Non-Stationary Reinforcement Learning

Ruihao Zhu

MIT IDSS

Joint work with Wang Chi Cheung (NUS) and David Simchi-Levi (MIT)

Epidemic Control

A DM iteratively:

1. Pick a measure to contain the virus.
2. See the corresponding outcome.

Goal: Minimize the total infected cases.

Epidemic Control

A DM iteratively:

1. Pick a measure to contain the virus.
2. See the corresponding outcome.

Goal: Minimize the total infected cases.

Challenges:

- **Uncertainty:** effectiveness of each measure is unknown.
- **Bandit feedback:** no feedback for un-chosen measures.
- **Non-stationarity:** virus might mutate throughout.

Epidemic Control

The DM's action could have **long-term impact**.

- Quarantine lockdown stem the spread of virus to elsewhere, but also delayed key supplies from getting in.

Search **Bloomberg**

Prognosis

China Sacrifices a Province to Save the World From Coronavirus

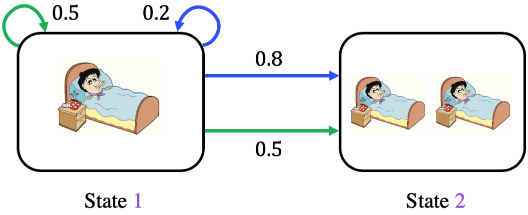
Bloomberg News

February 5, 2020, 11:01 AM EST

- Hubei province has seen 97% of all deaths from the virus
- Quarantine lockdown delayed key supplies from getting in

Model

Model epidemic control by a Markov decision process (MDP) *(Nowzari et al. 15, Kiss et al. 17)*.



For each time step $t = 1, \ldots, T$,

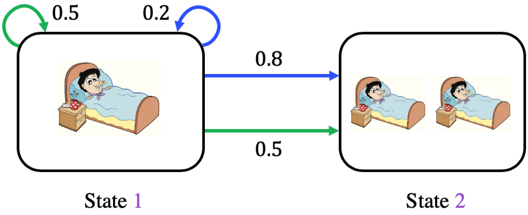
- Observe the current state $s_t = \{1, 2\}$, and receive a reward. For example

$$r(1) = 1 \text{ and } r(2) = 0.$$

- Pick an action $a_t \in \{B, G\}$, and transition to the next state $s_{t+1} \sim p_t(\cdot | s_t, a_t)$ (unknown).

Model

Model epidemic control by a Markov decision process (MDP) *(Nowzari et al. 15, Kiss et al. 17)*.



For each time step $t = 1, \ldots, T$,

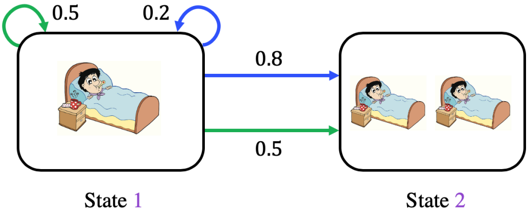
- Observe the current state $s_t = \{1, 2\}$, and receive a reward. For example

$$r(1) = 1 \text{ and } r(2) = 0.$$

- Pick an action $a_t \in \{B, G\}$, and transition to the next state $s_{t+1} \sim p_t(\cdot | s_t, a_t)$ (unknown).

Model

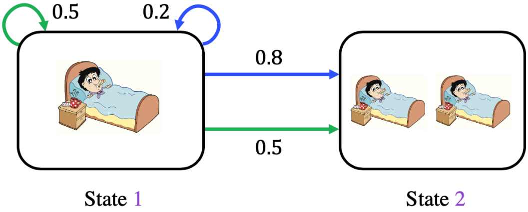
Model epidemic control by a Markov decision process (MDP) *(Nowzari et al. 15, Kiss et al. 17)*.



For each time step $t = 1, \ldots, T$,

- Observe the current state $s_t = \{1, 2\}$, and receive a reward. For example
$$r(1) = 1 \text{ and } r(2) = 0.$$
- Pick an action $a_t \in \{B, G\}$, and transition to the next state $s_{t+1} \sim p_t(\cdot | s_t, a_t)$ (unknown).

Model cont'd



- **Task:** Design a reward-maximizing policy π.

