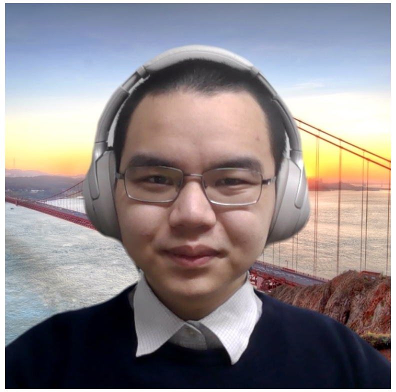
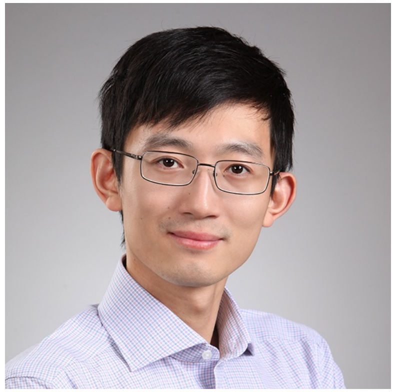
Provable Model-based Nonlinear Bandit and RL: Shelve Optimism, Embrace Virtual Curvature

Kefan Dong



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

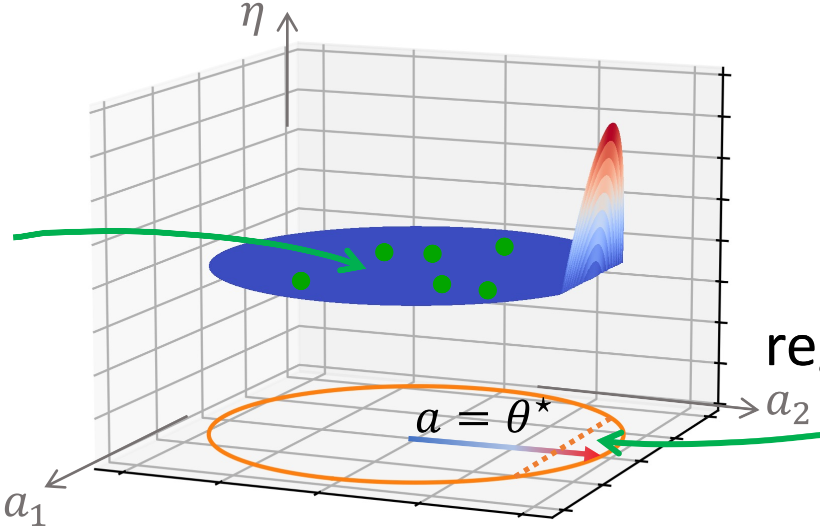
- Reward function $\eta(\theta, a)$
  - $\theta \in \Theta$: model parameter
  - $a \in \mathcal{A}$: continuous action
- Linear bandit: $\eta(\theta, a) = \theta^\top a$
- Neural net bandit: $\eta(\theta, a) = \mathrm{NN}_\theta(a)$
- Realizable and deterministic reward setting:
  - Agent observes ground-truth 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 $\mathbb{R}^d$
- $\eta(\theta, a) = \mathrm{relu}(\theta^\top a - 0.9)$, $\quad a^\star = \operatorname{argmax}_{||a||_2 \le 1} \mathrm{relu}(\theta^{\star\top} a - 0.9) = \theta^\star$

flat region



region with nonzero reward

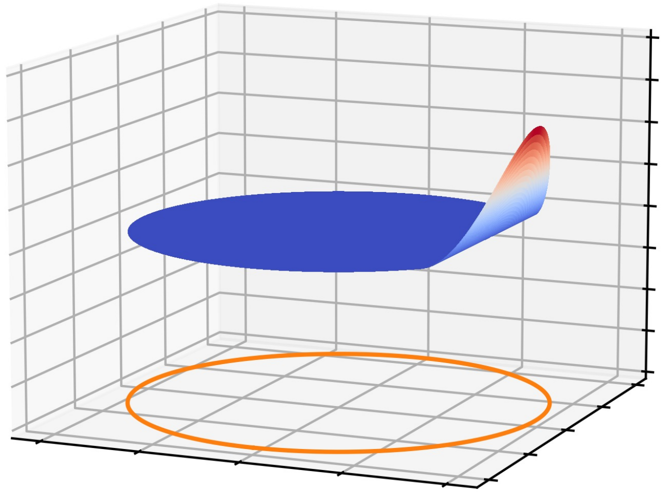
$\{a: \theta^{\star\top} a \ge 0.9\}$

$\exp(-d)$ prob. mass

needle in a haystack!

