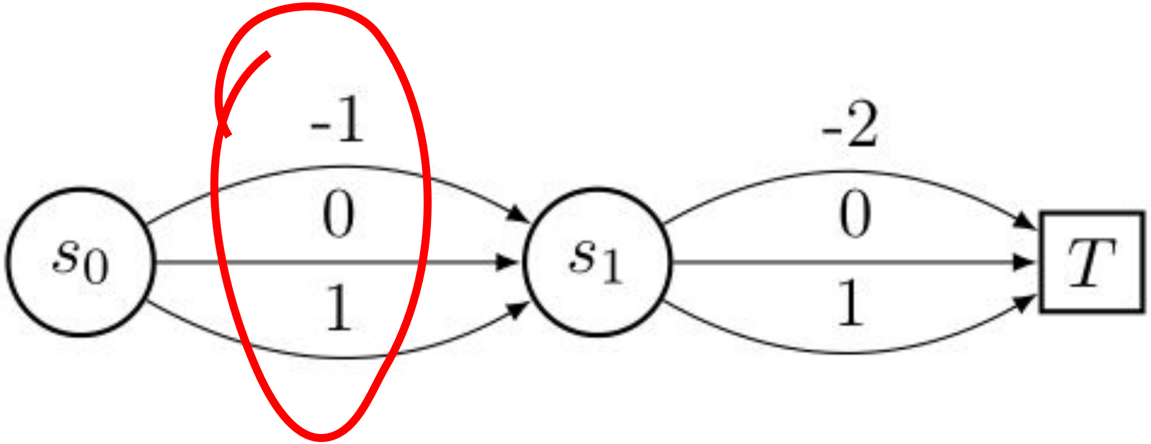
$$\forall s \in S, \mathrm{KL}(\mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot) \parallel \pi_{\text{old}}(\cdot \mid s)) \geq \mathrm{KL}(\mathcal{B}Q_\tau^{\pi_{\text{old}}}(s, \cdot) \parallel \pi_{\text{new}}(\cdot \mid s))$$

but $\forall a \in \mathcal{A}$, $Q_\tau^{\pi_{\text{new}}}(s', a) < Q_\tau^{\pi_{\text{old}}}(s', a)$.

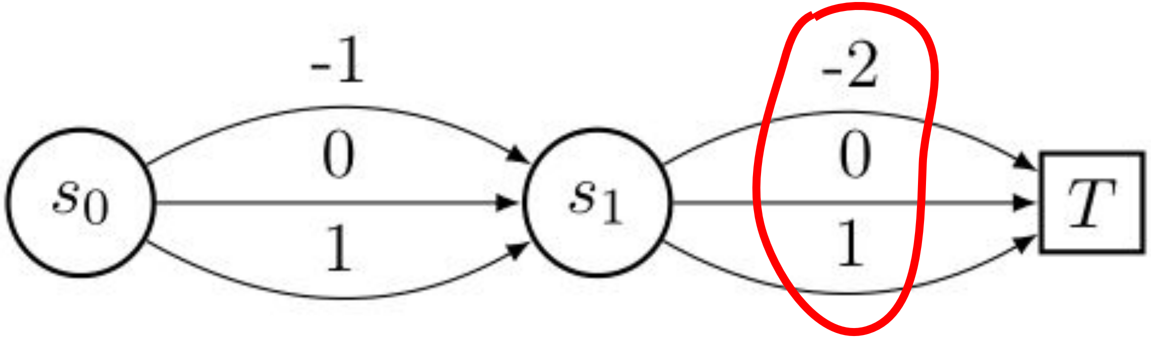
Proof idea



Proof idea



Proof idea



Can we do better?

What if we just reduced the FKL even more?

The FKL can induce policy improvement with a sufficiently large reduction

FKL Improvement

Proposition 3 (Policy Improvement for FKL with Sufficient Reduction). *Assume a discrete action space with $|\mathcal{A}| < \infty$, with a policy space Π that consists of policies where $\pi(a \mid s) > 0$ for all a. Let $C \geq 0$, $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$ be such that for a state s,*

$$\mathrm{KL}(\mathcal{B}Q^{\pi_{\text{old}}}(s,\cdot) \,\|\, \pi_{\text{new}}(\cdot \mid s)) + C \leq \mathrm{KL}(\mathcal{B}Q^{\pi_{\text{old}}}(s,\cdot) \,\|\, \pi_{\text{old}}(\cdot \mid s)), \tag{3.1}$$

where C additionally satisfies

$$C \geq \frac{1}{2}\sum_a \left(1 - \frac{1}{\pi_{\text{old}}(a \mid s)}\right)^2 \left(1 + \frac{Q^{\pi_{\text{old}}}(s,a)}{\tau} + \frac{\exp(\tau^{-1}Q^{\pi_{\text{old}}}(s,a))Q^{\pi_{\text{old}}}(s,a)^2}{2\tau^2}\right)$$
$$+ \frac{1}{2}\sum_a \exp(\tau^{-1}Q^{\pi_{\text{old}}}(s,a))\tau^{-2}Q^{\pi_{\text{old}}}(s,a)^2(1 - \pi_{\text{old}}(a \mid s))$$

with $\tau > 0$. Then,

$$\sum_a Q^{\pi_{\text{old}}}(s,a)\pi_{\text{old}}(a \mid s) \leq \sum_a Q^{\pi_{\text{old}}}(s,a)\pi_{\text{new}}(a \mid s).$$

FKL Improvement

Proposition 3 (Policy Improvement for FKL with Sufficient Reduction)**.** *Assume a discrete action space with $|\mathcal{A}| < \infty$, with a policy space Π that consists of policies where $\pi(a \mid s) > 0$ for all a. Let $C \geq 0$, $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$ be such that for a state s,*

$$\mathrm{KL}(\mathcal{B}Q^{\pi_{\text{old}}}(s,\cdot) \parallel \pi_{\text{new}}(\cdot \mid s)) + C \leq \mathrm{KL}(\mathcal{B}Q^{\pi_{\text{old}}}(s,\cdot) \parallel \pi_{\text{old}}(\cdot \mid s)), \tag{3.1}$$

where C additionally satisfies

$$C \geq \frac{1}{2}\sum_a \left(1 - \frac{1}{\pi_{\text{old}}(a \mid s)}\right)^2 \left(1 + \frac{Q^{\pi_{\text{old}}}(s,a)}{\tau} + \frac{\exp(\tau^{-1}Q^{\pi_{\text{old}}}(s,a))Q^{\pi_{\text{old}}}(s,a)^2}{2\tau^2}\right)$$
$$+\frac{1}{2}\sum_a \exp(\tau^{-1}Q^{\pi_{\text{old}}}(s,a))\tau^{-2}Q^{\pi_{\text{old}}}(s,a)^2(1-\pi_{\text{old}}(a \mid s))$$

with $\tau > 0$. Then,

$$\sum_a Q^{\pi_{\text{old}}}(s,a)\pi_{\text{old}}(a \mid s) \leq \sum_a Q^{\pi_{\text{old}}}(s,a)\pi_{\text{new}}(a \mid s).$$

FKL Improvement

Proposition 3 (Policy Improvement for FKL with Sufficient Reduction). *Assume a discrete action space with $|\mathcal{A}| < \infty$, with a policy space Π that consists of policies where $\pi(a \mid s) > 0$ for all a. Let $C \geq 0$, $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$ be such that for a state s,*

$$\mathrm{KL}(\mathcal{B}Q^{\pi_{\text{old}}}(s,\cdot) \parallel \pi_{\text{new}}(\cdot \mid s)) + C \leq \mathrm{KL}(\mathcal{B}Q^{\pi_{\text{old}}}(s,\cdot) \parallel \pi_{\text{old}}(\cdot \mid s)), \tag{3.1}$$

where C additionally satisfies

$$C \geq \frac{1}{2}\sum_a \left(1 - \frac{1}{\pi_{\text{old}}(a \mid s)}\right)^2 \left(1 + \frac{Q^{\pi_{\text{old}}}(s,a)}{\tau} + \frac{\exp(\tau^{-1}Q^{\pi_{\text{old}}}(s,a))Q^{\pi_{\text{old}}}(s,a)^2}{2\tau^2}\right)$$
$$+ \frac{1}{2}\sum_a \exp(\tau^{-1}Q^{\pi_{\text{old}}}(s,a))\tau^{-2}Q^{\pi_{\text{old}}}(s,a)^2(1 - \pi_{\text{old}}(a \mid s))$$

with $\tau > 0$. Then,

$$\sum_a Q^{\pi_{\text{old}}}(s,a)\pi_{\text{old}}(a \mid s) \leq \sum_a Q^{\pi_{\text{old}}}(s,a)\pi_{\text{new}}(a \mid s).$$

FKL Improvement

Proposition 3 (Policy Improvement for FKL with Sufficient Reduction). *Assume a discrete action space with $|\mathcal{A}| < \infty$, with a policy space Π that consists of policies where $\pi(a \mid s) > 0$ for all a. Let $C \geq 0$, $\pi_{\text{old}}, \pi_{\text{new}} \in \Pi$ be such that for a state s,*

$$\mathrm{KL}(\mathcal{B}Q^{\pi_{\text{old}}}(s,\cdot) \parallel \pi_{\text{new}}(\cdot \mid s)) + C \leq \mathrm{KL}(\mathcal{B}Q^{\pi_{\text{old}}}(s,\cdot) \parallel \pi_{\text{old}}(\cdot \mid s)), \tag{3.1}$$

where C additionally satisfies

$$C \geq \frac{1}{2}\sum_a \left(1 - \frac{1}{\pi_{\text{old}}(a \mid s)}\right)^2 \left(1 + \frac{Q^{\pi_{\text{old}}}(s,a)}{\tau} + \frac{\exp(\tau^{-1}Q^{\pi_{\text{old}}}(s,a))Q^{\pi_{\text{old}}}(s,a)^2}{2\tau^2}\right)$$
$$+ \frac{1}{2}\sum_a \exp(\tau^{-1}Q^{\pi_{\text{old}}}(s,a))\tau^{-2}Q^{\pi_{\text{old}}}(s,a)^2(1 - \pi_{\text{old}}(a \mid s))$$

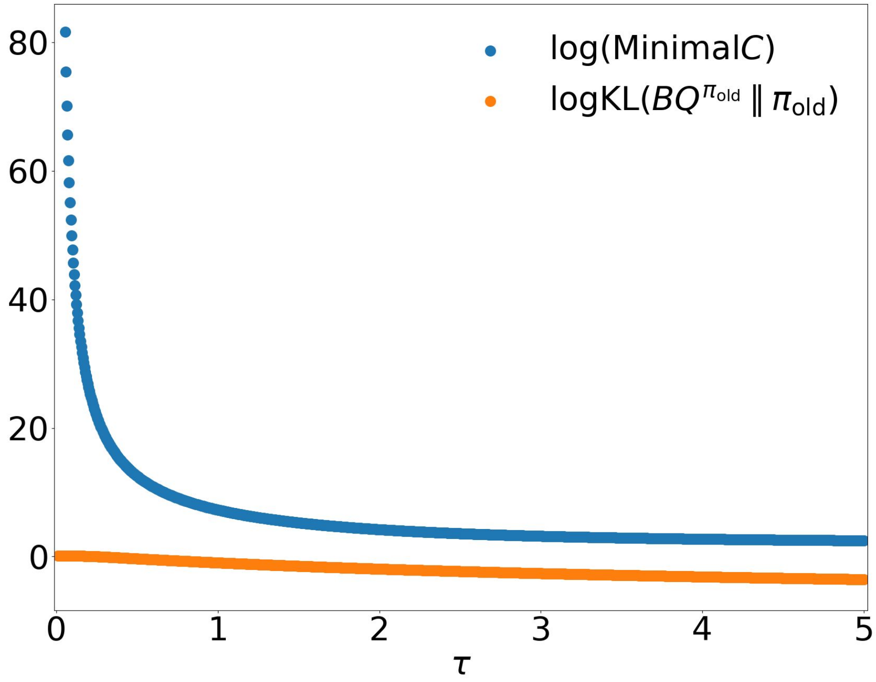
with $\tau > 0$. Then,

$$\sum_a Q^{\pi_{\text{old}}}(s,a)\pi_{\text{old}}(a \mid s) \leq \sum_a Q^{\pi_{\text{old}}}(s,a)\pi_{\text{new}}(a \mid s).$$

Proof idea

Apply **Taylor expansions** to extract the necessary terms for the conclusion

The condition can be quite strong!



FKL Improvement

FKL Improvement

Corollary 1. *Assume the following are true.*

$$\mathbb{E}_{d^{\pi_{new}},\pi_{old}}[Q_\tau^{\pi_{old}}(s,a)] \leq \mathbb{E}_{d^{\pi_{new}},\pi_{new}}[Q_\tau^{\pi_{old}}(s,a)],$$
$$\tau\mathbb{E}_{d^{\pi_{new}}}[\mathcal{H}(\pi_{new}(\cdot \mid s))] \geq \tau\mathbb{E}_{d^{\pi_{new}}}[\mathcal{H}(\pi_{old}(\cdot \mid s))].$$

Then $\eta_\tau(\pi_{new}) \geq \eta_\tau(\pi_{old})$. *This conclusion also holds for* $\tau = 0$.

FKL Improvement

Corollary 1. *Assume the following are true.*

$$\mathbb{E}_{d^{\pi_{new}},\pi_{old}}[Q_\tau^{\pi_{old}}(s,a)] \le \mathbb{E}_{d^{\pi_{new}},\pi_{new}}[Q_\tau^{\pi_{old}}(s,a)],$$

$$\tau\mathbb{E}_{d^{\pi_{new}}}[\mathcal{H}(\pi_{new}(\cdot \mid s))] \ge \tau\mathbb{E}_{d^{\pi_{new}}}[\mathcal{H}(\pi_{old}(\cdot \mid s))].$$

Then $\eta_\tau(\pi_{new}) \ge \eta_\tau(\pi_{old})$. *This conclusion also holds for* $\tau = 0$.

FKL Improvement

Corollary 1. *Assume the following are true.*

$$\mathbb{E}_{d^{\pi_{new}}, \pi_{old}}[Q_\tau^{\pi_{old}}(s,a)] \leq \mathbb{E}_{d^{\pi_{new}}, \pi_{new}}[Q_\tau^{\pi_{old}}(s,a)],$$
$$\tau \mathbb{E}_{d^{\pi_{new}}}[\mathcal{H}(\pi_{new}(\cdot \mid s))] \geq \tau \mathbb{E}_{d^{\pi_{new}}}[\mathcal{H}(\pi_{old}(\cdot \mid s))].$$

Then $\eta_\tau(\pi_{new}) \geq \eta_\tau(\pi_{old})$. *This conclusion also holds for* $\tau = 0$.

FKL Improvement

Corollary 1. *Assume the following are true.*

$$\mathbb{E}_{d^{\pi_{new}},\pi_{old}}[Q_\tau^{\pi_{old}}(s,a)] \le \mathbb{E}_{d^{\pi_{new}},\pi_{new}}[Q_\tau^{\pi_{old}}(s,a)],$$

$$\tau\mathbb{E}_{d^{\pi_{new}}}[\mathcal{H}(\pi_{new}(\cdot \mid s))] \ge \tau\mathbb{E}_{d^{\pi_{new}}}[\mathcal{H}(\pi_{old}(\cdot \mid s))].$$

Then $\eta_\tau(\pi_{new}) \ge \eta_\tau(\pi_{old})$. *This conclusion also holds for* $\tau = 0$.

