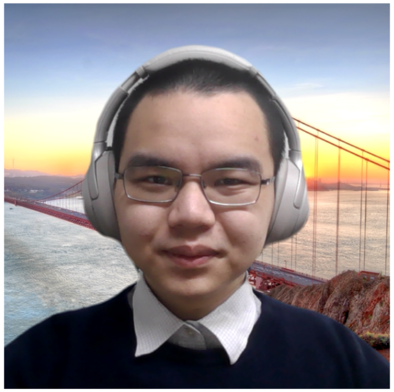
Provable Model-based Nonlinear Bandit (and RL)

Tengyu Ma

Stanford University

Kefan Dong
Stanford University



Jiaqi Yang
UC Berkeley

Toward a Theory for Deep Reinforcement Learning?

Is this the right timing?

- Q: why should I study deep RL theory
  - before understanding deep learning
  - before understanding out-of-domain generalization and uncertainty quantification with neural nets?
- My (debatable) answers:
  - Assuming computational oracle, deep RL theory may be easier than DL theory
  - Extrapolation to new domain in sequential setting may be easier than in static setting
    - online learning of neural nets is doable
    - but out-of-domain generalization for neural nets is challenging and requires assumptions on domain shift

State-of-the-art Analyses for RL (Until Recent 2-3 Months)

	B-Rank	B-Complete	W-Rank	Bilinear Class (this work)
Tabular MDP	✓	✓	✓	✓
Reactive POMDP [Krishnamurthy et al., 2016]	✓	✗	✓	✓
Block MDP [Du et al., 2019a]	✓	✗	✓	✓
Flambe / Feature Selection [Agarwal et al., 2020b]	✓	✗	✓	✓
Reactive PSR [Littman and Sutton, 2002]	✓	✗	✓	✓
Linear Bellman Complete [Munos, 2005]	✗	✓	✗	✓
Linear MDPs [Yang and Wang, 2019, Jin et al., 2020]	✓!	✓	✓!	✓
Linear Mixture Model [Modi et al., 2020b]	✗	✗	✗	✓
Linear Quadratic Regulator	✗	✓	✗	✓
Kernelized Nonlinear Regulator [Kakade et al., 2020]	✗	✗	✗	✓
$Q^\star$ "irrelevant" State Aggregation [Li, 2009]	✓	✗	✗	✓
Linear $Q^\star/V^\star$ (this work)	✗	✗	✗	✓
RKHS Linear MDP (this work)	✗	✗	✗	✓
RKHS Linear Mixture MDP (this work)	✗	✗	✗	✓
Low Occupancy Complexity (this work)	✗	✗	✗	✓
$Q^\star$ State-action Aggregation [Dong et al., 2020]	✗	✗	✗	✗
Deterministic linear $Q^\star$ [Wen and Van Roy, 2013]	✗	✗	✗	✗
Linear $Q^\star$ [Weisz et al., 2020]	Sample efficiency is not possible			

- **Claim:** none of these applies to even RL with general one-layer neural net approximation for dynamics (more evidence later)

[Bilinear Classes: A Structural Framework for Provable Generalization in RL. Du-Kakade-Lee-Lovett- Mahajan- Sun-Wang'21]

Neural Net Bandit: A Simplification With $H = 1$

- Reward function $\eta(\theta, a)$
  - $\theta \in \Theta$: model parameter
  - $a \in \mathcal{A}$: continuous action
  - Ex1: linear bandit: $\eta(\theta, a) = \theta^\top a$
  - Ex2: neural net bandit: $\eta(\theta, a) = \mathrm{NN}_\theta(a)$
- Realizable and deterministic reward setting:
  - Ground-truth $\theta^\star \in \Theta$
  - We observe the ground-truth reward $\eta(\theta^\star, a_t)$ after playing a_t
- Goal: to find the best arm