Neural Net Bandit is Statistically Hard!

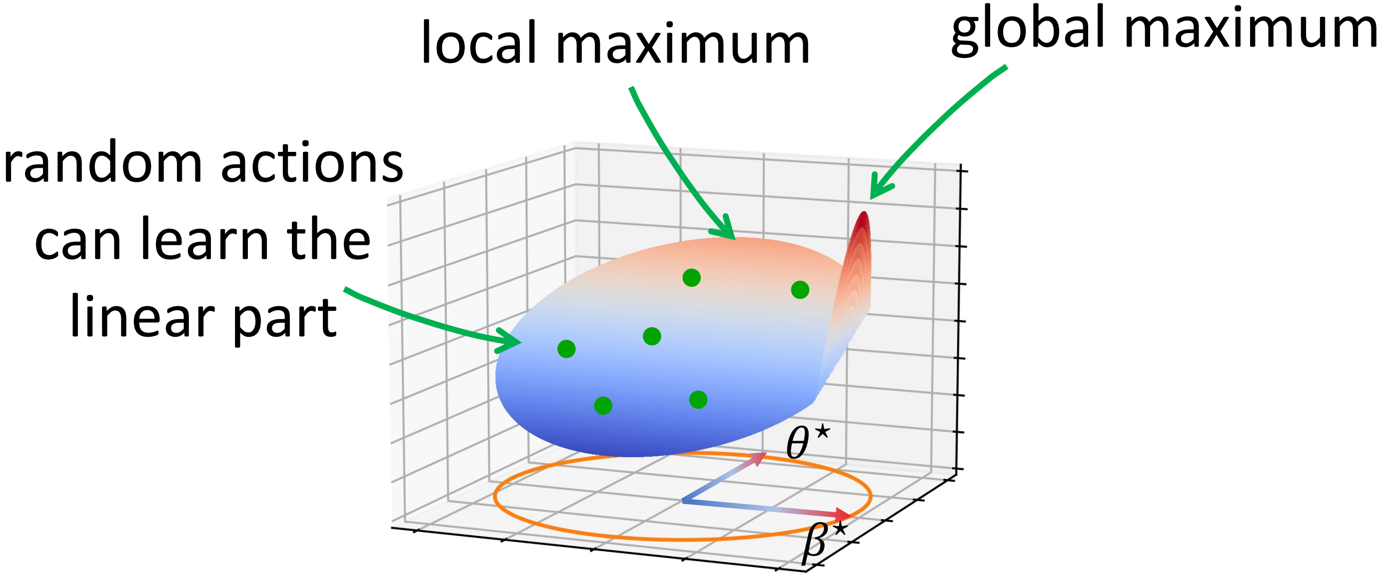
- $\Theta, \mathcal{A}$: unit ℓ_2-ball in $\mathbb{R}^d$
- $\eta(\theta, a) = \mathrm{relu}(\theta^\top a - 0.9)^2$



smoothed version

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 talk

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

Main Results

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

measures hardness of online learning w.r.t. model class

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

Baseline: Zero-order Optimization for Bandit

- True reward $f(a) = \eta(\theta^\star, a)$
- Zero-order optimization:
  - estimate gradient $\nabla f(a)$ by finite difference

    For $\xi \sim \mathcal{N}(0, I)$ and $\epsilon > 0$,
    $$\frac{1}{\epsilon}\mathbb{E}\left[\xi\big(f(a+\epsilon\xi) - f(a)\big)\right] \approx \mathbb{E}[\xi\xi^\top \nabla f(a)] = \nabla f(a)$$

