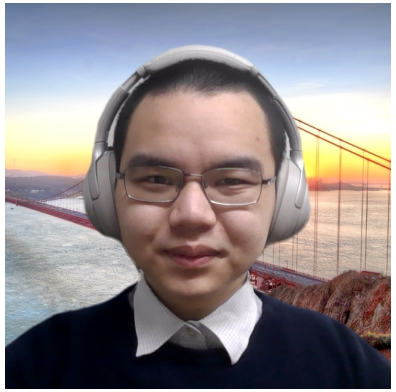
Provable Model-based Nonlinear Bandit and RL: Shelve Optimism, Embrace Virtual Curvature

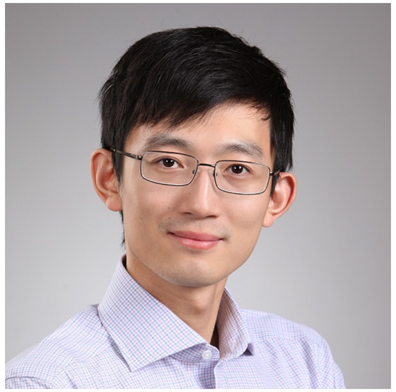
Kefan Dong

Stanford



Jiaqi Yang

Tsinghua University



Tengyu Ma

Stanford

Toward a Theory for Deep RL

Existing RL theory cannot apply to Neural Nets

- None of these give polynomial sample complexities for even one-layer NNs.

	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			

Du, Simon S., et al. "Bilinear Classes: A Structural Framework for Provable Generalization in RL."

Neural Net Bandit: A Simplification

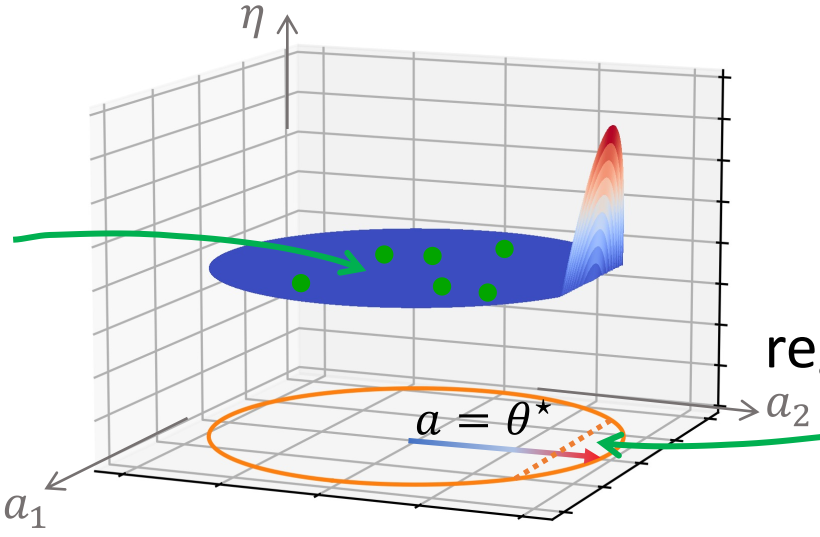
- Reward function $\eta(\theta, a)$
  - $\theta \in \Theta$: model parameter
  - $a \in \mathcal{A}$: continuous action
- Linear bandit: $\eta(\theta, a) = \theta^\top a$
- Neural net bandit: $\eta(\theta, a) = \mathrm{NN}_\theta(a)$
- Realizable and deterministic reward setting:
  - Agent observes ground-truth reward $\eta(\theta^\star, a)$ after playing action a
- Goal: finding the best action

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

Neural Net Bandit is Statistically Hard!

- $\Theta, \mathcal{A}$: unit ℓ_2-ball in $\mathbb{R}^d$
- $\eta(\theta, a) = \mathrm{relu}(\theta^\top a - 0.9)$, $\quad a^\star = \underset{||a||_2 \le 1}{\mathrm{argmax}}\, \mathrm{relu}(\theta^{\star\top} a - 0.9) = \theta^\star$