Proof idea

Just writing it out

Takeaways

1. The RKL has a stronger policy improvement result than the FKL

2. The FKL can fail to induce policy improvement

3. FKL policy improvement can follow with some strong assumptions.

Limitations

1. Assumed exact critic

2. Strong FKL conditions

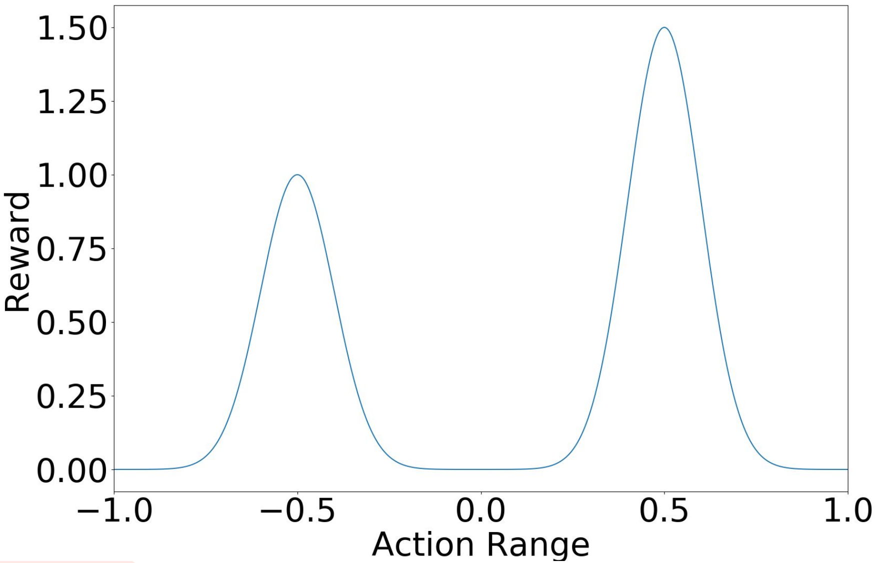
2 minute break

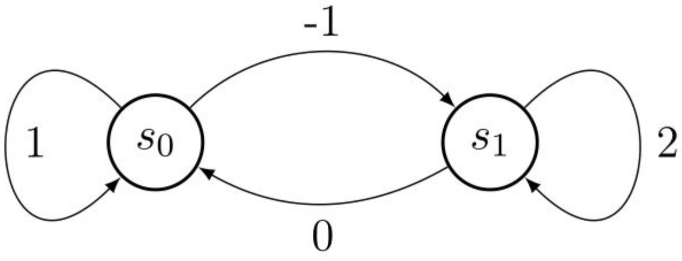
Small Experiments

Goals

Understand any policy improvement differences in simple environments

(Continuous-action) Environments





Implementation

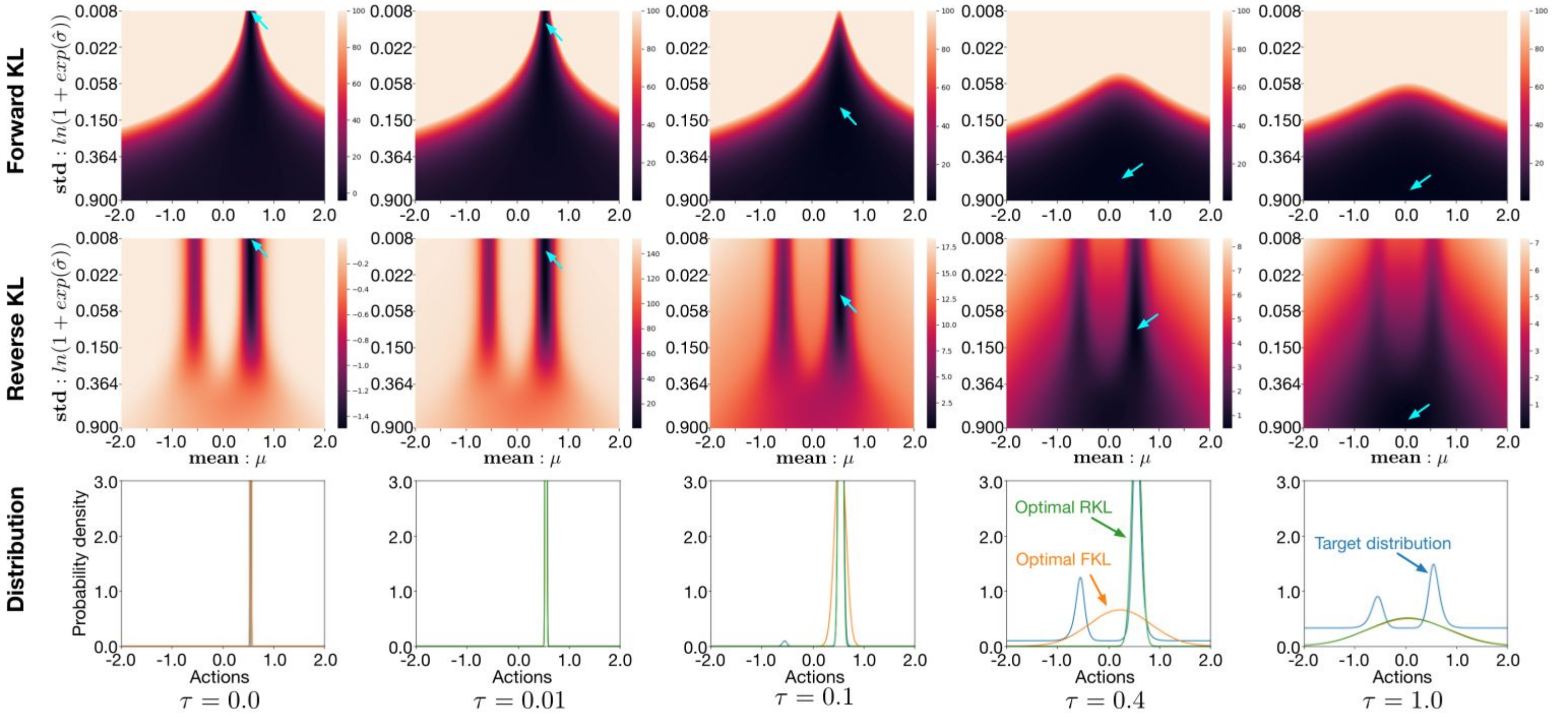
1. Tanh-Gaussian policy
2. Numerical integration

FKL has a smoother loss landscape on the Bimodal Bandit

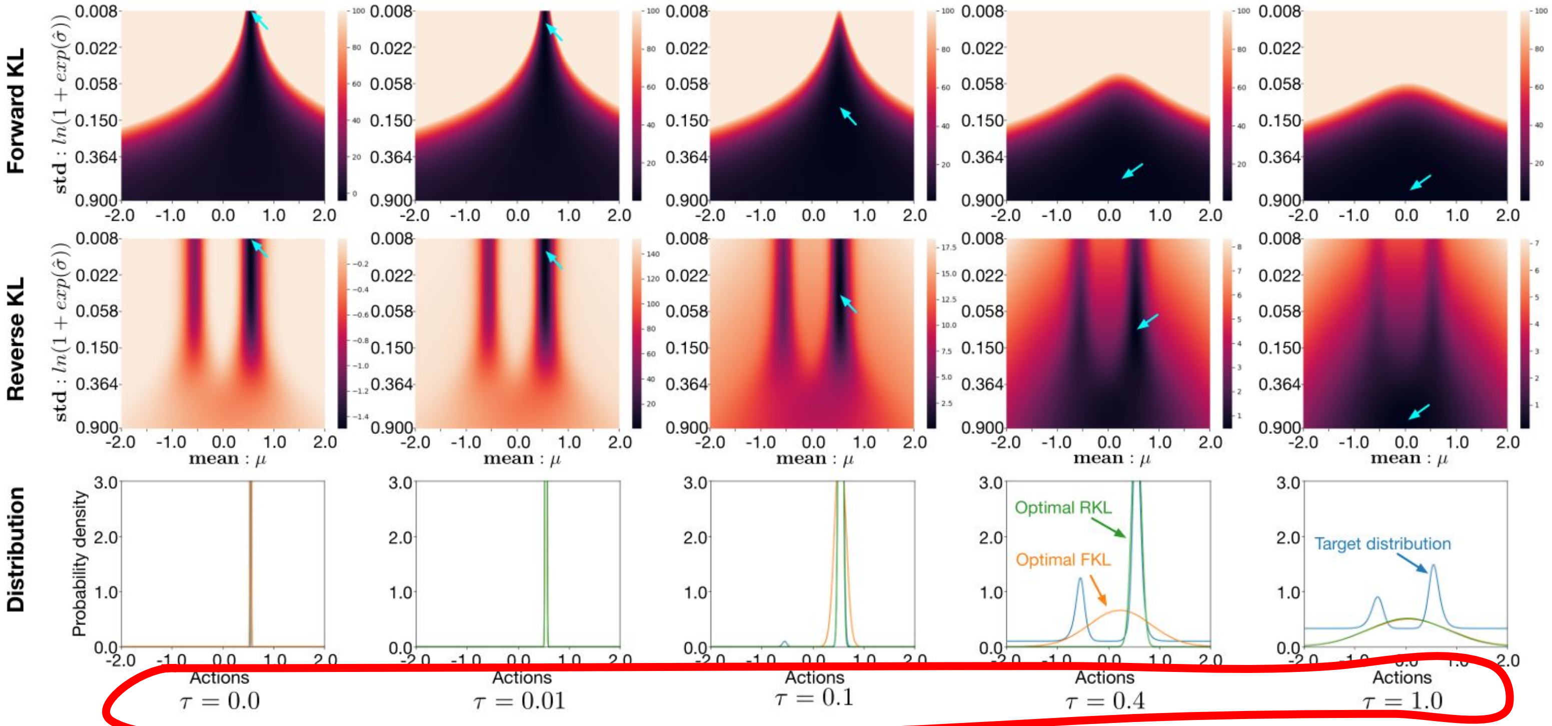
Bimodal Bandit KL Loss Heatmap

1. Each KL objective is a function of the policy parameters
2. Plot the value of the KL objective as we vary the policy parameters