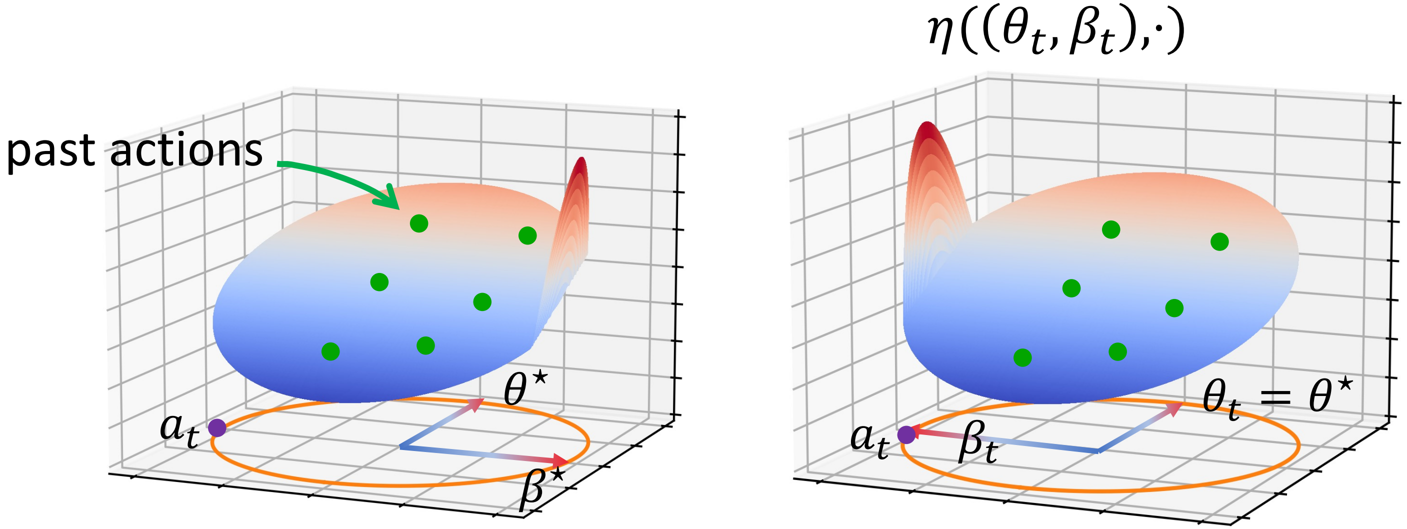
  - $\Omega(d)$ sample complexity

- Our key idea: leverage model extrapolation

Model-based UCB Does Even Not Converge To Local Max

$$a_t, \theta_t = \underset{a \in \mathcal{A}, \underbrace{\theta \in \Theta_t}_{\text{confidence region}}}{\operatorname{argmax}} \eta(\theta, a)$$

- Θ_t pins down $\theta^\star$ but has no clue about $\beta^\star$
- UCB keeps guessing β_t
- and choses $a_t = \beta_t$



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

- UCB over-explores and doesn't converge in polynomial steps
- In partice, deep RL methods with optimism also over-explore

Reviewing the Analysis of UCB

1. Optimization (high virtual reward):

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

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 $\Rightarrow$ $\eta(\theta^\star, a_t) \rightarrow \eta(\theta^\star, a^\star)$
- 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:

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

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:

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

  or the weight norm

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

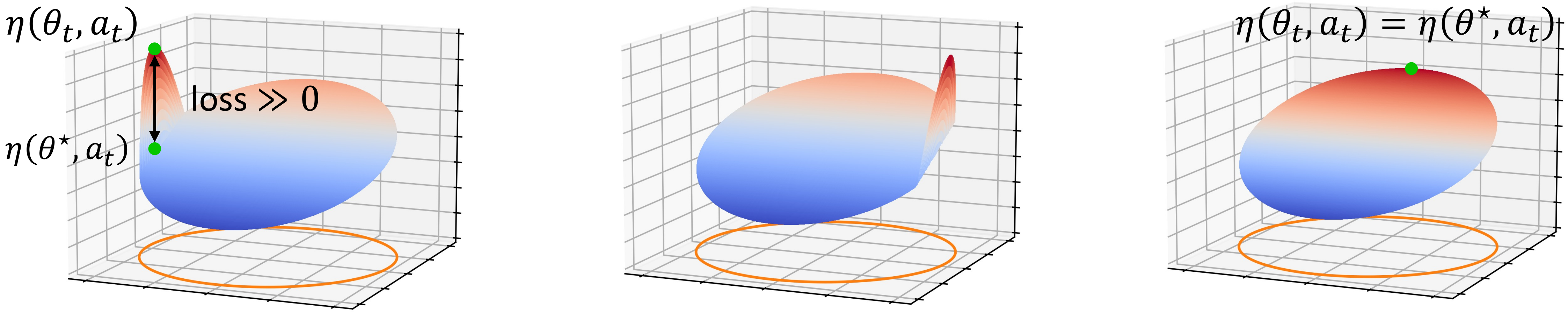
OL Oracle Extrapolates Optimally

Extrapolation error: $\sum_{t=1}^{T}\big(\eta(\theta_t, a_t) - \eta(\theta^\star, a_t)\big)^2$

UCB: loss $\gg$ 0

Ground truth

OL oracle: loss = 0
(for most of the times)