flat region



region with nonzero reward
$\{a: \theta^{\star\top} a \ge 0.9\}$

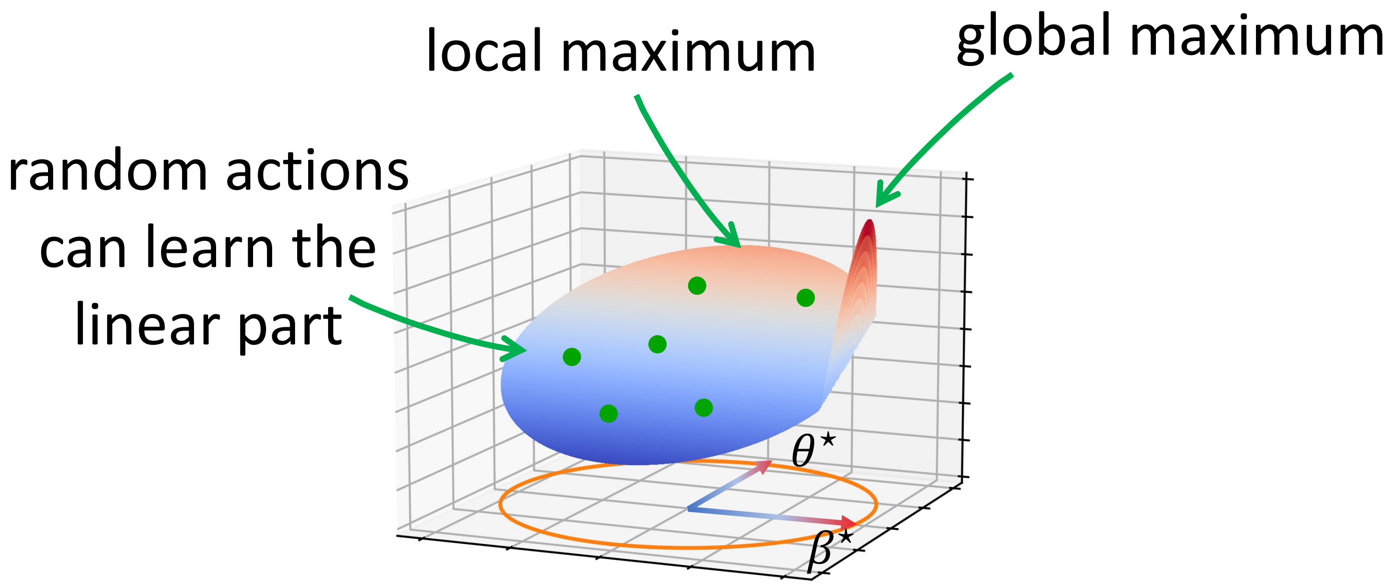
$\exp(-d)$ prob. mass

needle in a haystack!

Neural Net Bandit is Statistically Hard!

- Convergence to a global maximum is generally statistically intractable

- Existing RL theory cannot apply to NNs because they aim for global maximum



$$\eta((\theta, \beta), a) = \theta^\top a + 20 \cdot \mathrm{relu}(\beta^\top a - 0.9)$$

needle in a haystack!

A New Paradigm for Bandit/RL

1. Convergences to local maxima for general instances — This paper

2. Analysis of the landscape of the true reward $\eta(\theta^\star, \cdot)$

Main Results

- Theorem (informal): Under Lipschitz assumptions on η, our algorithm (ViOlin) converges to a ϵ-approximate local maxima in $\tilde{O}(\underbrace{R(\Theta)}_{\text{measures hardness of online learning w.r.t. model class}}\epsilon^{-8})$.

- Similar results for nonlinear RL (with many more assumptions and stochastic policies.)

Reviewing the Analysis of UCB

1. Optimization (high virtual reward):

$$\text{by optimism, } \eta(\theta_t, a_t) = \max_{\theta,a} \eta(\theta, a) \geq \eta(\theta^\star, a^\star)$$

2. Extrapolation (in average):

$$\sum_{t=1}^{T} \left(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\right)^2 \leq \sqrt{\underbrace{\mathrm{dim}_E}_{\text{Eluder dimension}}(\Theta) \cdot T}$$

- 1 + 2 $\Rightarrow$ $\eta(\theta^\star, a_t) \to \eta(\theta^\star, a^\star)$
- Step 2 fails for neural net models because $\mathrm{dim}_E(\Theta) \approx \exp(d)$

This result was independently proven in Li, Gene, Pritish Kamath, Dylan J. Foster, and Nathan Srebro. "Eluder Dimension and Generalized Rank."