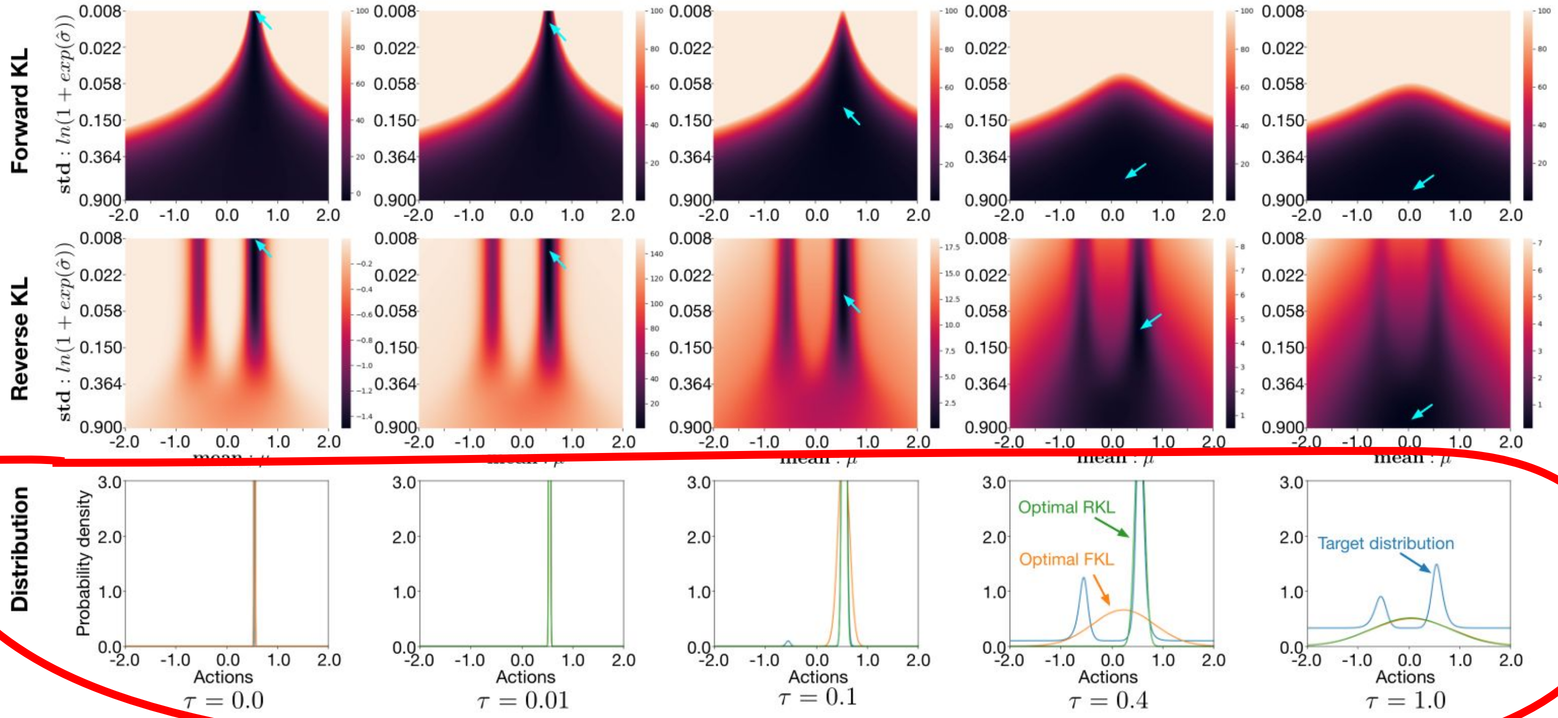
Bimodal Bandit KL Loss Heatmap



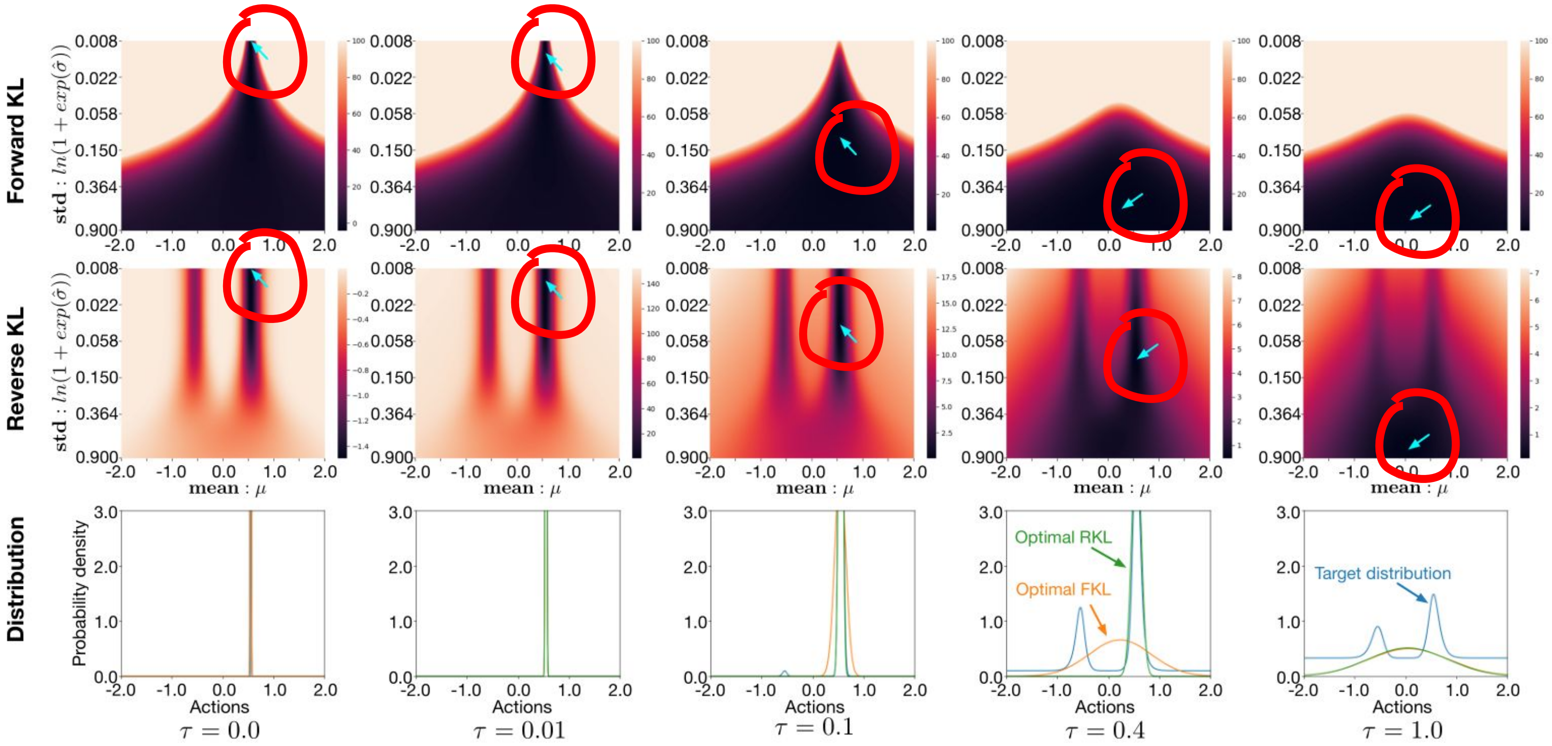
Bimodal Bandit KL Loss Heatmap



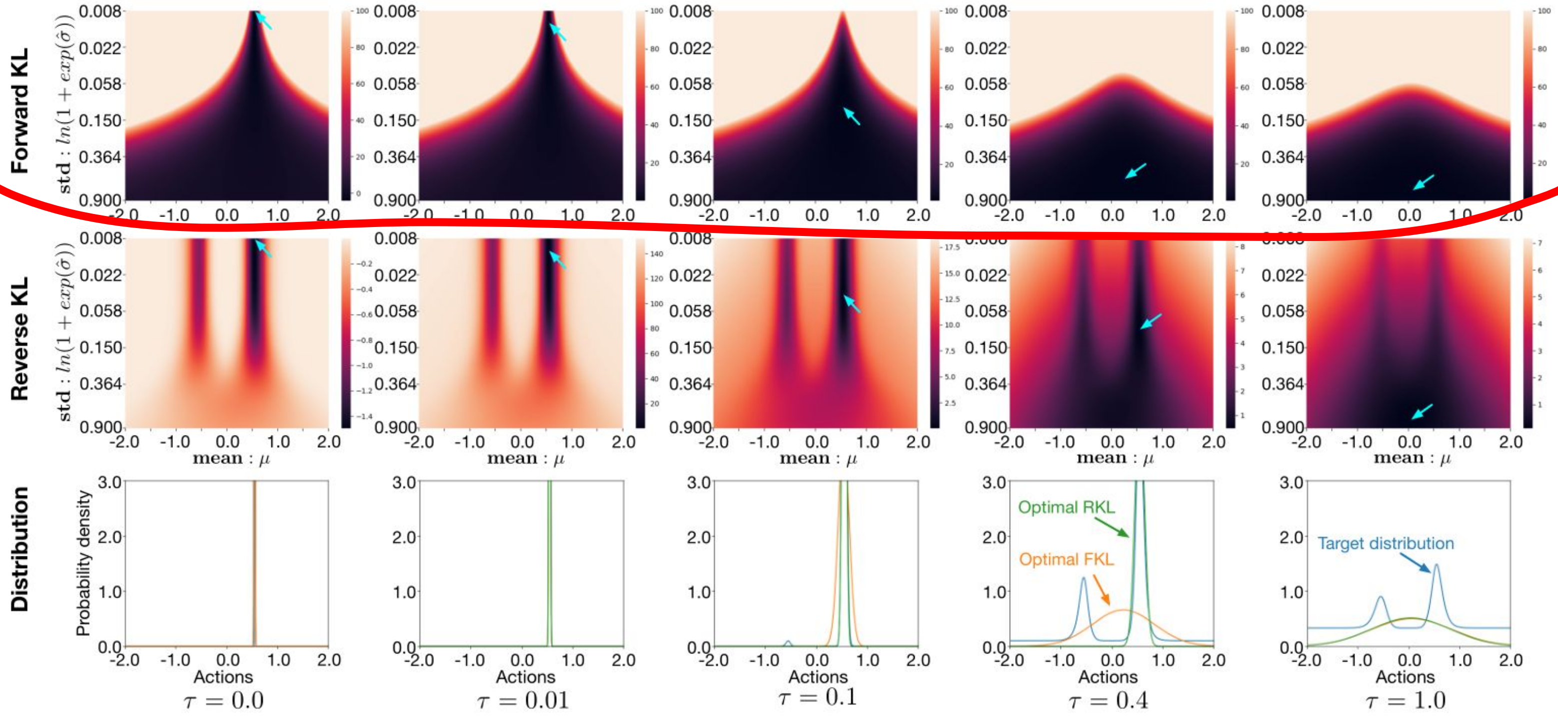
Bimodal Bandit KL Loss Heatmap



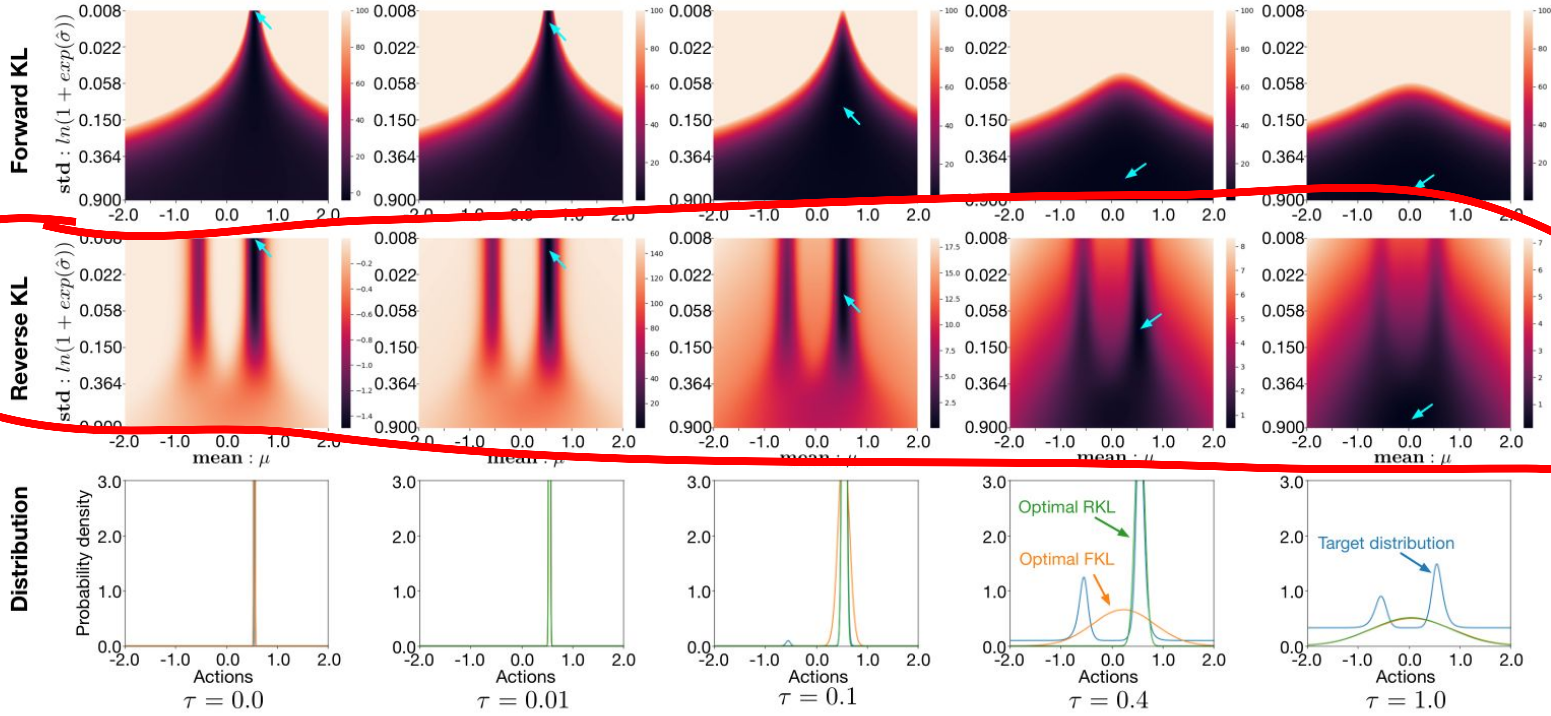
Bimodal Bandit KL Loss Heatmap



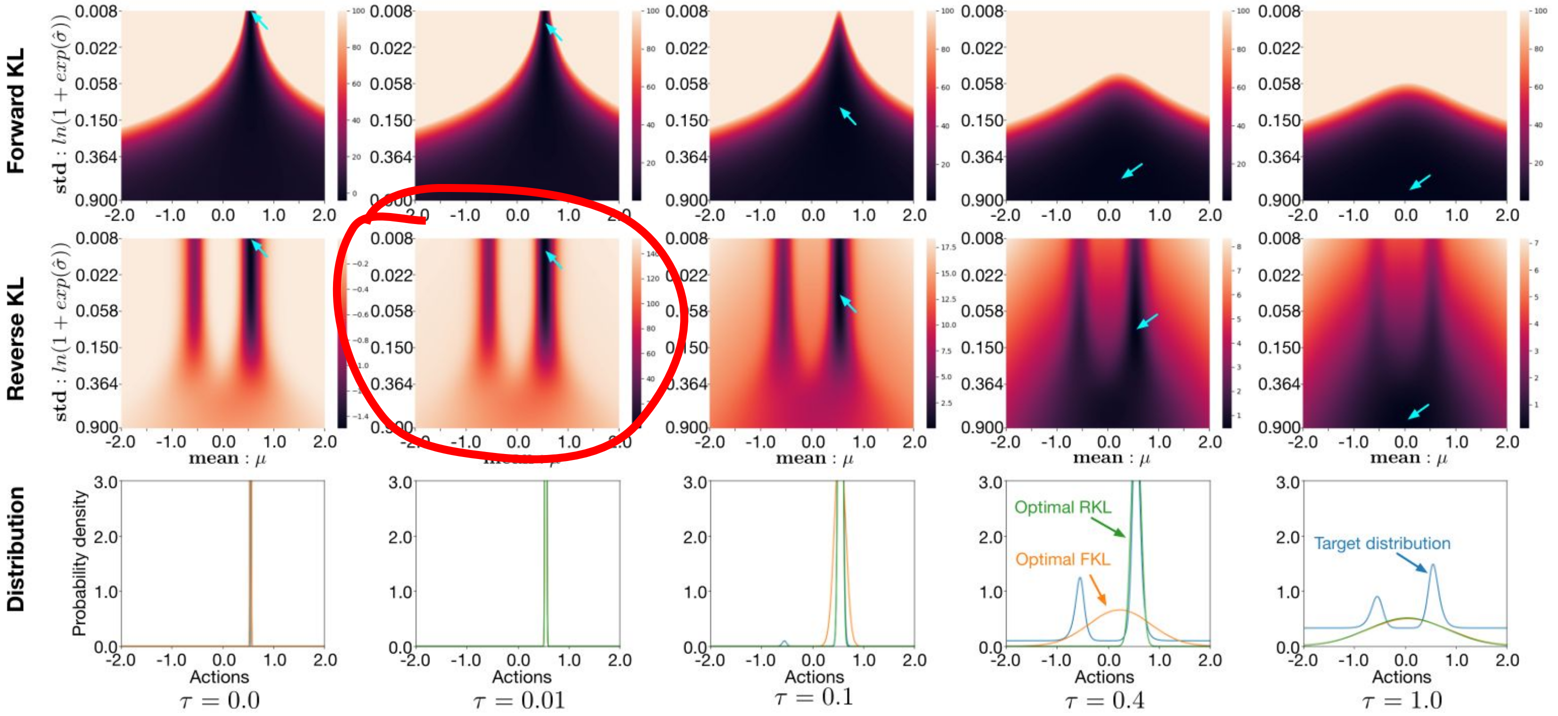
Bimodal Bandit KL Loss Heatmap



Bimodal Bandit KL Loss Heatmap



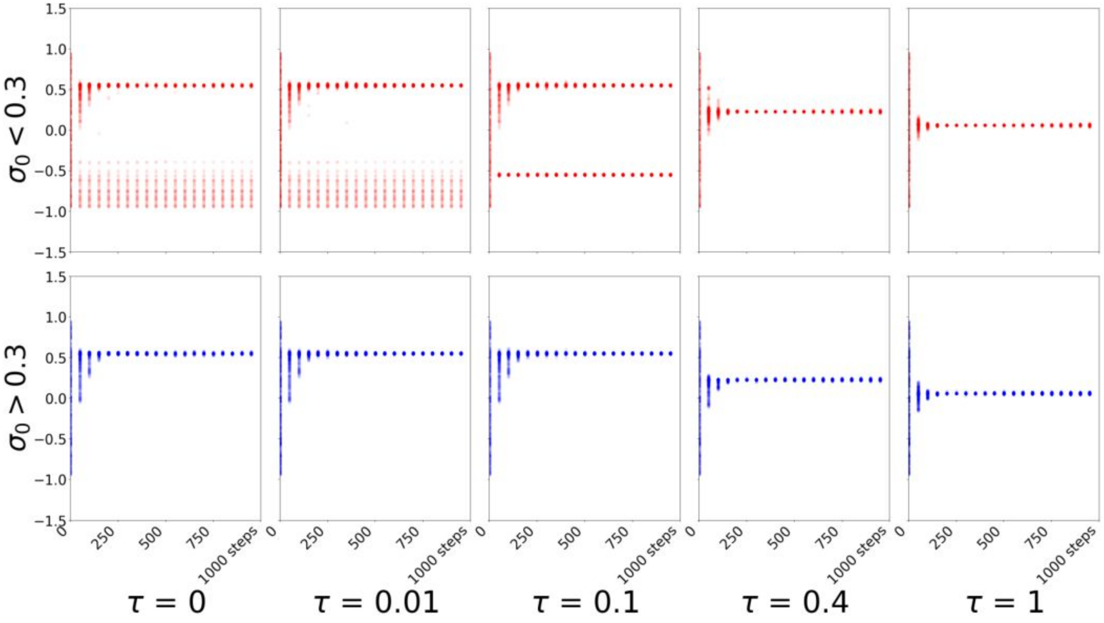
Bimodal Bandit KL Loss Heatmap



Tracking Bandit Iterates over Time

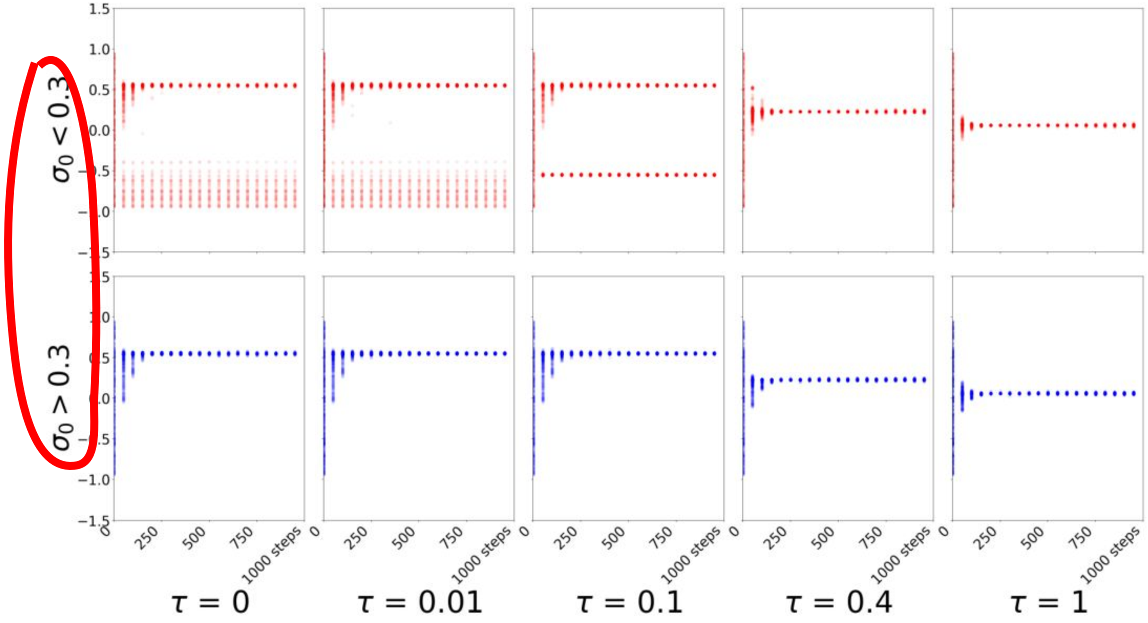
1. Randomly initialize policy parameters
2. Calculate target distribution with reward
3. Take gradient steps on the KL to update the parameters
4. Repeat (2) - (3) for N steps
5. Repeat (1) - (4) for 1000 initializations