  For every time step t : $\pi_t : \{1,2\} \to \{B, G\}$

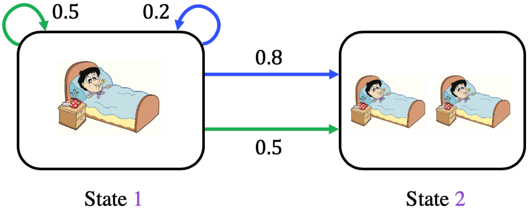
- **Dynamic regret** *(Besbes et al. 15)*:

$$\text{dym-reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} r(s_t(\underbrace{\pi_*}_{\text{knows } p_t\text{'s}}))\right] - \mathbb{E}\left[\sum_{t=1}^{T} r(s_t(\pi))\right].$$

- **Variation budget:**

$$\|p_1 - p_2\| + \|p_2 - p_3\| + \ldots + \|p_{T-1} - p_T\| \leq B_p.$$

Model cont'd



- **Task:** Design a reward-maximizing policy π.

  For every time step t : $\quad \pi_t : \{1, 2\} \to \{B, G\}$

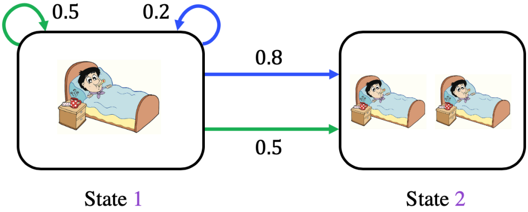
- **Dynamic regret** *(Besbes et al. 15)*:

$$\text{dym-reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} r(s_t(\underbrace{\pi_*}_{\text{knows } p_t\text{'s}}))\right] - \mathbb{E}\left[\sum_{t=1}^{T} r(s_t(\pi))\right].$$

- **Variation budget:**

$$\|p_1 - p_2\| + \|p_2 - p_3\| + \ldots + \|p_{T-1} - p_T\| \leq B_p.$$

Model cont'd



- **Task:** Design a reward-maximizing policy π.

  For every time step t : $\pi_t : \{1, 2\} \to \{B, G\}$

- **Dynamic regret** *(Besbes et al. 15)*:

$$\text{dym-reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} r(s_t(\underbrace{\pi_*}_{\text{knows } p_t\text{'s}}))\right] - \mathbb{E}\left[\sum_{t=1}^{T} r(s_t(\pi))\right].$$

- **Variation budget:**

$$\|p_1 - p_2\| + \|p_2 - p_3\| + \ldots + \|p_{T-1} - p_T\| \leq B_p.$$

Diameter of a MDP cont'd

- If the DM leaves state 1, she has to come back to state 1 to collect samples.

Diameter of a MDP cont'd

- If the DM leaves state 1, she has to come back to state 1 to collect samples.
- The longer it takes to commute between states, the harder the learning process.

Diameter of a MDP cont'd

- If the DM leaves state 1, she has to come back to state 1 to collect samples.
- The longer it takes to commute between states, the harder the learning process.

Definition ((*Jaksch et al. 10*) Informal)

Diameter $= \max\{\mathbb{E}[\text{min. time}(1 \to 2)], \mathbb{E}[\text{min. time}(2 \to 1)]\}$

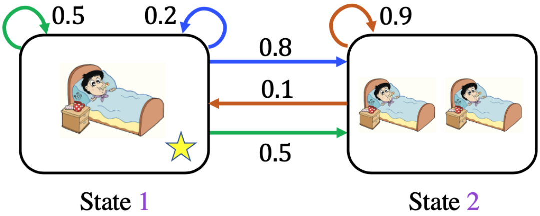
Diameter of a MDP cont'd

- If the DM leaves state 1, she has to come back to state 1 to collect samples.
- The longer it takes to commute between states, the harder the learning process.

Definition ((*Jaksch et al. 10*) Informal)

Diameter $= \max\{\mathbb{E}[\text{min. time}(1 \to 2)], \mathbb{E}[\text{min. time}(2 \to 1)]\}$

Example. Diameter $= \max\{1/0.8, 1/0.1\} = 10.$



Existing Works

	Stationary	Non-stationary
Multi-armed bandit	OFU*	Forgetting + OFU†
Reinforcement learning	OFU‡	? (Forgetting + OFU)