$$a^\star = \operatorname*{argmax}_{a \in \mathcal{A}} \eta(\theta^\star, a)$$

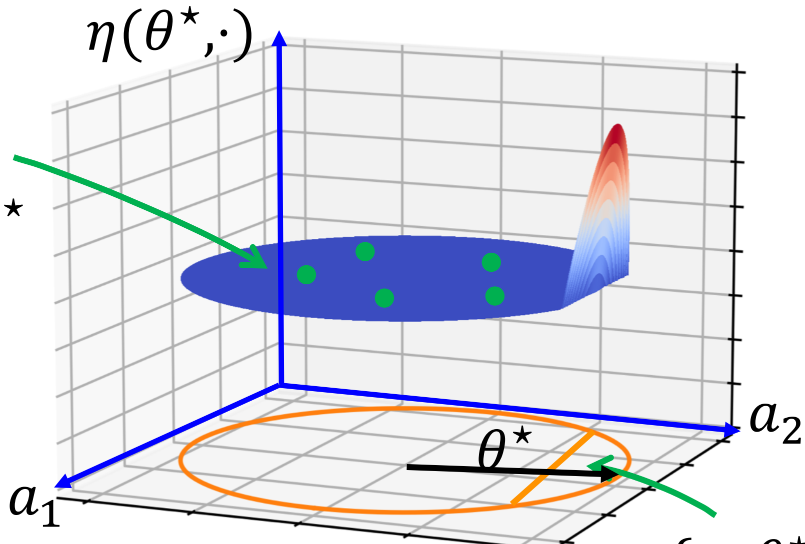
Even One-layer Neural Net Bandit is Statistically Hard!

- Θ and $\mathcal{A}$ are unit ℓ_2-balls in $\mathbb{R}^d$

$$\eta(\theta, a) = \mathrm{relu}(\theta^\top a - 0.9)$$

$$a^\star = \underset{||a||_2 \le 1}{\mathrm{argm}ax}\, \mathrm{relu}(\theta^{\star\top} a - 0.9) = \theta^\star$$

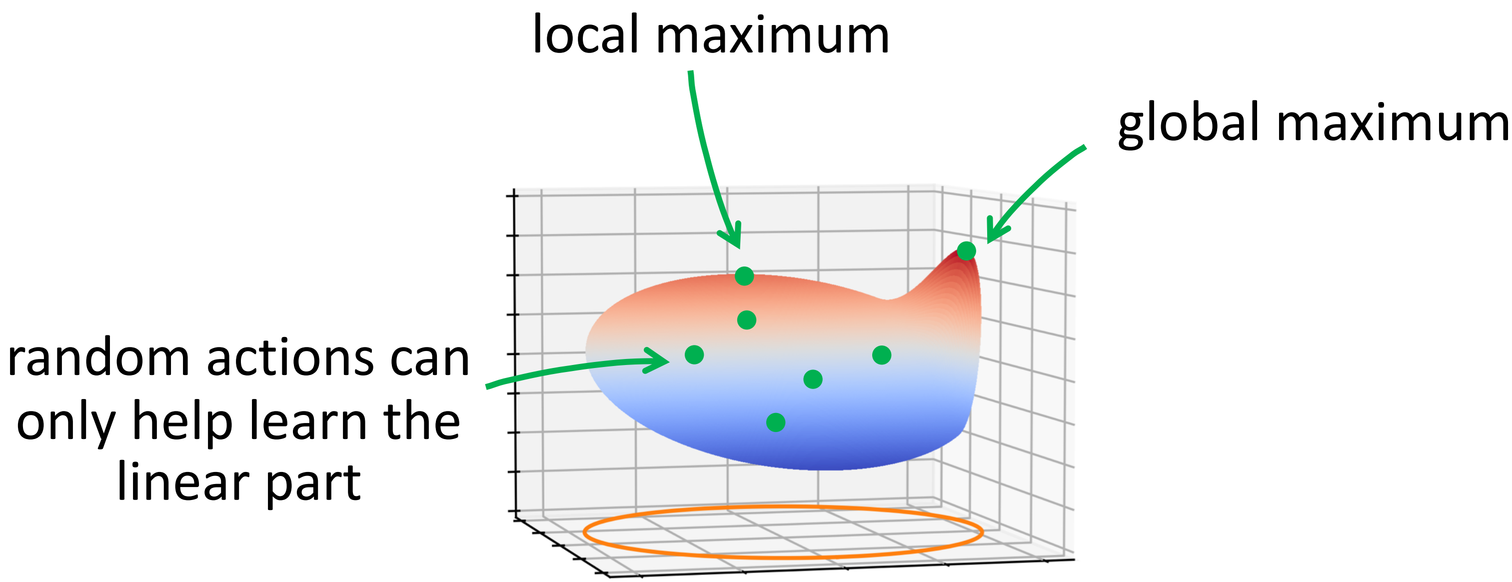
Random actions
have zero rewards
and no info about $\theta^\star$



$\{a\colon \theta^{\star\top} a \ge 0.9\}$ has
$\exp(-d)$ prob. mass

needle in a haystack!