Re-Prioritizing the Two Steps

1. Extrapolation by online learning (OL) oracles:

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\right)^2\right] \leq \tilde{O}\left(\sqrt{R(\Theta)\mathrm{T}}\right)$$

OL oracle outputs a distribution of θ_t

Sequential Rademacher Complexity
[Rakhlin-Sridharan-Tewari'15]

- For finite hypothesis Θ, $R(\Theta) = \log|\Theta|$
- For neural nets:

  $R(\Theta)$ = poly(d) vs. Eluder dim = $\exp(d)$

Rakhlin, Alexander, Karthik Sridharan, and Ambuj Tewari. "Online learning via sequential complexities." 2015.

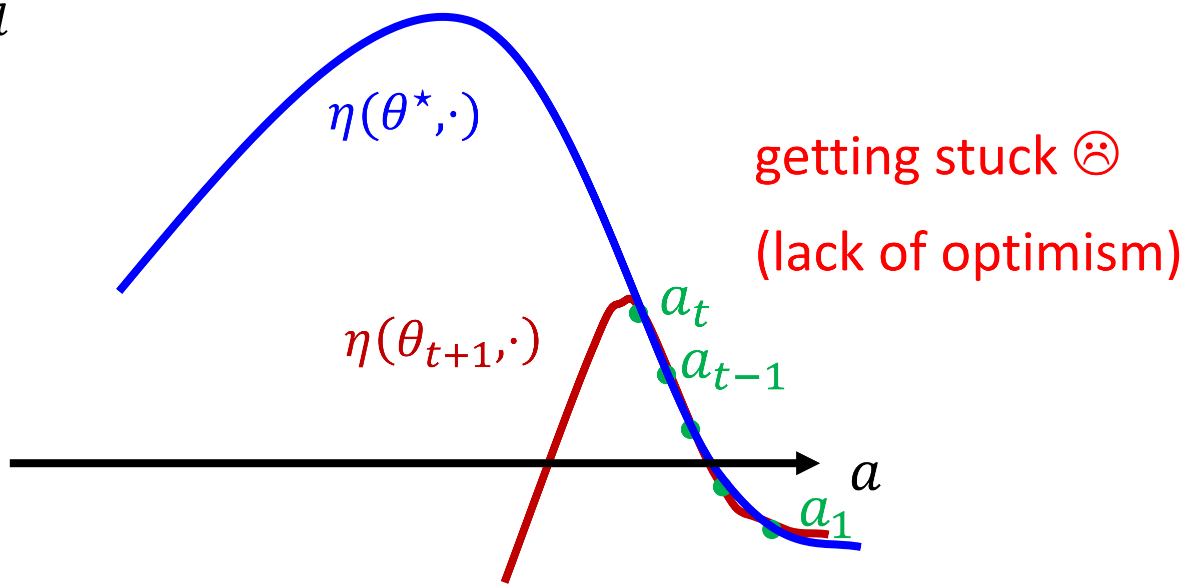
Re-Prioritizing the Two Steps

1. Extrapolation by online learning (OL) oracles

$$\mathbb{E}\left[\sum_{t=1}^{T}\left(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\right)^2\right] \le \sqrt{R(\Theta)T\ \text{polylog}(\mathrm{T})}$$