* *Auer et al. 03*

†*Besbes et al. 14, Cheung et al. 19*

‡*Jaksch et al. 10, Agrawal and Jia 20*

UCB for Stationary RL

1. Suppose at time t,

$$N_t(1, B) = 10: \quad 5 \times (1, B) \to 1, \quad 5 \times (1, B) \to 2$$

UCB for Stationary RL

1. Suppose at time t,

$$N_t(1, B) = 10: \quad 5 \times (1, B) \to 1, \quad 5 \times (1, B) \to 2$$

$$N_t(2, B) = 10: \quad 5 \times (2, B) \to 1, \quad 5 \times (2, B) \to 2$$

UCB for Stationary RL

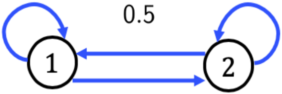
1. Suppose at time t,

$$N_t(1, B) = 10: \quad 5 \times (1, B) \to 1, \quad 5 \times (1, B) \to 2$$

$$N_t(2, B) = 10: \quad 5 \times (2, B) \to 1, \quad 5 \times (2, B) \to 2$$

Empirical state transition distribution:

$\hat{p}_t$



UCB for Stationary RL

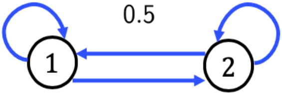
1. Suppose at time t,

$$N_t(1, B) = 10 : \quad 5 \times (1, B) \to 1, \quad 5 \times (1, B) \to 2$$

$$N_t(2, B) = 10 : \quad 5 \times (2, B) \to 1, \quad 5 \times (2, B) \to 2$$

   Empirical state transition distribution:

   $\hat{p}_t$



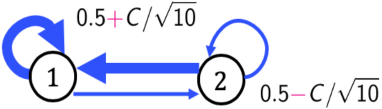
2. Confidence intervals:

$$\|\hat{p}_t(\cdot|1, B) - p(\cdot|1, B)\| \leq c_t(1, B) := C/\sqrt{10}$$

$$\|\hat{p}_t(\cdot|2, B) - p(\cdot|2, B)\| \leq c_t(2, B) := C/\sqrt{10}$$

UCB for Stationary RL

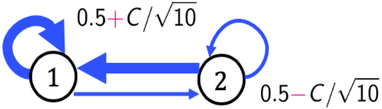
3. UCB of reward: find the $\mathring{p}$ that maximizes Pr(visiting state 1) within the confidence interval.



4. Execute the optimal policy w.r.t. the UCB until some termination criteria are met.

UCB for Stationary RL

3. UCB of reward: find the $\mathring{p}$ that maximizes Pr(visiting state 1) within the confidence interval.

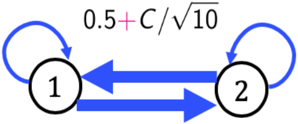


4. Execute the optimal policy w.r.t. the UCB until some termination criteria are met.

UCB for RL cont'd

Regret analysis:

- LCB of diameter: find the $\mathring{p}$ that maximizes Pr(commuting) within the confidence interval.



- Regret $\propto$ LCB $\times \left(\sum_{(s,a)} c_t(s,a)\right)$.
- Under stationarity, LCB of diameter $\leq$ Diameter(p).

Theorem

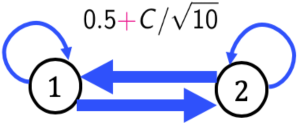
Denote $D :=$ Diameter(p), the regret of the UCB algorithm is $O(D\sqrt{T})$.

- **Summary:** UCB of reward + LCB of diameter $\Rightarrow$ low regret.

UCB for RL cont'd

Regret analysis:

- LCB of diameter: find the $\mathring{p}$ that maximizes Pr(commuting) within the confidence interval.



- Regret $\propto$ LCB $\times \left(\sum_{(s,a)} c_t(s,a)\right)$.
- Under stationarity, LCB of diameter $\leq$ Diameter(p).

Theorem

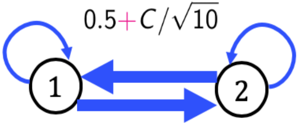
Denote $D :=$ Diameter(p), the regret of the UCB algorithm is $O(D\sqrt{T})$.

- **Summary:** UCB of reward + LCB of diameter $\Rightarrow$ low regret.

UCB for RL cont'd

Regret analysis:

- LCB of diameter: find the $\mathring{p}$ that maximizes Pr(commuting) within the confidence interval.



- Regret $\propto$ LCB $\times$ $\left(\sum_{(s,a)} c_t(s,a)\right)$.
- Under stationarity, LCB of diameter $\leq$ Diameter(p).

Theorem

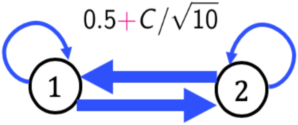
Denote $D :=$ Diameter(p), the regret of the UCB algorithm is $O(D\sqrt{T})$.