Hard Instances Can Also Have Smooth and Non-Sparse Rewards



$$\eta((\gamma, \beta), a) = \gamma^\top a + c_0 \cdot \sigma(\beta^\top a - 0.9)$$

- Extendable to RL with nonlinear family of dynamics and known reward

State-of-the-art Analyses for RL

	B-Rank	B-Complete	W-Rank	Bilinear Class (this work)
Tabular MDP	✓	✓	✓	✓
Reactive POMDP [Krishnamurthy et al., 2016]	✓	✗	✓	✓
Block MDP [Du et al., 2019a]	✓	✗	✓	✓
Flambe / Feature Selection [Agarwal et al., 2020b]	✓	✗	✓	✓
Reactive PSR [Littman and Sutton, 2002]	✓	✗	✓	✓
Linear Bellman Complete [Munos, 2005]	✗	✓	✗	✓
Linear MDPs [Yang and Wang, 2019, Jin et al., 2020]	✓!	✓	✓!	✓
Linear Mixture Model [Modi et al., 2020b]	✗	✗	✗	✓
Linear Quadratic Regulator	✗	✓	✗	✓
Kernelized Nonlinear Regulator [Kakade et al., 2020]	✗	✗	✗	✓
$Q^\star$ "irrelevant" State Aggregation [Li, 2009]	✓	✗	✗	✓
Linear $Q^\star$/$V^\star$ (this work)	✗	✗	✗	✓
RKHS Linear MDP (this work)	✗	✗	✗	✓
RKHS Linear Mixture MDP (this work)	✗	✗	✗	✓
Low Occupancy Complexity (this work)	✗	✗	✗	✓
$Q^\star$ State-action Aggregation [Dong et al., 2020]	✗	✗	✗	✗
Deterministic linear $Q^\star$ [Wen and Van Roy, 2013]	✗	✗	✗	✗
Linear $Q^\star$ [Weisz et al., 2020]	Sample efficiency is not possible			

- Claim: none of these applies to even RL with general one-layer neural net approximations for dynamics
  - It's just impossible!

What's the Path Forward?

- Empirically deep RL still works well largely---it's the limitation of theory

- Option 1: change / weaken the goal
- Option 2: restrict to realistic family of problem instances
  - E.g., two-layer neural nets without bias (and sample complexity depends on width) [Huang et al.'21]
- Option 3: combine option 1&2?

A Proposed Paradigm
(Analogous to Non-convex Optimization Literature)

1. Convergences to local maxima for general instances

Focus of this talk

2. Analysis of the quality of local maxima of the ground-truth $\eta(\theta^\star,\cdot)$
   - All local maxima are global or satisfactory enough?

some concave examples

Baselines for Converging to Local Maxima: Zero-order Optimization for Bandit and Policy Gradient for RL

- Let $\eta^\star(a) = \eta(\theta^\star, a)$
- Zero-order optimization: estimate gradient $\nabla\eta^\star(a)$ from $\eta^\star(a)$
  - Estimating $\nabla\eta^\star(a)$ doesn't help estimating $\nabla\eta^\star(a')$
  - at least $O(d)$ sample complexity where d = action dimension

Q: can we leverage the model extrapolation to improve sample efficiency?

- Model-based methods are largely believed to be more sample-efficient than model-free methods
  - model = reward parameterization for bandit
  - model = (dynamics model, reward) for RL

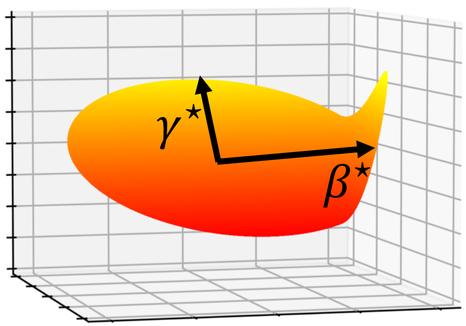
Main Results on Bandit

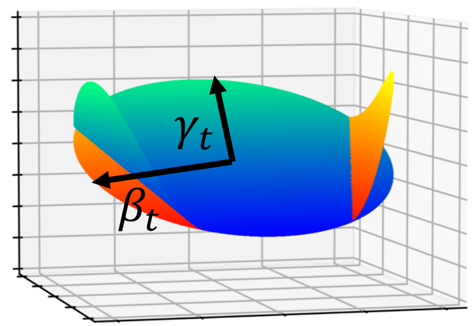
Theorem (informal): A model-based algorithm can converge to ϵ-approximate local maximum with $O(\mathfrak{R}(\Theta)/\epsilon^4)$ samples, where $\mathfrak{R}(\Theta)$ is a complexity measure of the model class $\{\theta: \eta(\theta,\cdot), \theta \in \Theta\}$.

