Corruption-robust exploration in episodic RL

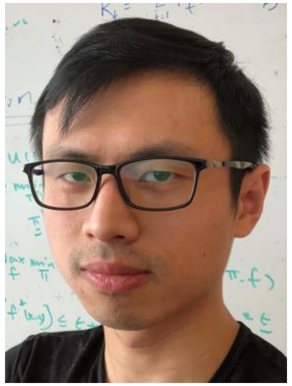
Alex Slivkins (Microsoft Research NYC)

joint work with

Thodoris Lykouris Max Simchowitz Wen Sun

MSR-NYC

Berkeley



Cornell

Prelude: from bandits to RL

	Bandits	RL
Global constraints	Limited-supply dynamic pricing, Bandits with knapsacks	Brantley, Dudik, Lykouris, Miryoosefi, Simchowitz, S., Sun: NeurIPS’20
Incentives	Incentivized exploration	Simchowitz & S., 2020
Lipschitz assumptions	Lipschitz bandits, adaptive discretization	Sinclair, Banerjee, Yu: Sigmetrics`20 Cao & Krishnamurthy: NeurIPS`20
Between IID & adversarial	adversarial corruptions Starting from [Lykouris et al., STOC`18]	This talk: RL with adversarial corruptions

PSA: Thodoris Lykouris will give a longer talk on this work on Nov 3 in *Virtual RL Theory Seminar*

Episodic RL with adversarial corruptions

- Fixed and unknown nominal MDP
  - known state space, action space;
  - randomized (& unknown) rewards & transitions
- K episodes of H steps each, $T = KH$ steps total.
- At each episode k: algorithm commits to a policy π_k,
  executes π_k in the MDP for H steps, observes state-actions-rewards trajectory.
  - policy maps histories to actions, can be randomized
  - bandit feedback: only for (current state, chosen action)
- $\text{Regret} = K \cdot rew(\pi^*) - \sum_k rew(\pi_k)$ w.r.t. best policy π^*
  - e.g., $\text{poly}(H) \cdot \sqrt{\#\text{states} \cdot \#\text{actions} \cdot T}$ (Azar et al.`17, Jin et al.`19)

C = #corrupted episodes not known in advance

Adversary corrupts the MDP

Goals: scale well with C, approx. state-of-art for $C = 0$

Our results

K episodes, H steps each, $T = KH$
S states, A actions
$C \geq 1$ (unknown) #corrupted episodes

Tabular RL: regret $C \cdot \mathrm{poly}(H) \cdot \sqrt{SAT}$

- $\sqrt{SAT}$ dependence is optimal, even for IID

First non-trivial guarantees for RL with non-IID transitions & bandit feedback

- Also: first computationally efficient guarantees for any feedback model

Linear RL: regret $\mathrm{poly}(H) \left(C\sqrt{(d^3+dA) \cdot T} + C^2\sqrt{dT} \right)$

no dependence on S

expected rewards and transition probs are linear in (known) d-dim feature vectors

- optimal dependence on T, state-of-art dependence on d, even for IID

Transformation: (some) algorithms for IID environment → corruption-robust algorithms

Provable guarantees known only for Tabular and Linear variants of episodic RL

e.g., well-defined for deep RL

Prior work

Bandits: stochastic vs adversarial

- Classic papers: UCB1 and EXP3
- Best of both worlds
  Bubeck & S. `12; Seldin & S. `14; Auer & Chiang `16; Seldin & Lugosi `17; Wei & Luo `18
  - intermediate regimes
    starting from Seldin & S. `14
- Adversarial corruptions
  Lykouris-Mirrokni- Paes Leme `18
  improved regret bounds
  Gupta-Koren-Talwar `19, Zimmert & Seldin `19
  many extensions
  LLS19, CKW19, BJS20, KLPS20, AAKLM20

Episodic RL

- Stochastic: optimistic value iteration
  starting from Jaksch-Ortner-Auer'10
  worst-case optimal regret rates
  Azar et al.'17, Dann et al. `17
  instance-dependent regret rates
  Zanette &Brunskill `19, Simchowitz &Jamieson `19
- Adversarial rewards: full feedback
  transition probabilities known (Even-Dar+ `10),
  unknown (Rosenberg+ `19),
  or adversarial (Abbasi-Yadkori+ `13)

  ... bandit feedback

  trans. probs known (Neu+ `10) or not (Jin+ `19)

Prior work: how to resolve uncertainty?