- **Summary:** UCB of reward + LCB of diameter $\Rightarrow$ low regret.

UCB for RL cont'd

Regret analysis:

- LCB of diameter: find the $\mathring{p}$ that maximizes Pr(commuting) within the confidence interval.



- Regret $\propto$ LCB $\times \left(\sum_{(s,a)} c_t(s,a)\right)$.
- Under stationarity, LCB of diameter $\leq$ Diameter(p).

Theorem

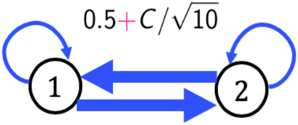
Denote $D :=$ Diameter(p), the regret of the UCB algorithm is $O(D\sqrt{T})$.

- **Summary:** UCB of reward + LCB of diameter $\Rightarrow$ low regret.

UCB for RL cont'd

Regret analysis:

- LCB of diameter: find the $\mathring{p}$ that maximizes Pr(commuting) within the confidence interval.



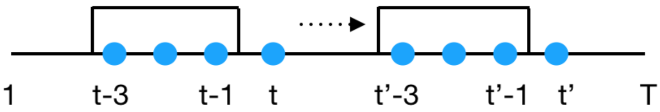
- Regret $\propto$ LCB $\times \left(\sum_{(s,a)} c_t(s,a)\right)$.
- Under stationarity, LCB of diameter $\leq$ Diameter(p).

Theorem

Denote $D :=$ Diameter(p), the regret of the UCB algorithm is $O(D\sqrt{T})$.

- **Summary:** UCB of reward + LCB of diameter $\Rightarrow$ low regret.

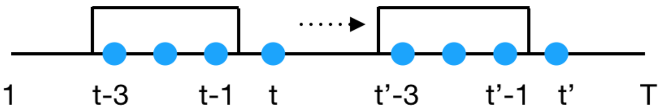
SWUCB for RL



According to *(Cheung et al. 19)*:

- **SWUCB for RL:** UCB for RL with W most recent samples.

SWUCB for RL



According to *(Cheung et al. 19)*:

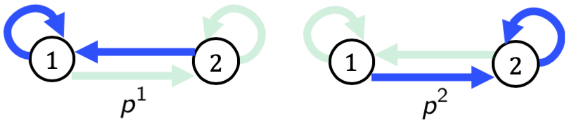
- **SWUCB for RL:** UCB for RL with W most recent samples.
- **The perils of drift:** Under non-stationarity,

$$\text{LCB of diameter} \gg \text{Diameter}(p_s)$$

for all $s \in [T]$.

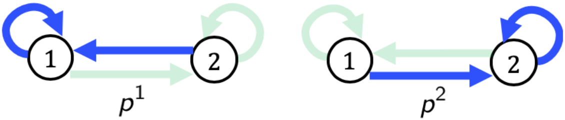
Perils of Non-Stationarity in RL

Non-stationarity: The DM faces time-varying environment.

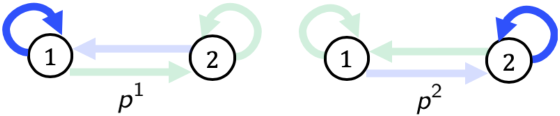


Perils of Non-Stationarity in RL

Non-stationarity: The DM faces time-varying environment.

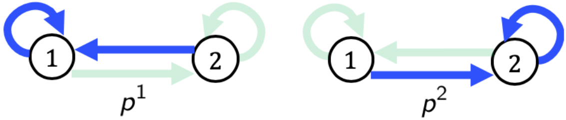


Bandit feedback: The DM is not seeing everything.

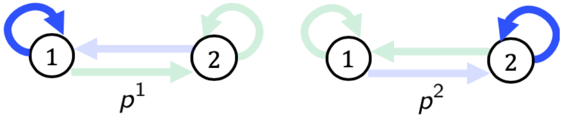


Perils of Non-Stationarity in RL

Non-stationarity: The DM faces time-varying environment.



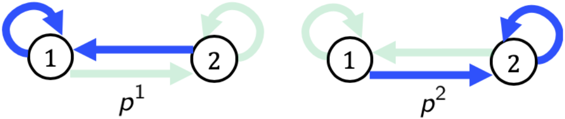
Bandit feedback: The DM is not seeing everything.