- complexity measure = sequential Rademacher complexity (which appears to be similar to standard Rademacher complexity)

Does Classical Model-based UCB Converge to Local Max?

$$a_t, \theta_t = \underset{\substack{a \in \mathcal{A} \\ \theta \text{ fits past observations}}}{\operatorname{argmax}} \eta(\theta, a)$$





- Easy to learn γ
- UCB keeps optimistically guessing $(\gamma_t, \beta_t) = (\gamma^\star, \beta)$ and $a = \beta$ for some random β

$$\eta((\gamma, \beta), a) = \gamma^\top a + c_0 \cdot \sigma(\beta^\top a - 0.9)$$

- UCB fundamentally aims for global maximum and keeps exploring
- It also fails for deep RL empirically because the optimistic model fantasizes too much (anecdotal, [Luo et al.'18])

Where Does UCB Analysis Break?

- virtual reward: $\eta(\theta_t,\cdot)$
- real reward: $\eta(\theta^\star,\cdot)$

1. Exploration (virtual reward $\geq$ optimal reward)

by def. of optimism, $\eta(\theta_t, a_t) \geq \eta(\theta^\star, a^\star)$

2. Extrapolation (i.e., virtual $\approx$ real):

$$\sum_{t=1}^{T}\big(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\big)^2 \leq \sqrt{\underbrace{\mathrm{dim}(\Theta)}\cdot T}$$

- e.g., Eluder dim

- 1 + 2 $\Rightarrow$ $\eta(\theta^\star, a_t) \approx \eta(\theta_t, a_t) \geq \eta(\theta^\star, a^\star)$

- Step 2 fails for neural nets because
  - $\mathrm{dim}_{\mathrm{Eluder}}(\Theta) = \exp(d)$
  - b.c. learning θ_t suboptimally: we only know that θ_t fits past data

[Russo-Van Roy'13, Eluder Dimension and the Sample Complexity of Optimistic Exploration]

Our Idea: Re-Prioritizing the Two Steps

2. Extrapolation by an online learning (OL) algorithm

$$\underbrace{\sum_{t=1}^{T}\big(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\big)^2 \leq \mathrm{SRC}_T(\Theta)}$$

$$\sum_{t=1}^{T}\big(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\big)^2 \leq \underbrace{\mathrm{SRC}_T(\Theta)}_{\text{Sequential Rademacher Complexity}}$$

[Rakhlin-Sridharan-Tewari'15]

- For finite hypothesis Θ, $\mathrm{SRC_T}(\Theta) = \sqrt{\log|\Theta| \cdot T}$
- For neural nets:

  SRC = poly$(d) \cdot \sqrt{T}$ vs. Eluder dim = $\exp(d)$

  - SRC can be dimension-free and only depend on the weight norm
- Source of gains: OL oracle chooses θ_t better than UCB by stochastic predictions that hedges risks

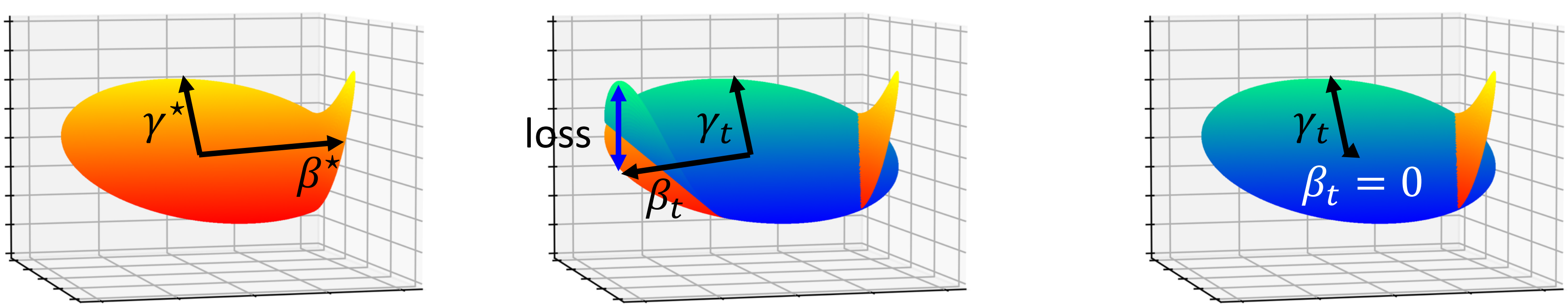
OL Oracle Extrapolates Better

$$\text{loss} = \sum\left(\ell(\theta_t, a_t) - \ell(\theta^\star, a_t)\right)^2$$

ground-truth $\eta(\theta^\star, \cdot)$

OL: $\beta_t = 0$ (to hedge the risk)

loss = 0 at action $a_t = \gamma_t$



UCB:

β_t is random (to be optimistic)

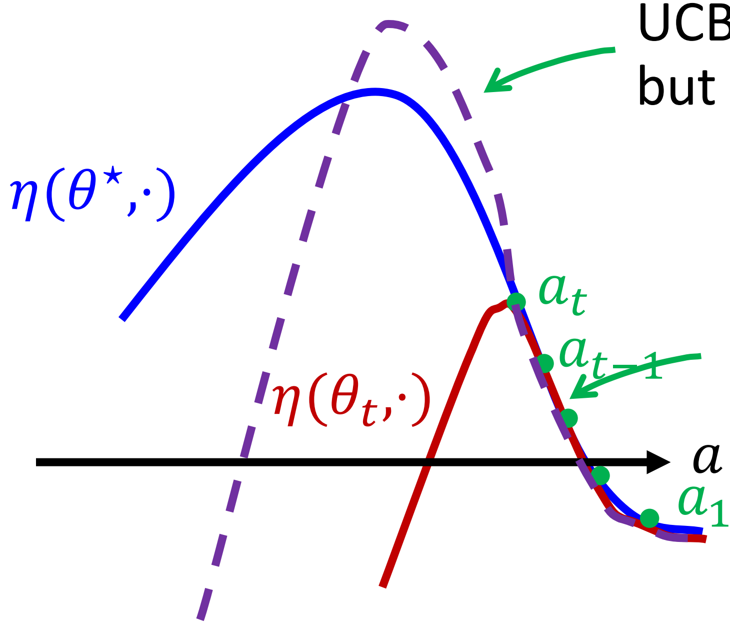
loss $\gg$ 0 at action $a_t = \beta_t$

Our Idea: Re-Prioritizing the Two Steps

1. Exploration ~~(virtual reward ≥ optimal reward)~~

2. Extrapolation by an online learning (OL)

$$\sum_{t=1}^{T}\big(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\big)^2 \leq \mathrm{SRC}_T(\Theta)$$



UCB will pick optimistic models but too optimistic

an arbitrary model that fits past observation

Our Idea: Re-Prioritizing the Two Steps

1. Exploration ~~(virtual reward ≥ optimal reward)~~

Local, model-based exploration: virtual reward increases incrementally

2. Extrapolation by an online learning (OL)

$$\sum_{t=1}^{T}\left(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\right)^2 \leq \mathrm{SRC}_T(\Theta)$$

- Step 1: modify the loss to predict directional reward gradient

$$\left(\eta(\theta, a) - \eta(\theta^\star, a)\right)^2 + \langle \nabla\eta(\theta, a') - \nabla\eta(\theta^\star, a'), u\rangle^2$$

- Step 2: take the best action according to the virtual reward
  - accurate gradient estimation guarantees local first-order improvements (exploration)

- Model-based learning of gradient is more sample-efficient than model-free estimate