Forward KL



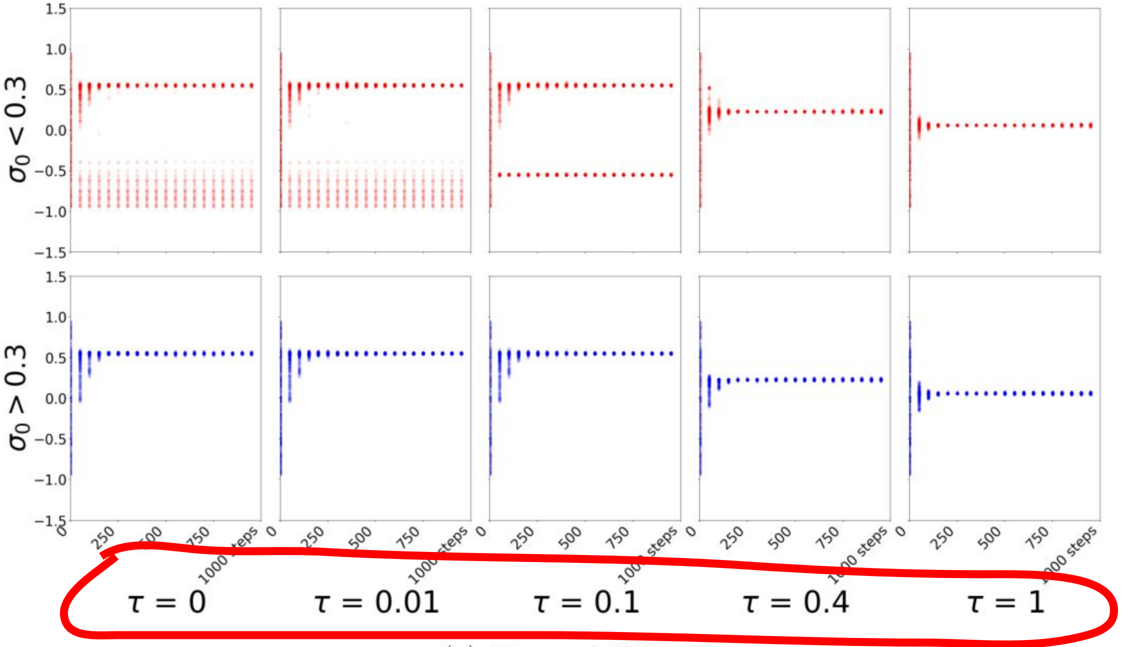
(a) Forward KL.

Forward KL



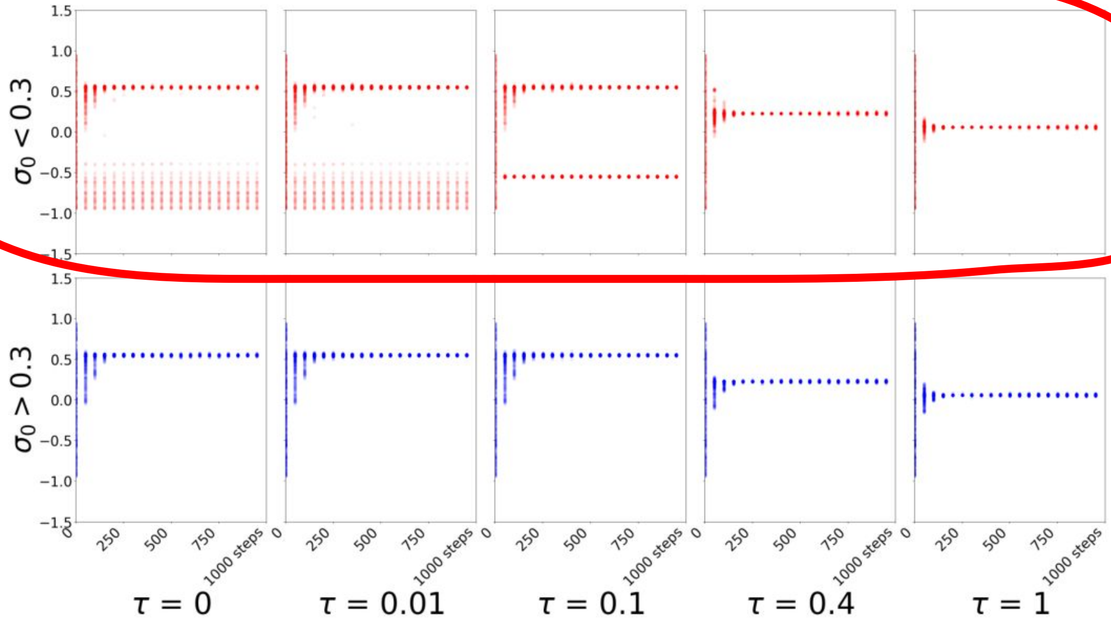
(a) Forward KL.

Forward KL



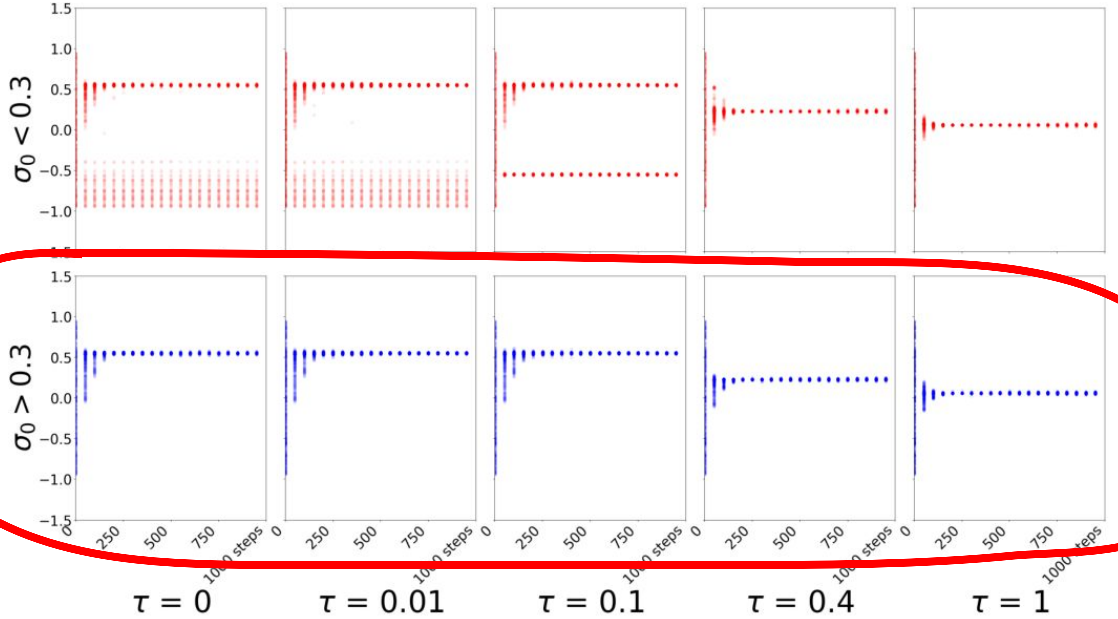
(a) Forward KL.

Forward KL



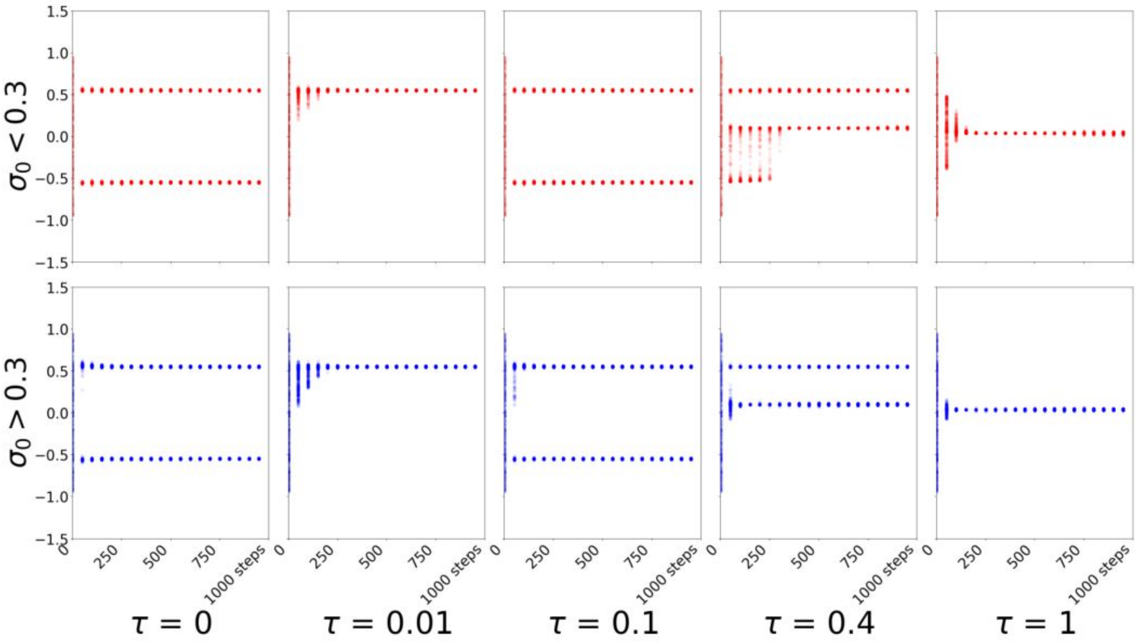
(a) Forward KL.

Forward KL



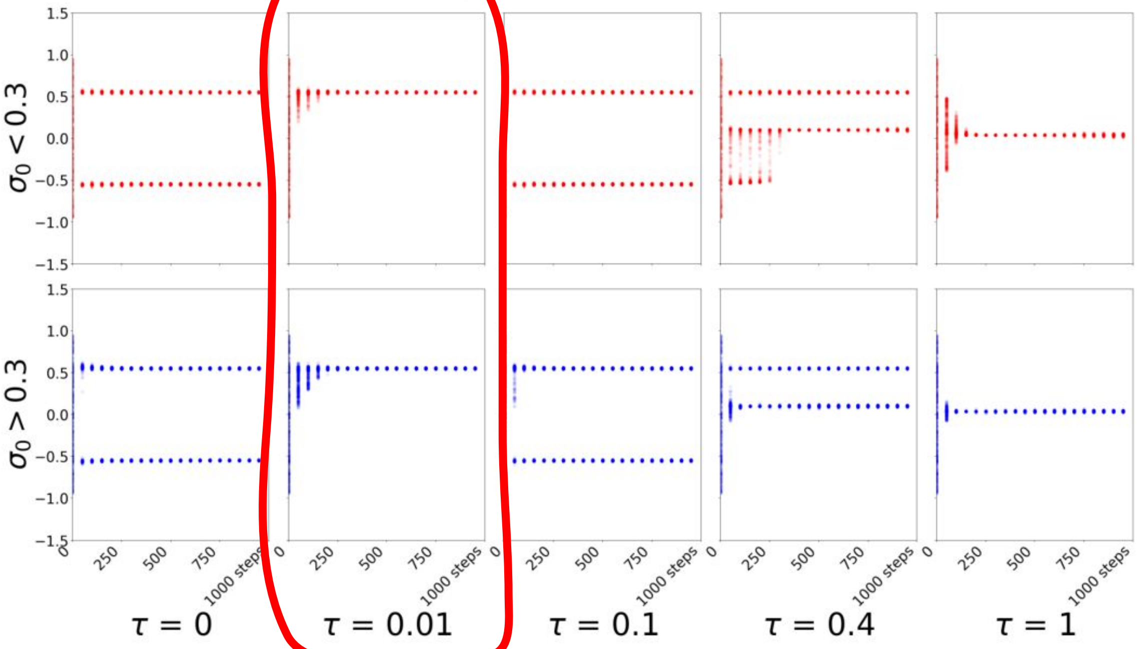
(a) Forward KL.

Reverse KL



(b) Reverse KL.

Reverse KL



(b) Reverse KL.

The FKL solution is more suboptimal on Switch-Stay

Experimental question

After optimizing a policy under either KL for some time, what is the quality of the resulting policy?