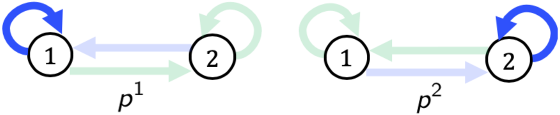
Collected data: $\{(1, B) \to 1,\ (2, B) \to 2\}$

Perils of Non-Stationarity in RL

Non-stationarity: The DM faces time-varying environment.



Bandit feedback: The DM is not seeing everything.

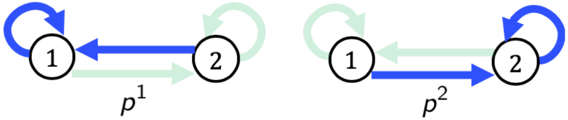


Collected data: $\{(1, B) \to 1,\ (2, B) \to 2\}$

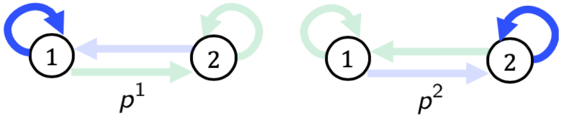
Empirical state transition $\hat{p}_t$:

Perils of Non-Stationarity in RL

Non-stationarity: The DM faces time-varying environment.



Bandit feedback: The DM is not seeing everything.



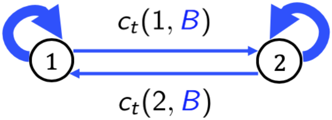
Collected data: $\{(1, B) \to 1,\ (2, B) \to 2\}$

Empirical state transition $\hat{p}_t$:

Diameter explodes!

Perils of Non-Stationarity in RL

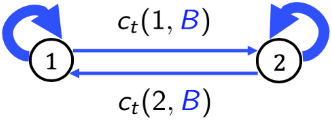
But let's still check the "LCB" of diameter:



- For a window size W, $c_t(1, B)$ and $c_t(2, B)$ can be as small as $\Theta(1/\sqrt{W})$ *(Cheung et al. 20)*.
- Hence, the "LCB" of diameter can be as large as $\Theta(\sqrt{W})$.
- **Recall:** diameters of p^1 and p^2 are $1 \ll \Theta(\sqrt{W})$.
- The "LCB" is no longer a valid LCB under non-stationarity.
- SWUCB incurs $\Theta(T)$ dynamic regret.

Perils of Non-Stationarity in RL

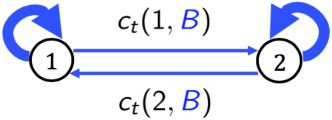
But let's still check the "LCB" of diameter:



- For a window size W, $c_t(1, B)$ and $c_t(2, B)$ can be as small as $\Theta(1/\sqrt{W})$ *(Cheung et al. 20)*.
- Hence, the "LCB" of diameter can be as large as $\Theta(\sqrt{W})$.
- **Recall:** diameters of p^1 and p^2 are $1 \ll \Theta(\sqrt{W})$.
- The "LCB" is no longer a valid LCB under non-stationarity.
- SWUCB incurs $\Theta(T)$ dynamic regret.

Perils of Non-Stationarity in RL

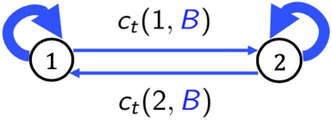
But let's still check the "LCB" of diameter:



- For a window size W, $c_t(1, B)$ and $c_t(2, B)$ can be as small as $\Theta(1/\sqrt{W})$ *(Cheung et al. 20)*.
- Hence, the "LCB" of diameter can be as large as $\Theta(\sqrt{W})$.
- **Recall:** diameters of p^1 and p^2 are $1 \ll \Theta(\sqrt{W})$.
- The "LCB" is no longer a valid LCB under non-stationarity.
- SWUCB incurs $\Theta(T)$ dynamic regret.

Perils of Non-Stationarity in RL

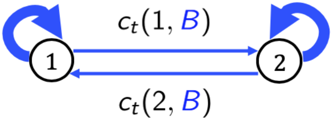
But let's still check the "LCB" of diameter:



- For a window size W, $c_t(1, B)$ and $c_t(2, B)$ can be as small as $\Theta(1/\sqrt{W})$ *(Cheung et al. 20)*.
- Hence, the "LCB" of diameter can be as large as $\Theta(\sqrt{W})$.
- **Recall:** diameters of p^1 and p^2 are $1 \ll \Theta(\sqrt{W})$.
- The "LCB" is no longer a valid LCB under non-stationarity.
- SWUCB incurs $\Theta(T)$ dynamic regret.