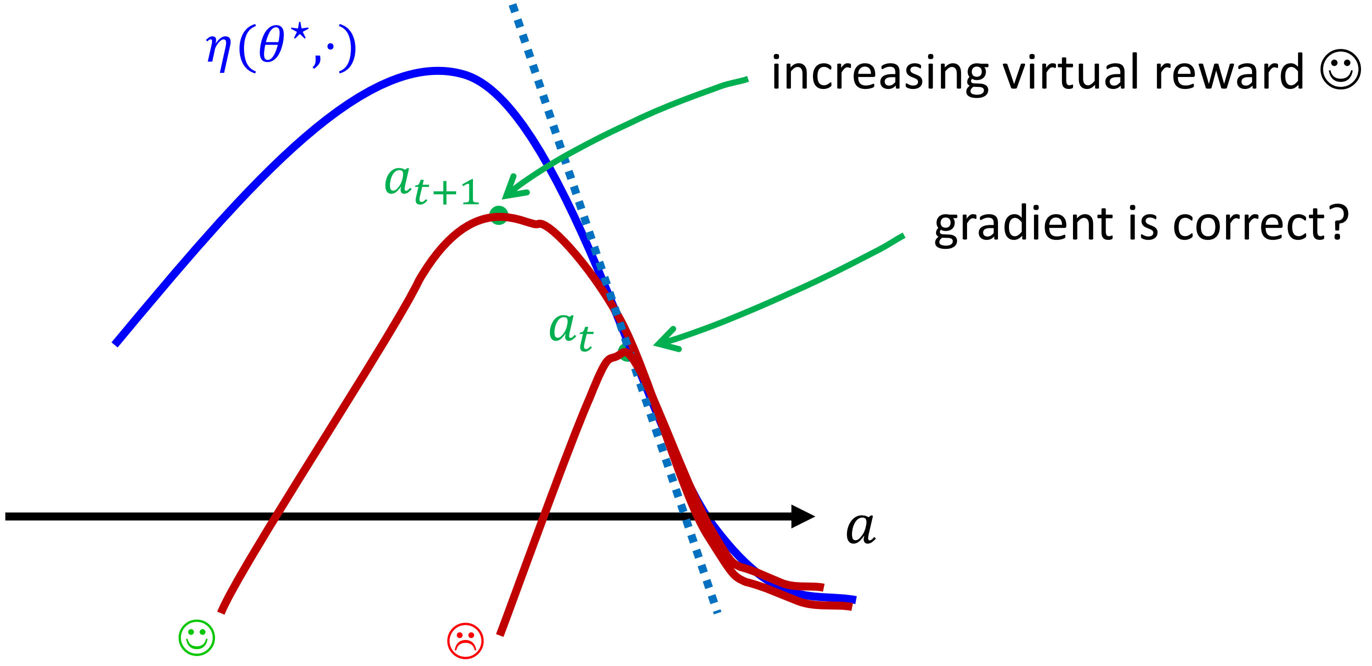
2. High virtual reward:

best attempt: $a_t = \underset{a\in\mathcal{A}}{\operatorname{argmax}}\, \mathbb{E}[\eta(\theta_t, a)]$



Embrace Virtual Curvature

- Need the online learner to work harder to guarantee an increasing virtual reward
- Estimating the curvature: learn θ_t such that
  1. $\eta(\theta_t, a_t) \approx \eta(\theta^\star, a_t)$
  2. $\nabla_a \eta(\theta_t, a_t) \approx \nabla_a \eta(\theta^\star, a_t)$
  3. $\nabla_a^2 \eta(\theta_t, a_t) \approx \nabla_a^2 \eta(\theta^\star, a_t)$



Algorithm and Theorem

$$\ell_t(\theta) = \left(\eta(\theta, a_t) - \eta(\theta^\star, a_t)\right)^2 + \left(\eta(\theta, a_{t-1}) - \eta(\theta^\star, a_{t-1})\right)^2$$
$$+\langle \nabla\eta(\theta, a_{t-1}) - \nabla\eta(\theta^\star, a_{t-1}), u_t\rangle^2$$

Computed by finite difference

- ViOlin (**Vi**rtual Ascent with **Onlin**e Model Learner)

1. Sample $u_t \sim \mathcal{N}(0, I)$
2. Use OL to minimize losses ℓ_t and get a distribution of θ_t
3. Take $a_t = \mathrm{argmax}_a \, \mathbb{E}_{\theta_t}[\eta(\theta_t, a)]$

- Theorem (informal): Under Lipschitz assumptions on η, ViOlin converges to a ϵ-approximate local maxima in $\tilde{O}(R(\Theta)\epsilon^{-8})$.

Instantiations

- Linear bandit with structured model family: $\eta(\theta, a) = \theta^\top a$
  - Θ is finite: $\mathrm{poly}(\log|\Theta|)$ sample complexity
  - Θ contains s-sparse vectors: $\mathrm{poly}(s, \log d)$ sample complexity
  - local maximum are global because $\eta(\theta^\star, \cdot)$ is concave.
  - only hold for deterministic reward

- Neural net bandit: $\eta(W, a) = w_2^\top \sigma(W_1 a)$
  - assume $O(1)$ norms bounds on $||w_2||_1, \|W_1\|_{\infty\to\infty}$
  - $R(W) \leq \tilde{O}(1)$
  - sample complexity for local max = $\tilde{O}(1)$
  - Local maximum are global for input-concave neural nets

Summary

- Global convergence for nonlinear models is statistically intractable
- ViOlin: convergence to a local maximum with sample complexity that only depends on the model class
- Similar results for nonlinear RL (with many more assumptions and stochastic policies.)

Thank you for your attention ☺