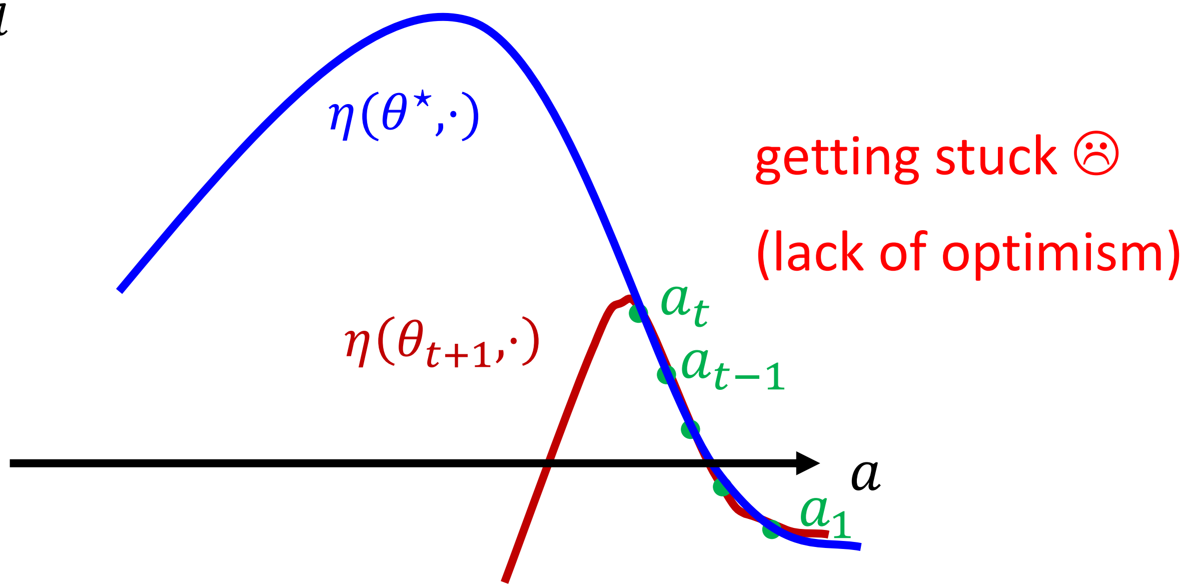
Re-Prioritizing the Two Steps

1. Extrapolation by online learning (OL) oracles

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

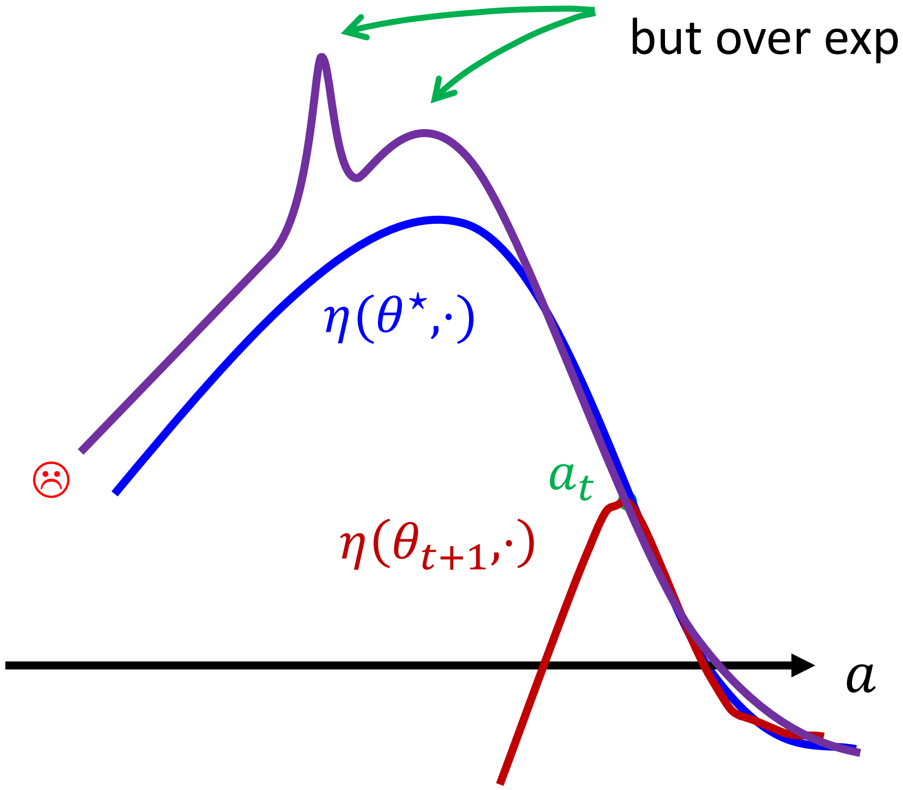
2. High virtual reward:

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



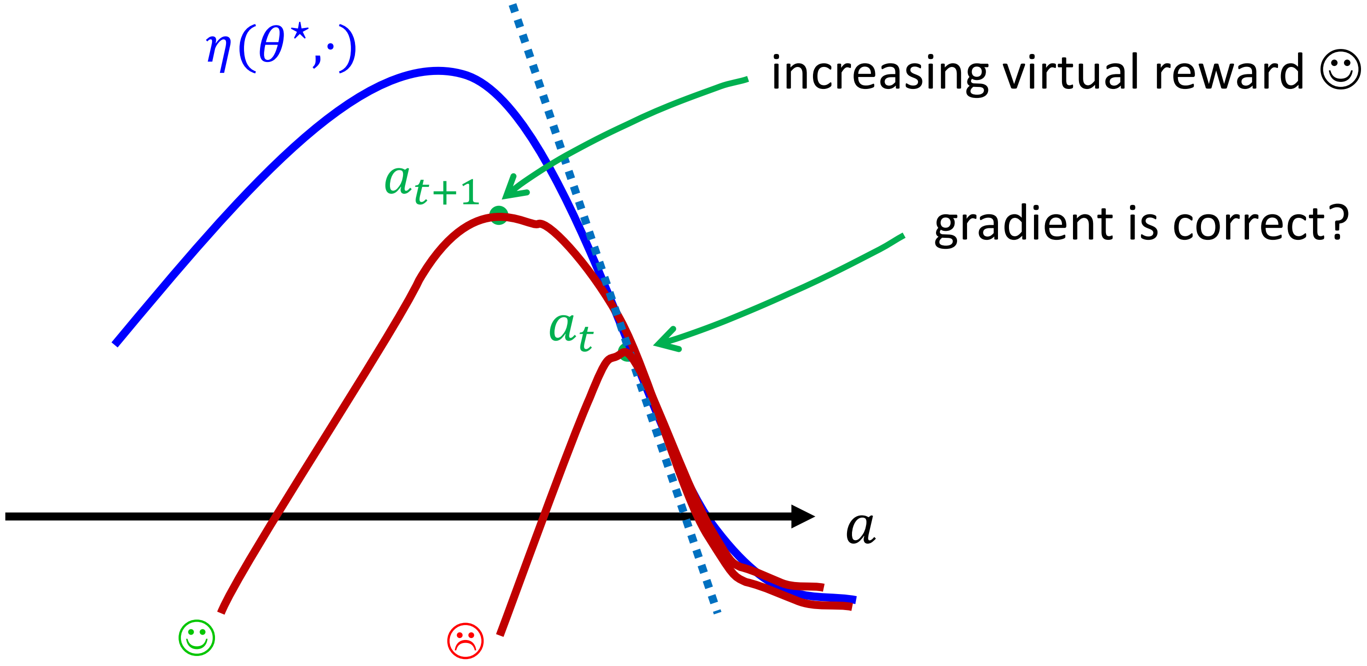
Re-Prioritizing the Two Steps

UCB will pick optimistic models but over explores



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. $\nabla_a^2 \eta(\theta_t, a_t) \approx \nabla_a^2 \eta(\theta^\star, a_t)$



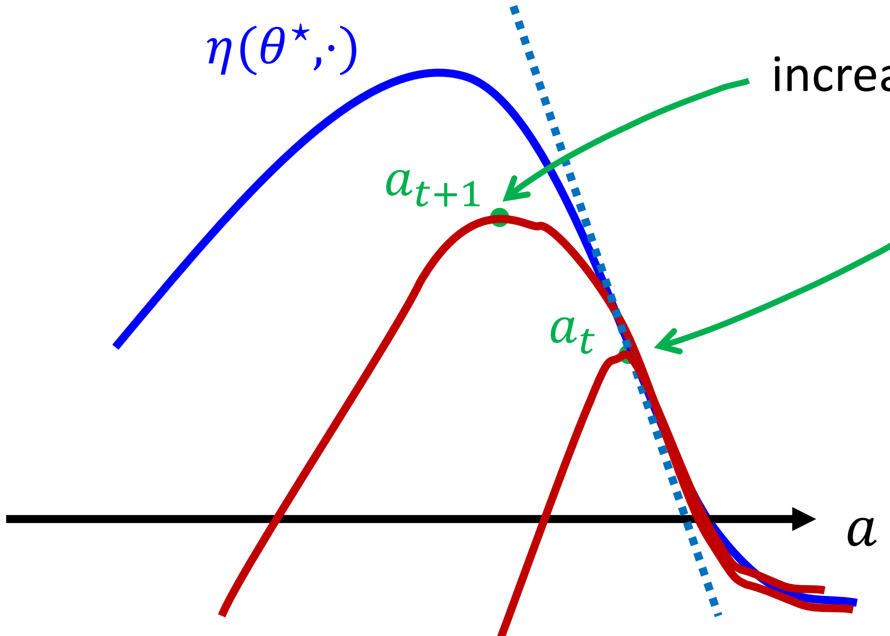
Virtual Improvements With Curvature Estimates

- Assume $\eta(\theta,\cdot)$ is smooth (with bounded 2nd derivative)
- $\eta(\theta^\star, a_{t+1}) \approx \mathbb{E}_{\theta_{t+1}}[\eta(\theta_{t+1}, a_{t+1})]$

$$\geq \mathbb{E}_{\theta_{t+1}}[\eta(\theta_{t+1}, a_t)] + \Omega(||\mathbb{E}_{\theta_{t+1}}[\nabla\eta(\theta_{t+1}, a_t)]||_2^2)$$

$$\approx \eta(\theta^\star, a_t) + \Omega(||\nabla\eta(\theta^\star, a_t)||_2^2)$$