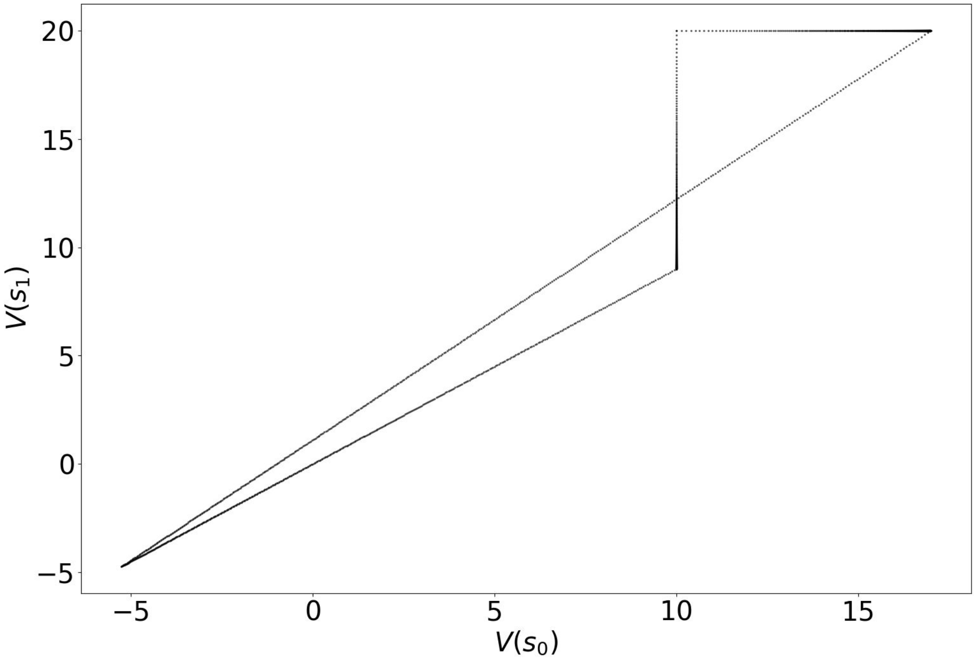
Experiment description

1. Randomly initialize policy parameters
2. Calculate value function of policy
3. Take gradient step with respect to mean KL divergence
4. Repeat (2) - (3) for 1000 steps
5. Plot value function of the final policy
6. Repeat for (1) - (5) for 1000 initializations

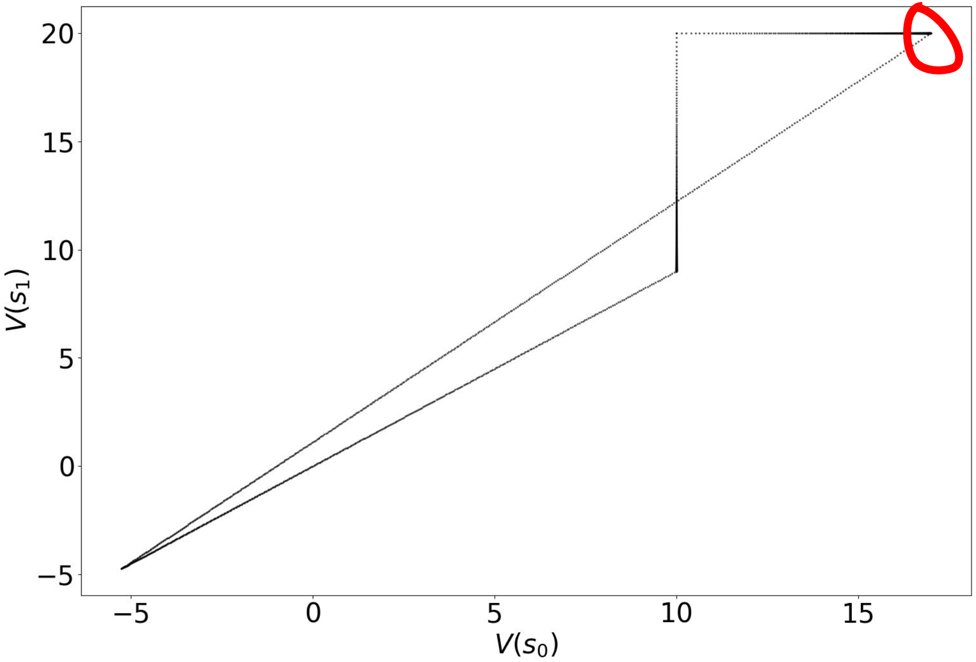
Value function space

The value function polytope is the space of all possible value functions for an MDP

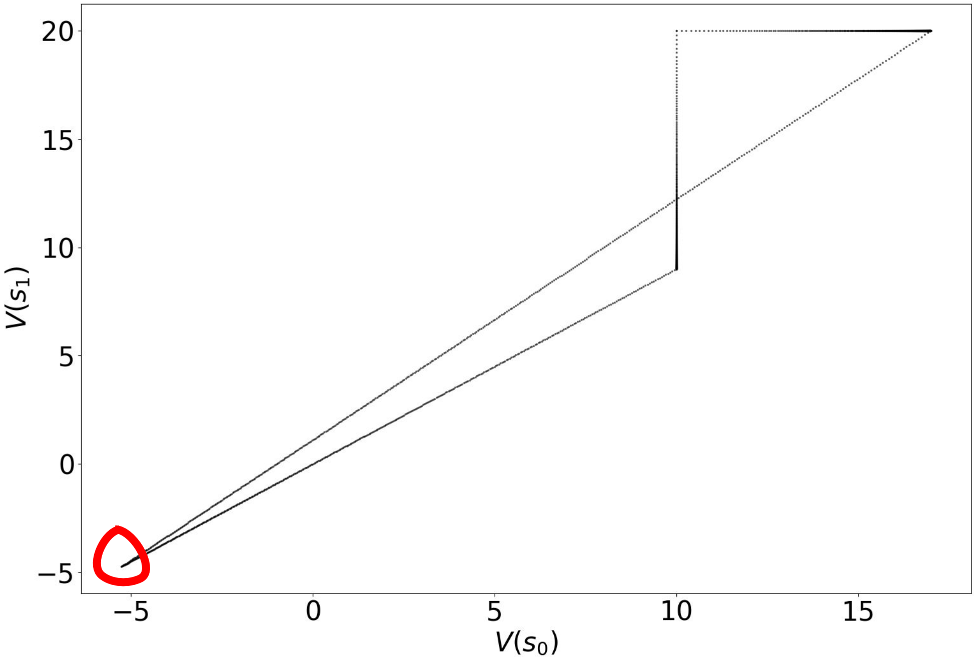
Switch–Stay Polytope Boundary



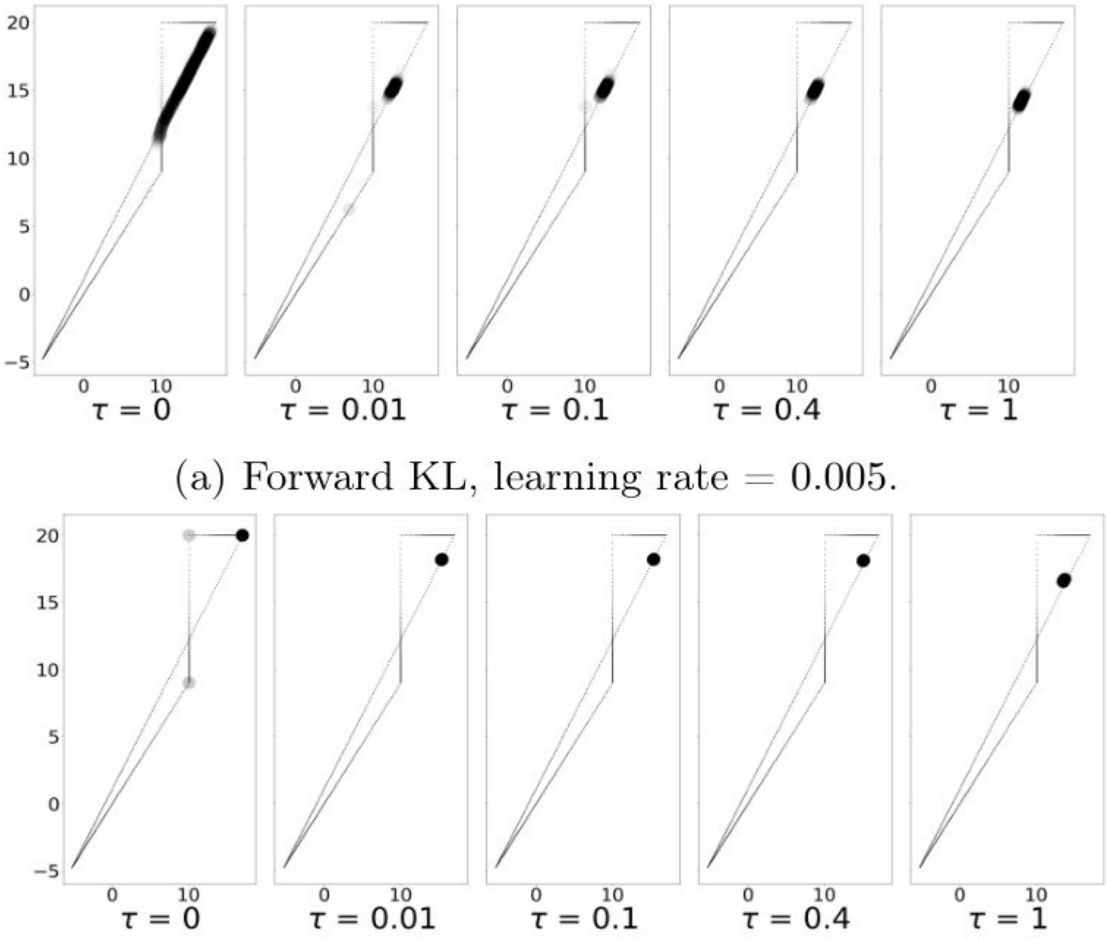
Switch–Stay Polytope Boundary



Switch-Stay Polytope Boundary



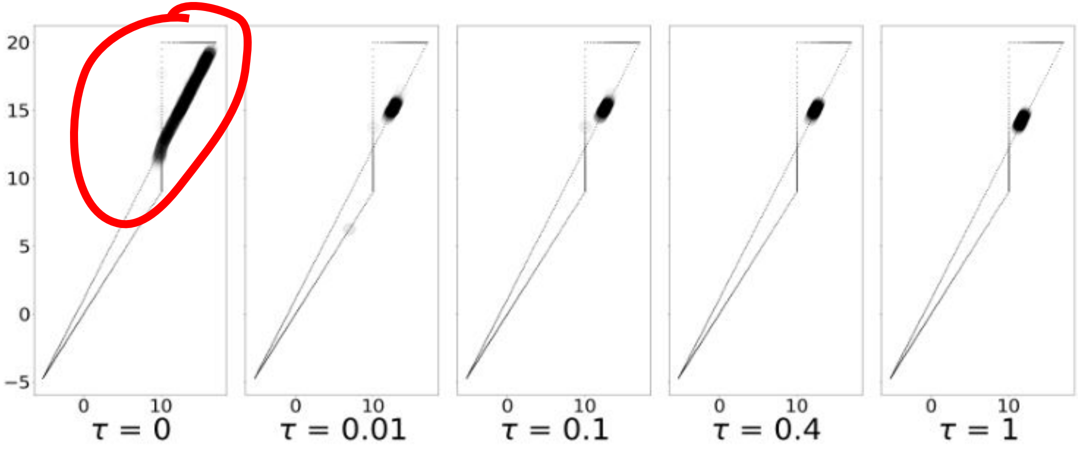
Switch-Stay



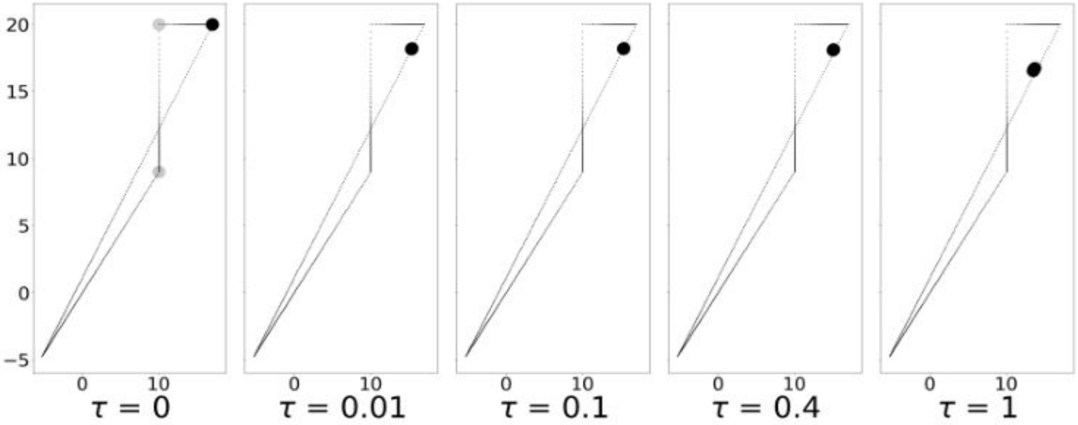
(a) Forward KL, learning rate = 0.005.

(b) Reverse KL, learning rate = 0.005.

Switch-Stay

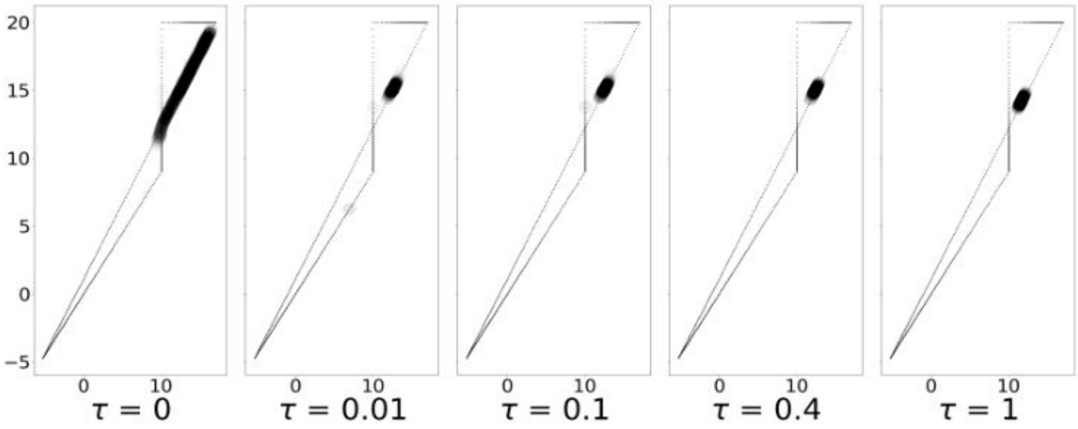


(a) Forward KL, learning rate = 0.005.

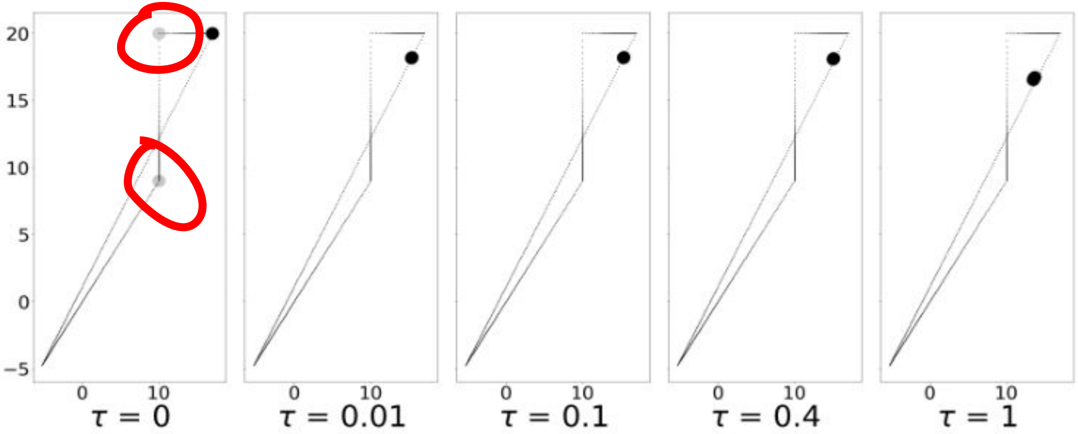


(b) Reverse KL, learning rate = 0.005.

Switch-Stay

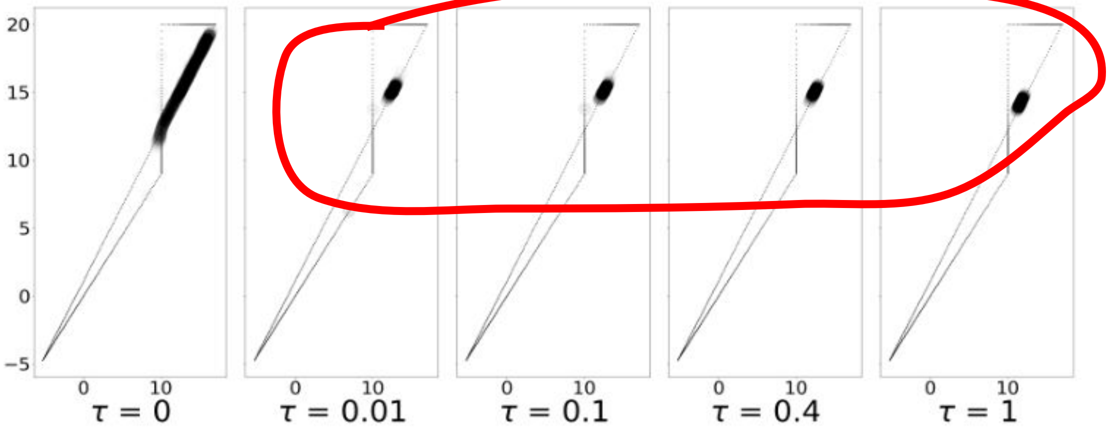


(a) Forward KL, learning rate = 0.005.