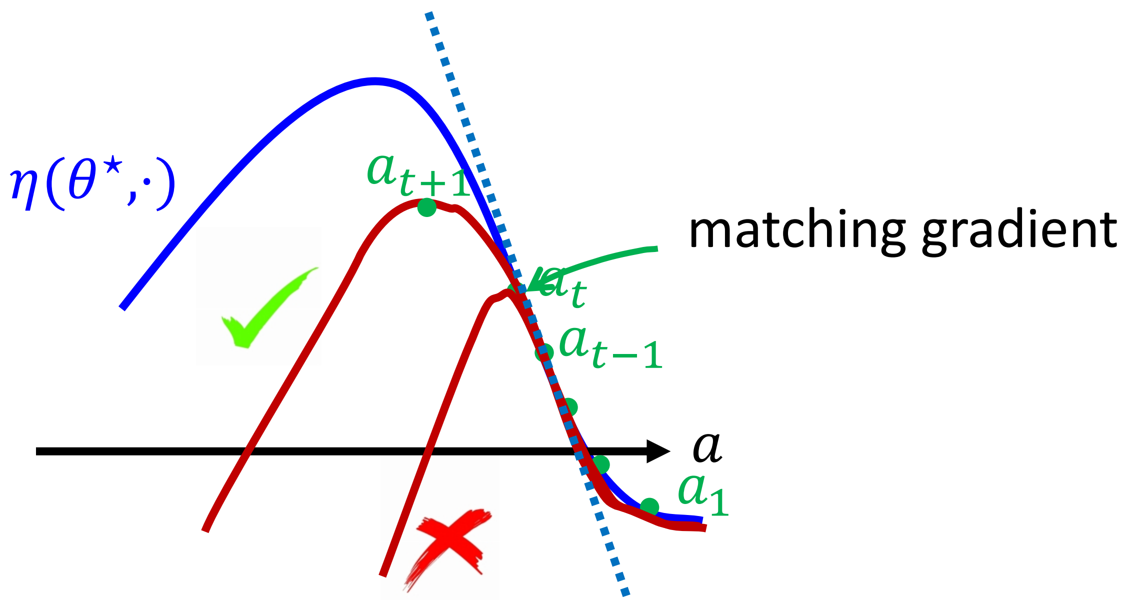
Our Idea: Re-Prioritizing the Two Steps

1. Exploration ~~(virtual reward ≥ optimal reward)~~

Local, model-based exploration: virtual reward increases incrementally

2. Extrapolation by an online learning (OL)

$$\sum_{t=1}^{T}\left(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\right)^2 \le \mathrm{SRC}_T(\Theta)$$



A Proposed Paradigm
(Analogous to Non-convex Optimization Literature)

1. Convergences to local maxima for general instances

Focus of this talk

2. Analysis of the quality of local maxima of the ground-truth $\eta(\theta^\star, \cdot)$
   - All local maxima are global or satisfactory enough?

some concave examples

Implications of the Theorem Where All Local Max are Global Max

- Linear bandit with structured model family: $\eta(\theta, a) = \theta^\top a$
  - Θ is finite: $O(\log|\Theta|)$ sample complexity
    - squareUCB [Foster-Rakhlin'21] depends on action dimension
  - Θ contains s-sparse vectors or only has s-degree of freedom: $O(s \log d)$ sample complexity

- Negative-weights neural net bandit: $\eta(W, a) = w_2^\top \sigma(W_1 a)$
  - assume $O(1)$ norms bounds on $||w_2||_1, |W_1|_1$
  - $\eta(W, \cdot)$ is concave in a --- all local max are global
  - SRC $\leq O\left(\sqrt{T}\right)$, sample complexity = $\tilde{O}(1)$
  - with general weights then can only find local max
    - conjecture: with random weights local max perhaps are very good?
  - NB: recovering the neural nets parameters does NOT seem to be easy (the learning loss is nonconvex)

A First-Cut Extension to Model-based RL

- Dynamics T_θ and policy π_ψ
- $\eta(\theta, \psi)$ = total expected return of policy π_ψ on dynamics T_θ
- Goal: find local max of $\eta(\theta^\star, \cdot)$

Challenge:

How does learning dynamics help estimate the $\eta(\theta^\star, \cdot)$ and its gradient?

- A result for stochastic policies

$$|\eta(\theta, \psi) - \eta(\theta^\star, \psi)| \lesssim \mathbb{E}_{s,a\sim T_{\theta^\star}, \pi_\psi}[\|T_\theta(s,a) - T_{\theta^\star}(s,a)\|^2]$$
$$\|\nabla\eta(\theta, \psi) - \nabla\eta(\theta^\star, \psi)\| \lesssim \mathbb{E}_{s,a\sim T_{\theta^\star}, \pi_\psi}[\|T_\theta(s,a) - T_{\theta^\star}(s,a)\|^2]$$
$$\|\nabla^2\eta(\theta, \psi) - \nabla^2\eta(\theta^\star, \psi)\| \lesssim \mathbb{E}_{s,a\sim T_{\theta^\star}, \pi_\psi}[\|T_\theta(s,a) - T_{\theta^\star}(s,a)\|^2]$$

- With many assumptions:
  - Value functions are Lipschitz/smooth in states and policy paramaters
  - $\nabla \log \pi_\psi$ is bounded in various ways
  - Not vacuous: e.g., $T(s,a) = NN_\theta(s+a)$ and linear policy can work

Summary

- Global regret for nonlinear models is statistically intractable
- ViOL converges to a local maximum with sample complexity that only depends on the model class complexity

Open questions:

- Bandit with stochastic rewards
- Faster convergence rate / smaller regret
- Analyze Q-learning algorithms?
- Analyze more special instances with global convergence

Thank you!