OL guarantee



increasing virtual reward ☺

gradient is correct

"descent lemma":

$$\max_a f(a) \geq f(a_0) + \Omega(||\nabla f(a_0)||_2^2)$$

Virtual Improvements With Curvature Estimates

- Consider online learning problem with the idealized $\bar{\ell}_t$

$$\bar{\ell}_t(\theta) = \left(\eta(\theta, a_t) - \eta(\theta^\star, a_t)\right)^2 + \left(\eta(\theta, a_{t-1}) - \eta(\theta^\star, a_{t-1})\right)^2$$
$$+ ||\nabla\eta(\theta, a_{t-1}) - \nabla\eta(\theta^\star, a_{t-1})||_2^2$$

- ViOL (**Vi**rtual Ascent with **O**nline Model **L**earner)

1. Use OL to minimize losses $\bar{\ell}_t$ (assuming $\bar{\ell}_t$ is accessible) and get a distribution of θ_t
2. Take $a_t = \mathrm{argmax}_a\, \mathbb{E}_{\theta_t}[\eta(\theta_t, a)]$

Lemma: If online learning for $\bar{\ell}_t$ has regret

$$\mathbb{E}\left[\sum_{t=1}^{T} \bar{\ell}_t\,(\theta_t) - \min_\theta \sum_{t=1}^{T} \bar{\ell}_t\,(\theta)\right] = \mathbb{E}\left[\sum_{t=1}^{T} \bar{\ell}_t\,(\theta_t)\right] = o(T)$$

Then a_t converges to a critical point of the reward $\eta(\theta^\star, \cdot)$

Learning Gradients With Model Extrapolation

$$\bar{\ell}_t(\theta) = \left(\eta(\theta, a_t) - \eta(\theta^\star, a_t)\right)^2 + \left(\eta(\theta, a_{t-1}) - \eta(\theta^\star, a_{t-1})\right)^2 + ||\nabla\eta(\theta, a_{t-1}) - \underbrace{\nabla\eta(\theta^\star, a_{t-1})}_{\text{not observed}}||_2^2$$

- $||\nabla\eta(\theta, a) - \nabla\eta(\theta^\star, a)||_2^2 = \mathbb{E}_u[\langle\nabla\eta(\theta, a) - \nabla\eta(\theta^\star, a), u\rangle^2]$

where $u \sim \mathcal{N}(0, I)$

- Directional gradient $\langle\nabla\eta(\theta^\star, a), u\rangle$ can be computed by two actions

$$\langle\nabla\eta(\theta^\star, a), u\rangle \approx \frac{\eta(\theta^\star, a + \alpha u) - \eta(\theta^\star, a)}{\alpha} \quad (\alpha \to 0)$$

- Similarly to Johnson–Lindenstrauss, it requires complexity(Θ) samples of u.
- Zero order optimization requires $\Omega(d)$ samples.

Algorithm and Theorem

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

- ViOL (**Vi**rtual Ascent with **O**nline Model **L**earner)

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 $a_t = \text{argmax}_a \, \mathbb{E}_{\theta_t}[\eta(\theta_t, a)]$

- Theorem (informal): Under Lipschitz assumptions on η, ViOL 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 $||w_2||_1, \|W_1\|_{\infty\to\infty}$
  - $R(W) \le \tilde{O}(1)$
  - sample complexity for local max = $\tilde{O}(1)$
  - Local maximum are global for input-concave neural nets

First-cut Extension to RL

RL	Bandit with Continous Actions
Dynamics T_θ	Model parameter θ
Policy π_ψ	Action a
Total return $\eta(T_\theta, \pi_\psi)$	Reward function $\eta(\theta, a)$

- Caveat: $\{\eta(T_\theta,\cdot): \theta \in \Theta\}$ has high complexity
- A result for stochastic policies (with many Lipschitz conditions)

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

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

- Summary: online learning T_θ implies predicting the curvature of η

Summary

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

- Check out our paper for more detail: https://arxiv.org/abs/2102.04168

Thank you for your attention ☺