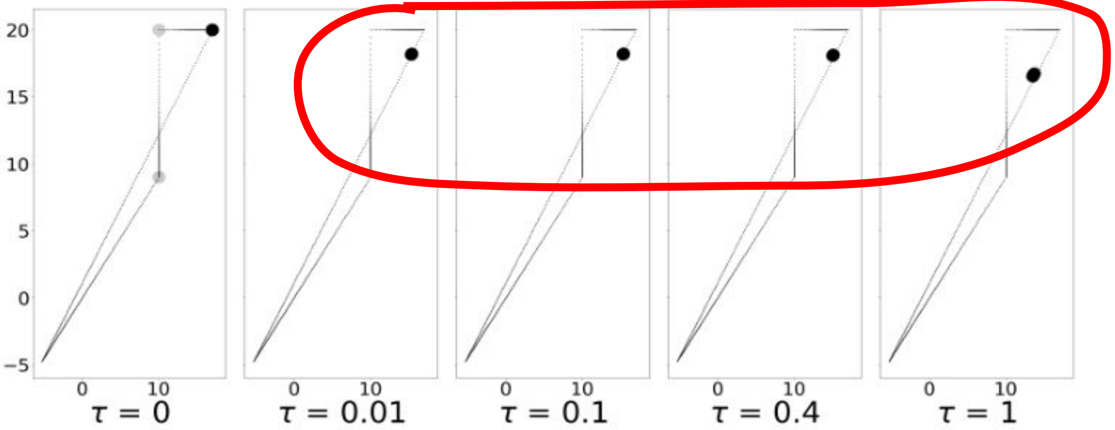


(b) Reverse KL, learning rate = 0.005.

Switch-Stay

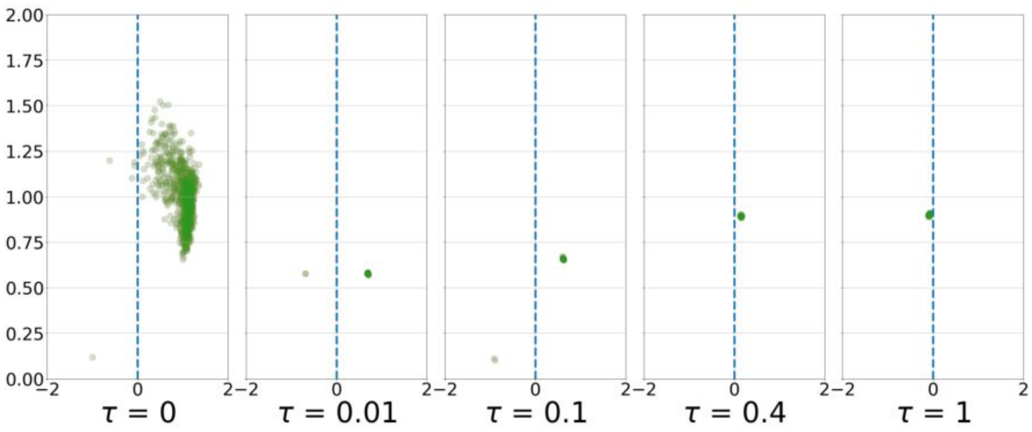


(a) Forward KL, learning rate = 0.005.

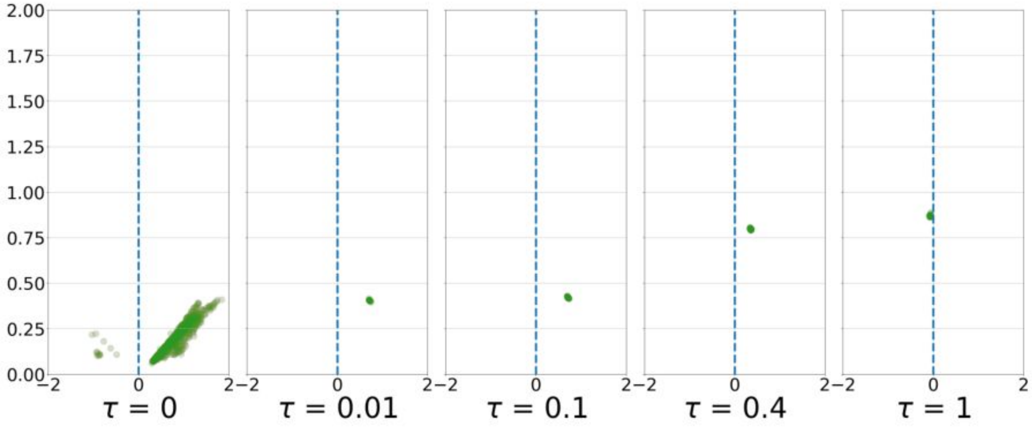


(b) Reverse KL, learning rate = 0.005.

Switch-Stay

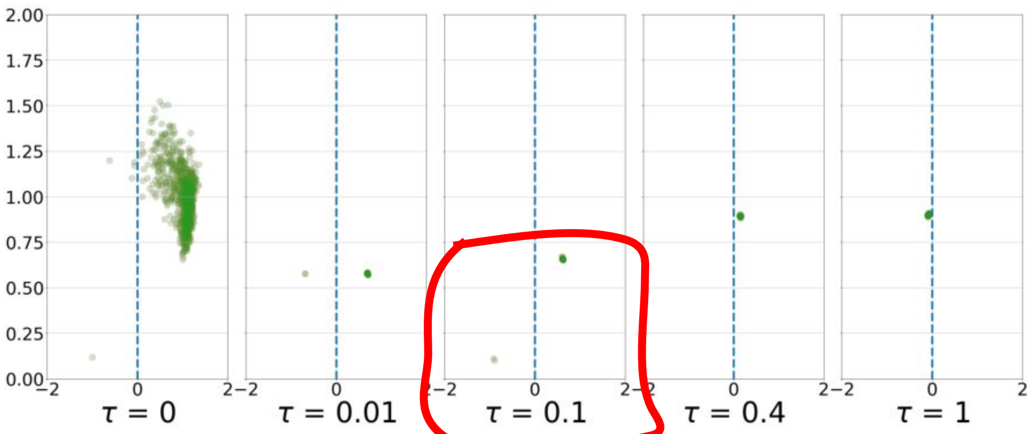


(a) Forward KL on state 0.

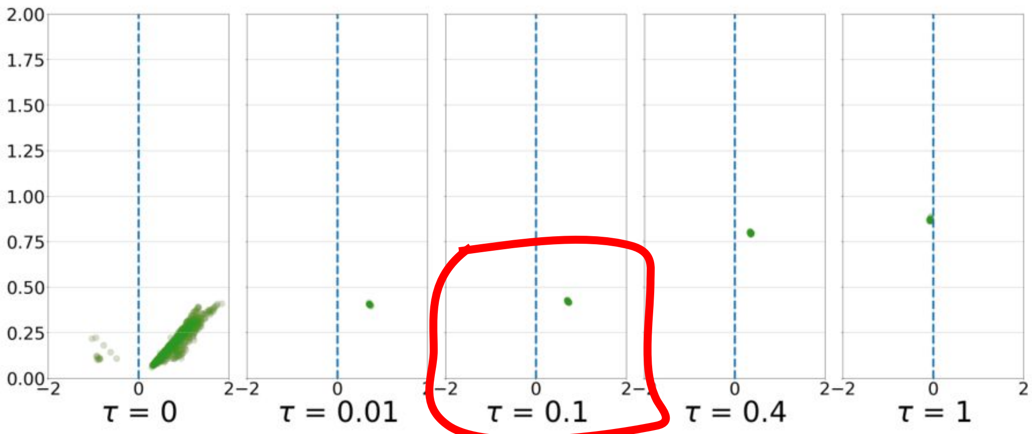


(b) Reverse KL on state 0.

Switch-Stay

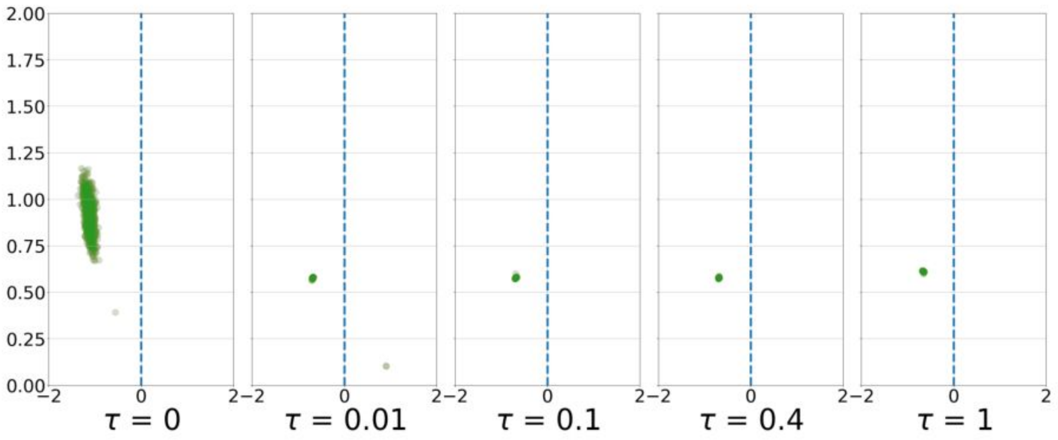


(a) Forward KL on state 0.

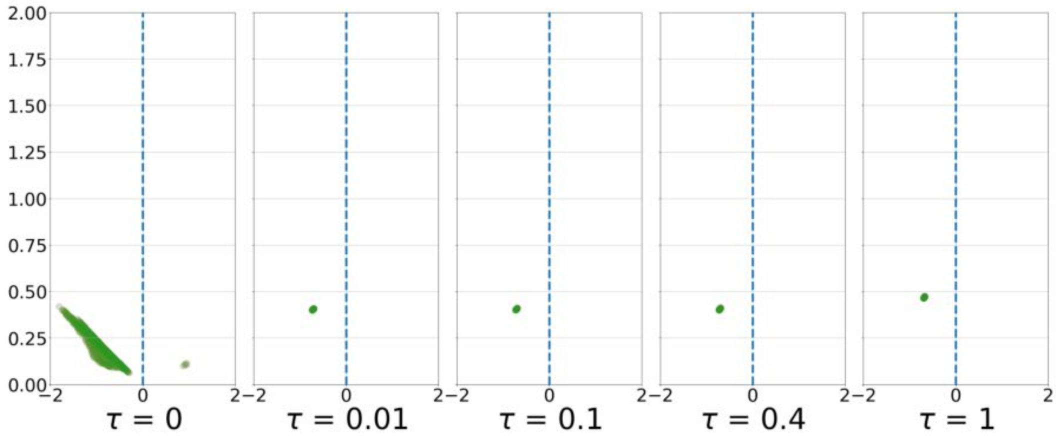


(b) Reverse KL on state 0.

Switch-Stay



(a) Forward KL on state 1.



(b) Reverse KL on state 1.

The FKL may be more robust to stochasticity

Implementation

- Repeat Switch-Stay experiments w/ Monte Carlo sampling

Switch–Stay, 10 points

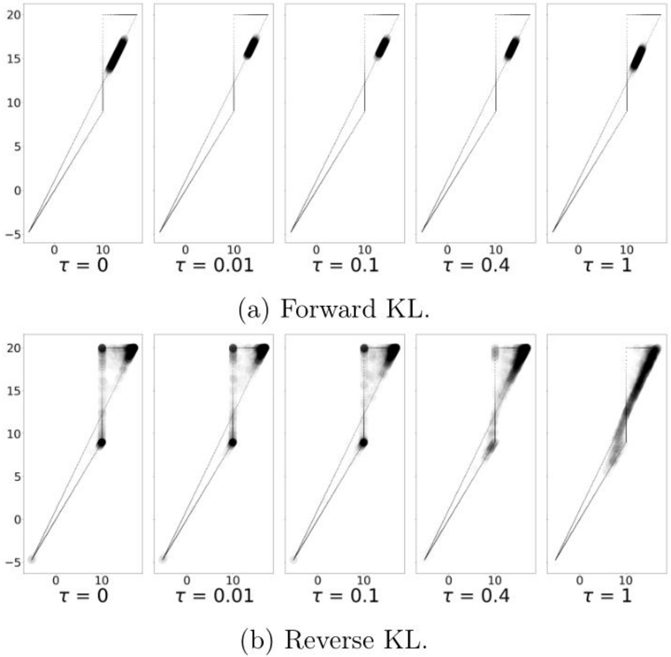
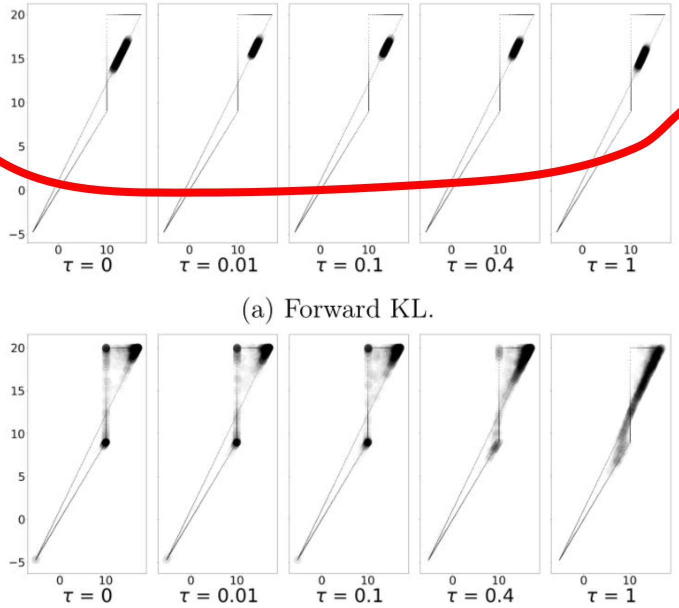


Figure 4.14: Switch-stay with 10 sample points, learning rate = 0.01, with RMSprop.

Switch-Stay, 10 points

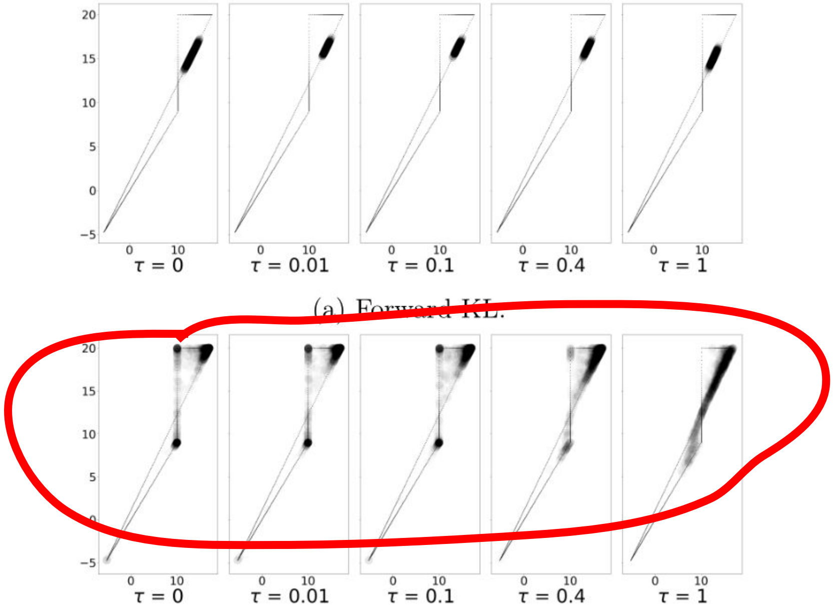


(a) Forward KL.