Active sets

update active set = {plausibly optimal actions}, choose uniformly from this set

- works for bandits
- underlies the corruption-robust algorithm in Lykouris et al. `18

Fails for RL: "any reasonable version" suffers regret $\min(K, A^H)$ on a "combination lock instance"

K episodes of H steps each, A actions

Optimism

pick alternative with best optimistic estimate: most favorable estimate consistent with data

- works for RL: optimistic value iteration

Bellman updates with optimistic estimates

- vast majority of Episodic RL algorithms
  except Jin et al.'19 and Russo'19

Fails for corruptions, even for bandits

Suffices to corrupt $O(\log T)$ rounds: reward 0 each time algorithm picks best arm

Optimistic Value Iteration with active sets

For each step h from H down to 1

- update Q_h using V_{h+1}, rewards & transition probs
  - UCB via optimistic reward estimates
  - LCB via pessimistic reward estimates
  - use both "local" and "global" data
- update π^* using Q_h
  - use UCBs
  - restrict to active sets
- update V_h using Q_h
  - compute UCBs and LCBs
  - recompute active sets (of actions)

Starting at state x, action a, step h
$Q_h(x,a)$: value if continued optimally
$V_h(x) = \max_a Q_h(x,a)$
$\pi^*_h(x) = \operatorname{argmax}_a Q_h(x,a)$

Value iteration (VI)

Optimistic VI

Optimistic VI with active sets

"Base Algorithm"

Full algorithm: Base Learners

Each Base Learner (BL) ℓ runs a separate instance of Base Algorithm

- robust against a given level of corruption $C = 2^{\ell}$
- "local data": data assigned to this BL

  "global data": union of data from all BLs

Need "global data" because different BLs may traverse different trajectories across state space

At each step of each episode: randomly switch to a more robust BL (larger ℓ)

- carefully chosen, data-independent probs
- sufficient prob of switching to a more robust BL for the rest of the episode
- episode's data assigned to the most robust BL used in this episode

More robust BL provide supervision for less robust BL via "global data"

Analysis

General framework to analyze Base Learners with active sets

- beyond UCB selection (or uniform selection)

Bellman errors $\hat{Q}_h(x,a) - \left(r^*(x,a) + \hat{V}_{h+1} \cdot\, p^*(x,a)\right)$

states x, actions a, steps h

Error in Bellman update $\hat{Q}_h(x,a) - \left(\hat{r}(x,a) + \hat{V}_{h+1} \cdot \hat{p}(x,a)\right)$

Decomposition: express regret in terms of Bellman Errors

Compare policy π with UCB policy

policy π': what if we switch to UCB after step h

Visitation ratio $\max_{\text{steps } h<\tau} \max_{x,a} \frac{\mathcal{M}'(x,a)}{\mathcal{M}(x,a)}$

$\mathcal{M}, \mathcal{M}'$occupancy measures for π, π' at step $\tau > h$

$\Pr[(x_\tau, a_\tau) = (x,a)]$

Zoom out

RL challenge: inject enough exploration into a complex behavior

- optimism = best available hammer

e.g., one that ensures corruption-robustness

Design principle: randomly switch to a (more) reliable version of optimism

- general framework for analysis
- proof of concept: a new algorithm for "stochastic" episodic RL,
  start with active sets & uniform exploration, inject optimism => optimal regret
- this machinery could be applicable to other domains

e.g., more robust Base Learner

Extensions & Open Questions

K episodes, H steps each, $T = KH$
S states, A actions
$C \geq 1$ (unknown) #corrupted episodes

Instance-dependent regret bounds: $C \cdot \text{poly}(H) \cdot \frac{AS}{\text{MinGap}} \cdot \log(SAT)$

$\text{MinGap} = \min_{\text{states } x, \text{ actions } a} \text{Gap}(x, a)$

Improves to $AS + \frac{1}{\text{MinGap}}$ if all but few actions are bad

- constant C: matches state-of-art for the IID case (Simchowitz & Jamieson`19)

Open Q: mitigate the linear dependence on C

- make it additive rather than multiplicative?
- non-trivial guarantees for $C > \sqrt{T}$?
- $o(C)$ dependence, preferably $\sqrt{C}$
  ... if we only count regret for non-corrupted rounds?

Yes for bandits
Gupta-Koren-Talwar '19;
Zimmert & Seldin `20

PSA: Thodoris Lykouris will give a longer talk on this work on Nov 3 in *Virtual RL Theory Seminar*

Link to the paper: https://arxiv.org/abs/1911.08689 .