(b) Reverse KL.

Figure 4.14: Switch-stay with 10 sample points, learning rate = 0.01, with RMSprop.

Switch–Stay, 10 points

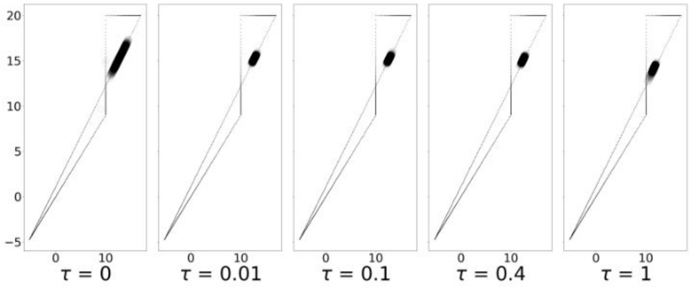


(a) Forward KL.

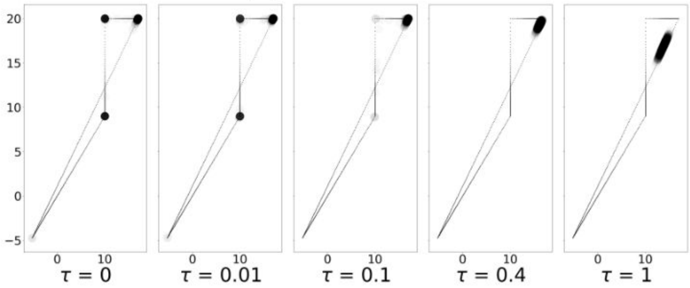
(b) Reverse KL.

Figure 4.14: Switch-stay with 10 sample points, learning rate = 0.01, with RMSprop.

Switch-Stay, 500 points



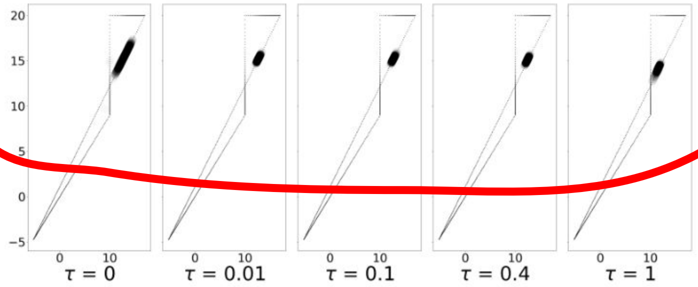
(a) Forward KL.



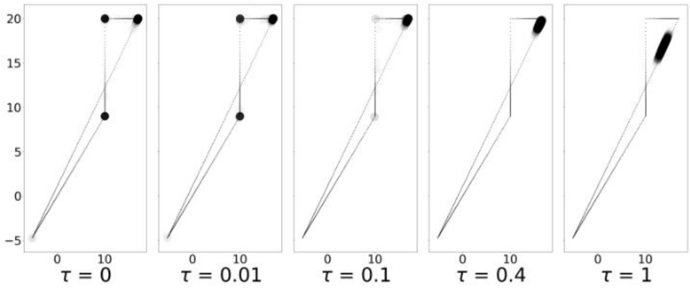
(b) Reverse KL.

Figure 4.15: Switch-stay with 500 sample points, learning rate = 0.01, with RMSprop.

Switch-Stay, 500 points



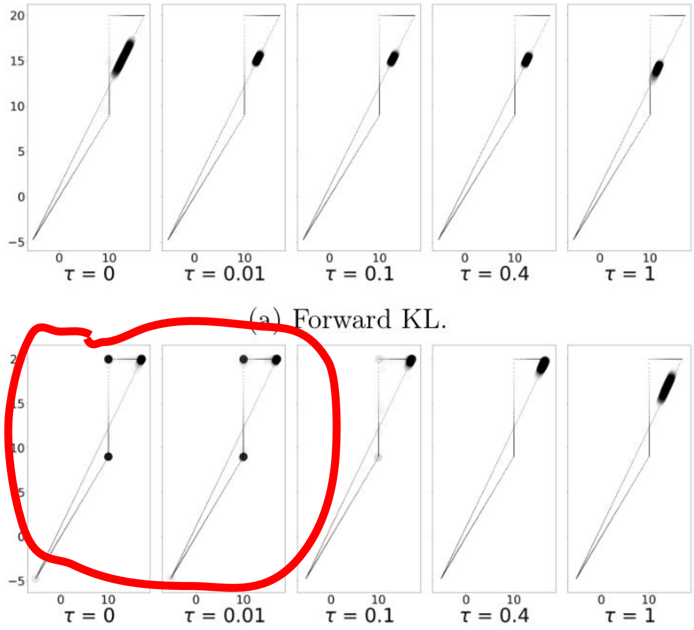
(a) Forward KL.



(b) Reverse KL.

Figure 4.15: Switch-stay with 500 sample points, learning rate = 0.01, with RMSprop.

Switch-Stay, 500 points



(a) Forward KL.

(b) Reverse KL.

Figure 4.15: Switch-stay with 500 sample points, learning rate = 0.01, with RMSprop.

The differences are negligible with discrete actions

Implementation

- Two-armed bandit
- Softmax policy

Two-armed Bandit Heatmap

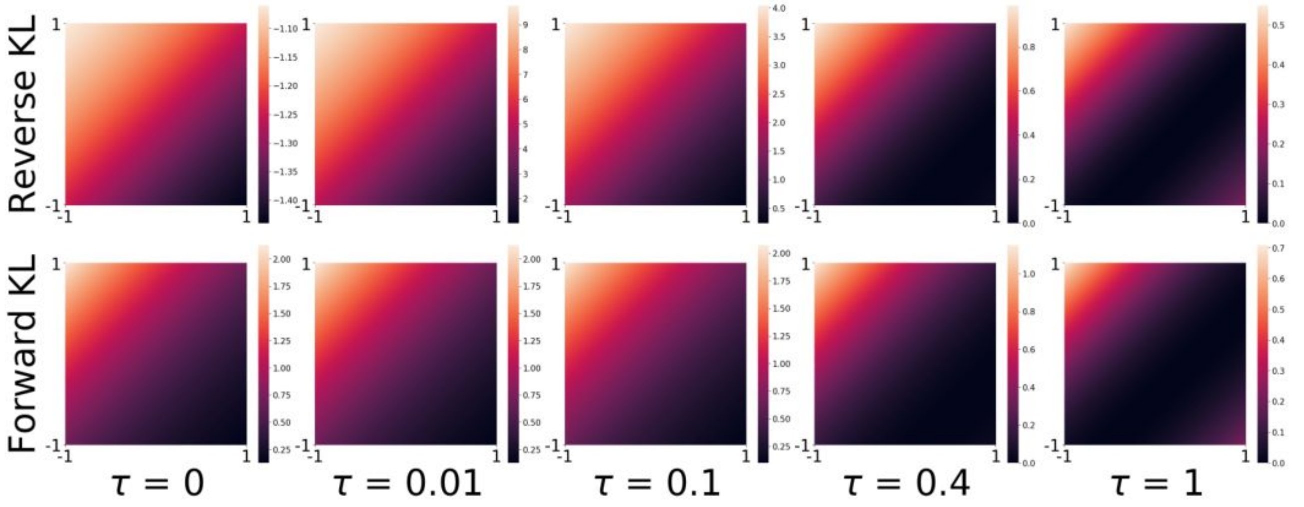
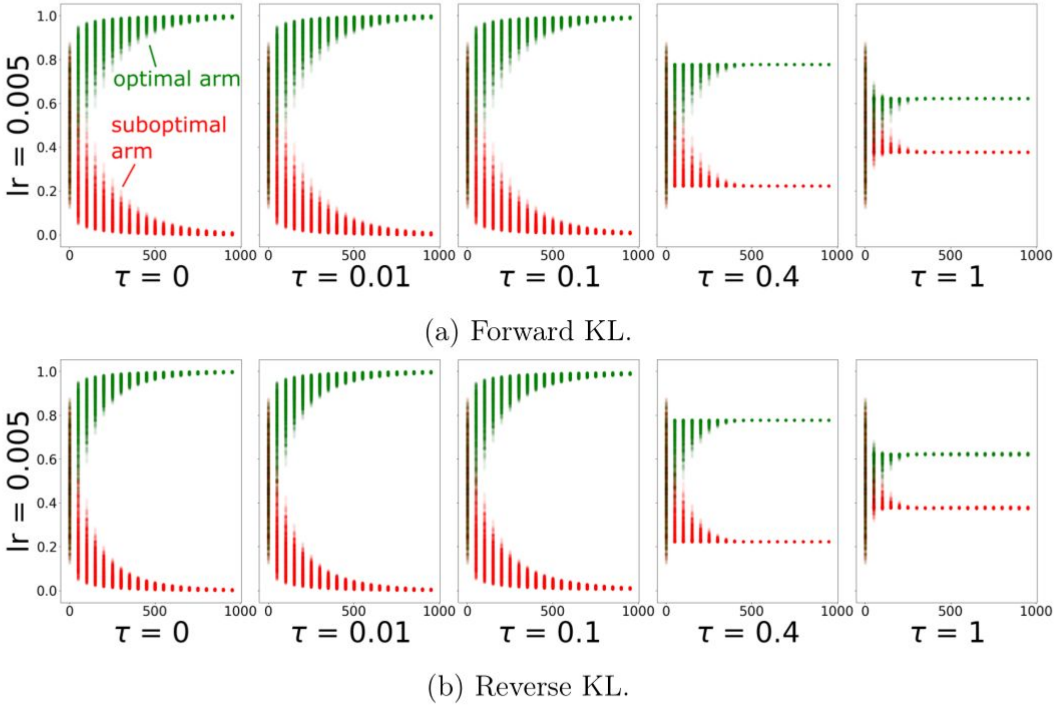
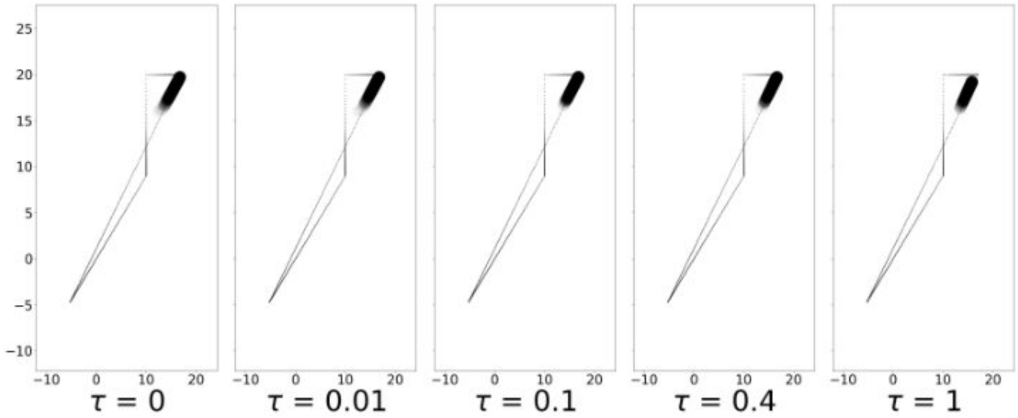


Figure 4.9: Heatmap for the KLs on the discrete bandit. In a given subplot, the x-axis is the logit for the optimal arm and the y-axis is the logit for the suboptimal arm.

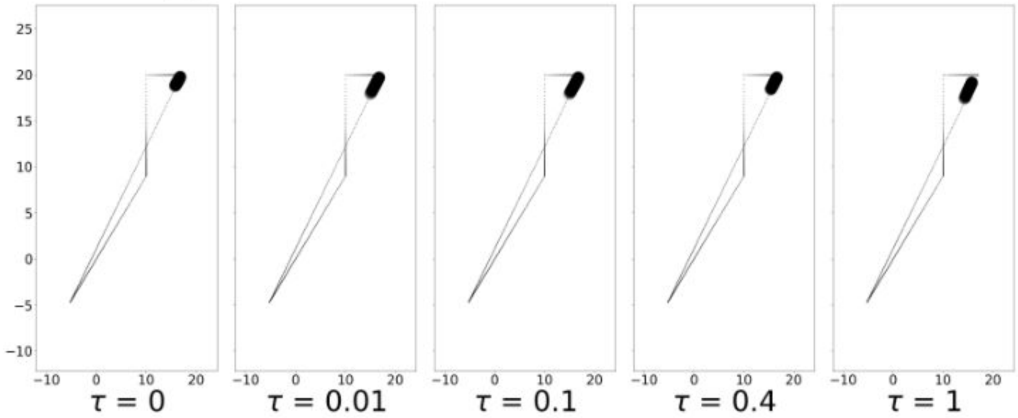
Tracking Bandit Iterates over Time



Switch–Stay



(a) Forward KL, learning rate $= 0.005$.



(b) Reverse KL, learning rate $= 0.005$.

Takeaways

1. Policy parameterisation is important
2. FKL has a smoother landscape
3. FKL solution may be more suboptimal
4. FKL more robust to stochasticity

Limitations

1. Exact critic was used
2. No function approximation
3. No stochastic rewards

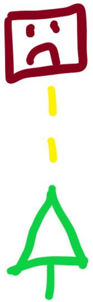
Large Experiments

Goals

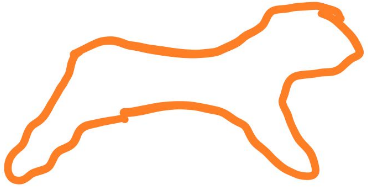
Understand what is true in more complicated environments

Environments

OpenAI Gym



MinAtar



Mujoco

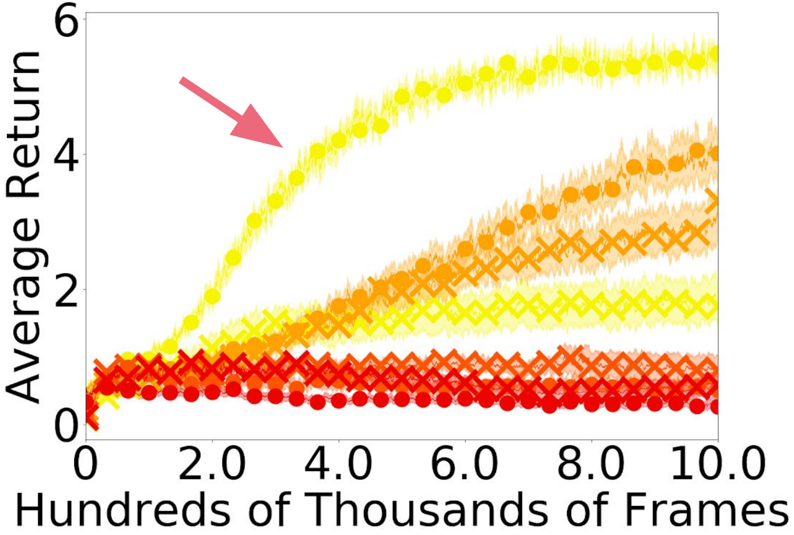
Implementation

1. 11 environments
2. Swept learning rates
3. 30 runs
4. Different network sizes for discrete-action setting
5. RMSprop
6. Last half of AUC

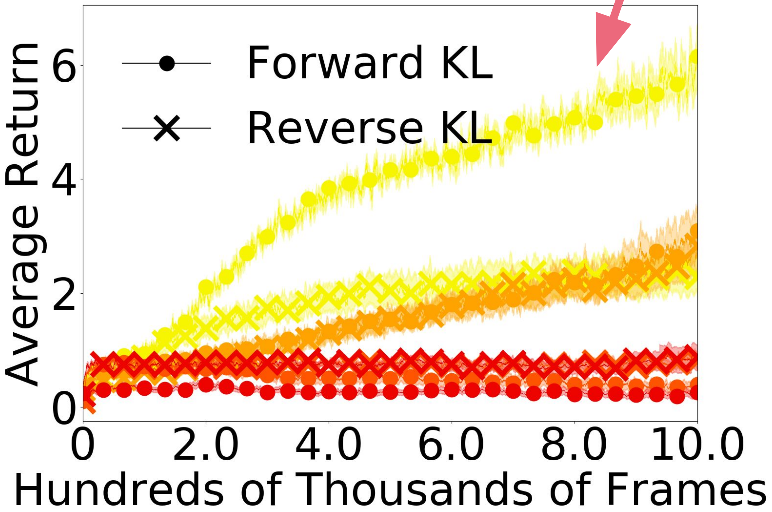
The Hard FKL performs surprisingly well sometimes

Seaquest

Hidden layer = 32

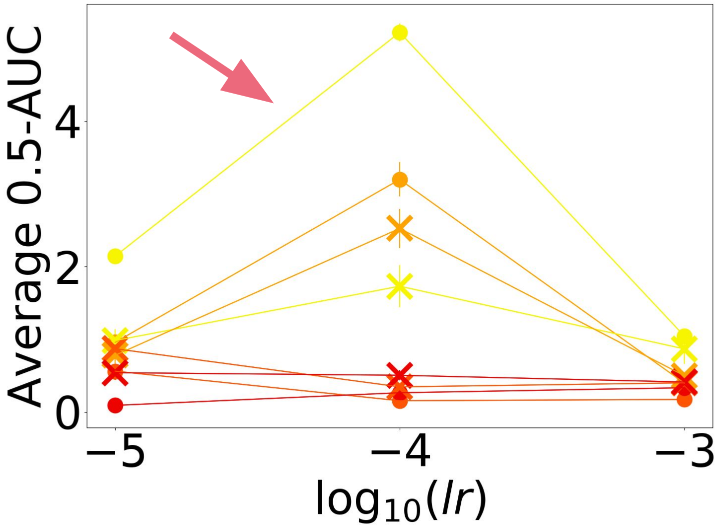


Hidden layer = 128

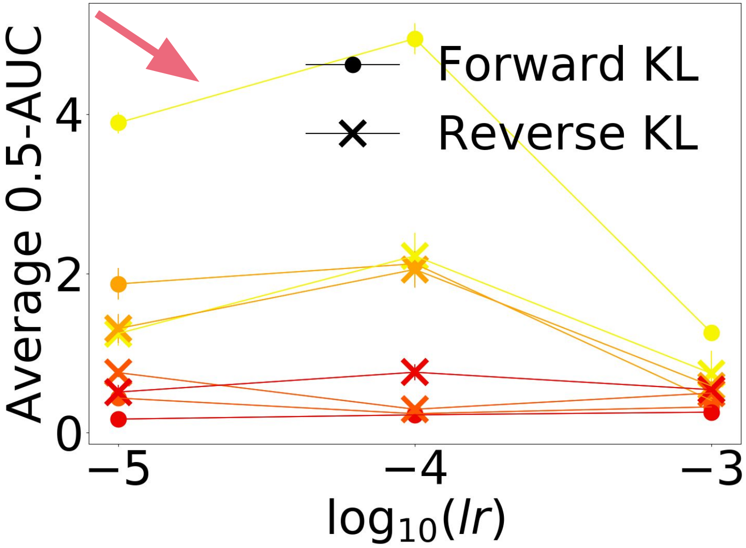


Seaquest

Hidden layer size = 32

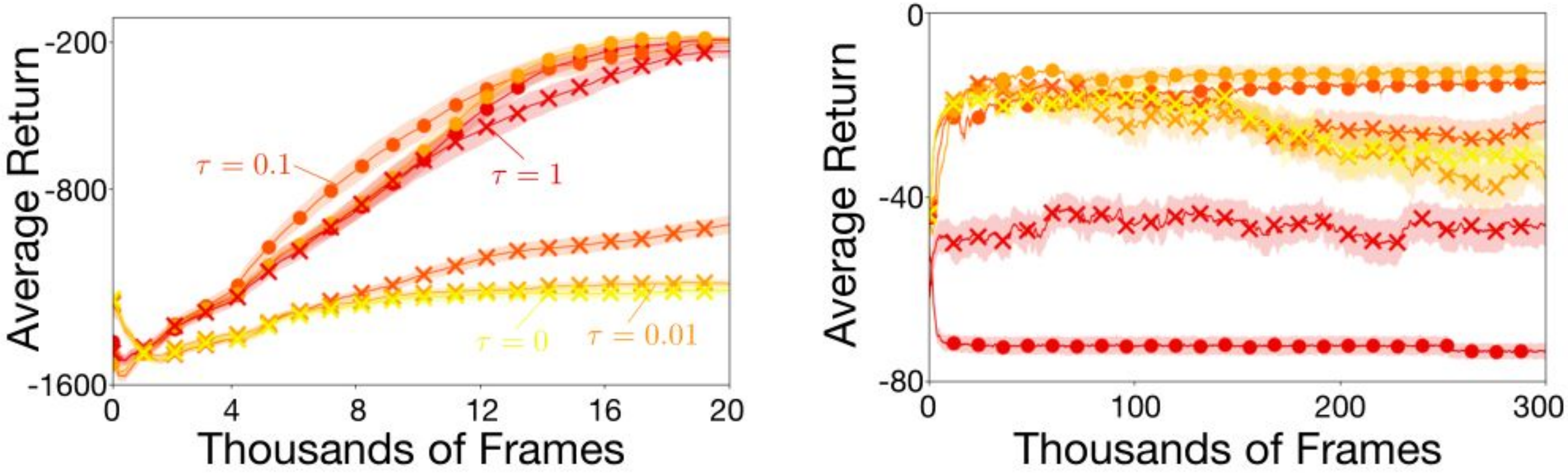


Hidden layer size = 128



The FKL might have a similar effect as entropy regularisation

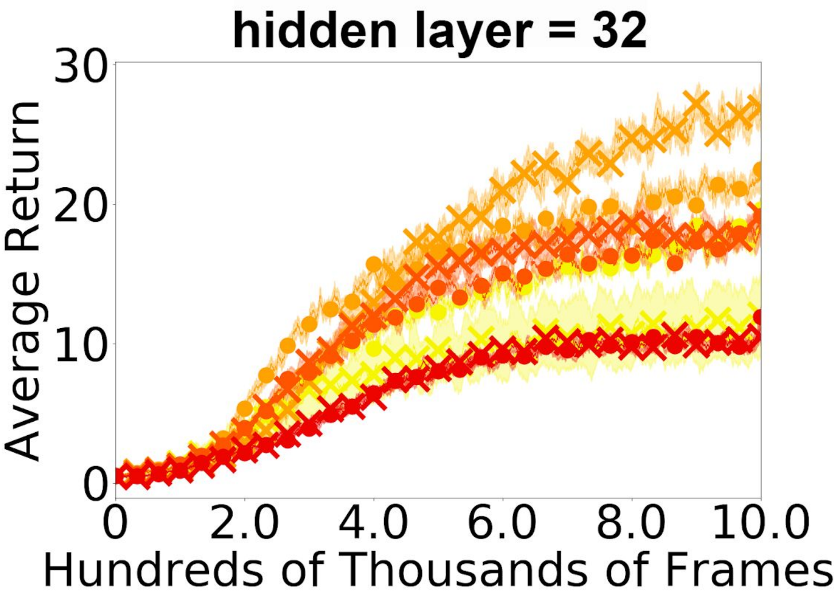
Mujoco



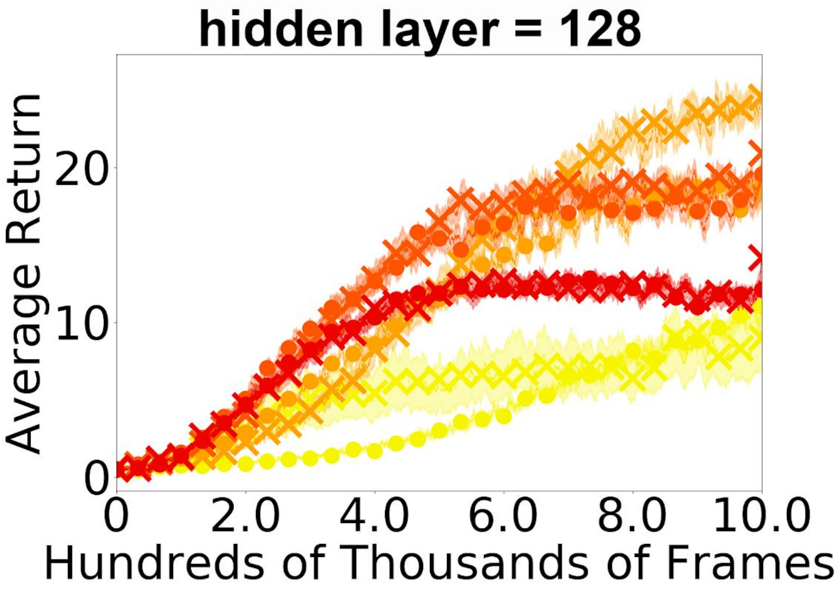
(a) Pendulum (b) Reacher

Neither KL seems generally superior

Asterix



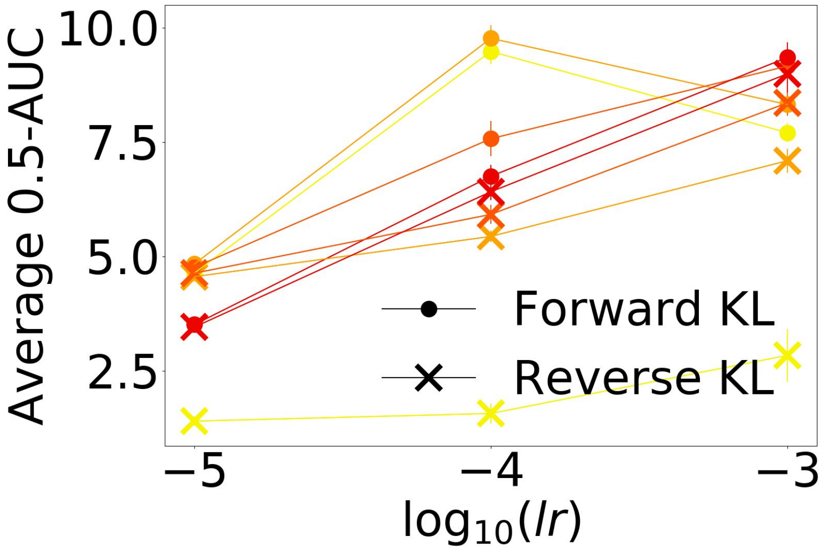
hidden layer = 32
Average Return
30
20
10
0
0
2.0
4.0
6.0
8.0
10.0
Hundreds of Thousands of Frames



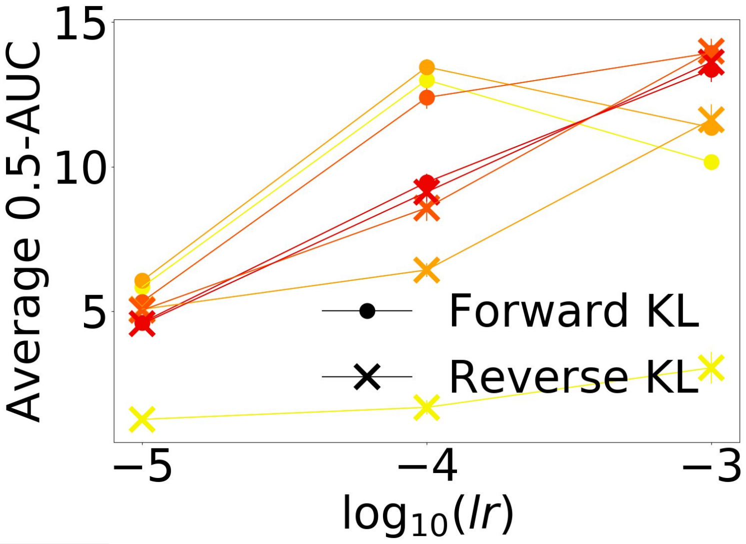
hidden layer = 128
Average Return
20
10
0
0
2.0
4.0
6.0
8.0
10.0
Hundreds of Thousands of Frames

Breakout

Hidden layer size = 32

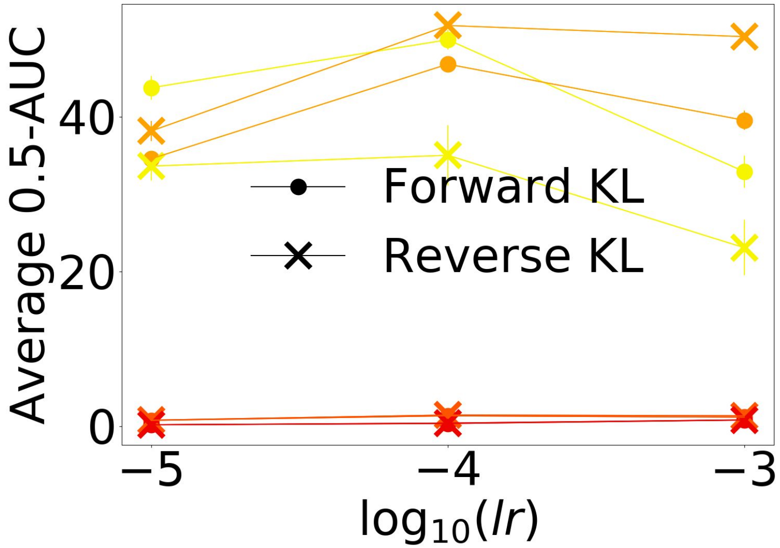


Hidden layer size = 128

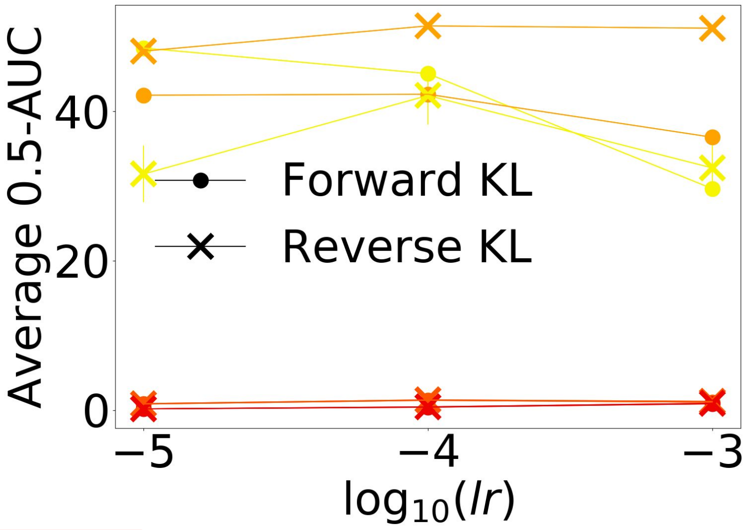


Freeway

Hidden layer = 32



Hidden layer = 128



Takeaways

1. The Hard FKL can perform surprisingly well
2. Neither KL seems generally superior

Limitations

1. Only RMSprop tested
2. Used tanh
3. Large range of environments

Concluding Thoughts

The FKL is promising

Reward structure

Inaccurate action-value estimates

Policy parameterizations

Target distributions

Thank You!

Email: achan4@ualberta.ca
Site: https://achan.ca
Twitter: @_achan96_