Perils of Non-Stationarity in RL

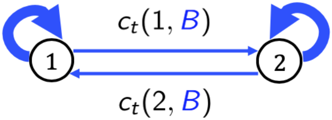
But let's still check the "LCB" of diameter:



- For a window size W, $c_t(1, B)$ and $c_t(2, B)$ can be as small as $\Theta(1/\sqrt{W})$ *(Cheung et al. 20)*.
- Hence, the "LCB" of diameter can be as large as $\Theta(\sqrt{W})$.
- **Recall:** diameters of p^1 and p^2 are $1 \ll \Theta(\sqrt{W})$.
- The "LCB" is no longer a valid LCB under non-stationarity.
- SWUCB incurs $\Theta(T)$ dynamic regret.

Perils of Non-Stationarity in RL

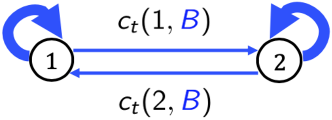
But let's still check the "LCB" of diameter:



- For a window size W, $c_t(1, B)$ and $c_t(2, B)$ can be as small as $\Theta(1/\sqrt{W})$ *(Cheung et al. 20)*.
- Hence, the "LCB" of diameter can be as large as $\Theta(\sqrt{W})$.
- **Recall:** diameters of p^1 and p^2 are $1 \ll \Theta(\sqrt{W})$.
- The "LCB" is no longer a valid LCB under non-stationarity.
- SWUCB incurs $\Theta(T)$ dynamic regret.

Perils of Non-Stationarity in RL

But let's still check the "LCB" of diameter:



- For a window size W, $c_t(1, B)$ and $c_t(2, B)$ can be as small as $\Theta(1/\sqrt{W})$ *(Cheung et al. 20)*.
- Hence, the "LCB" of diameter can be as large as $\Theta(\sqrt{W})$.
- **Recall:** diameters of p^1 and p^2 are $1 \ll \Theta(\sqrt{W})$.
- The "LCB" is no longer a valid LCB under non-stationarity.
- SWUCB incurs $\Theta(T)$ dynamic regret.

Confidence Widening

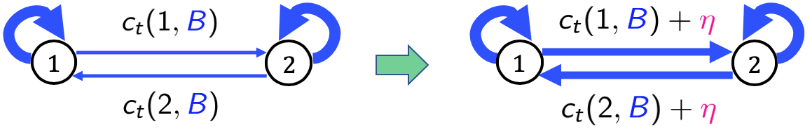
- This caveat stems from the estimation.

Confidence Widening

- This caveat stems from the estimation.
- We can refine the design principle of UCB.

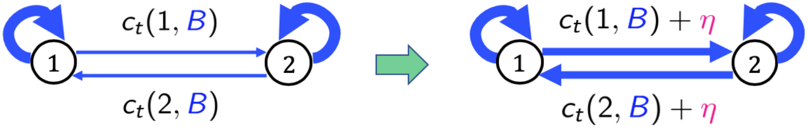
Confidence Widening

- This caveat stems from the estimation.
- We can refine the design principle of UCB.
- **Confidence widening:** increase each confidence interval by η.

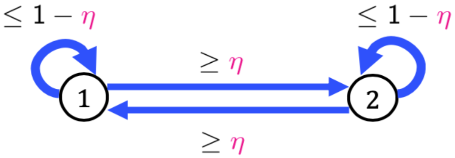


Confidence Widening

- This caveat stems from the estimation.
- We can refine the design principle of UCB.
- **Confidence widening:** increase each confidence interval by η.

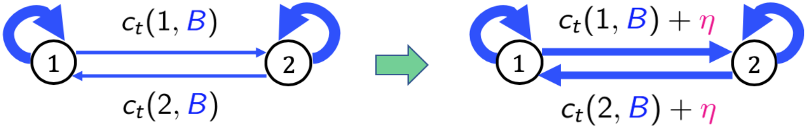


- $c_t \geq 0 \implies \Pr(\text{commuting}) \geq \eta$

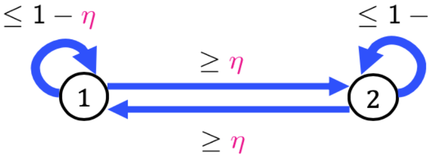


Confidence Widening

- This caveat stems from the estimation.
- We can refine the design principle of UCB.
- **Confidence widening:** increase each confidence interval by η.



- $c_t \geq 0 \implies \Pr(\text{commuting}) \geq \eta$



- New "LCB" $\leq 1/\eta$.

Confidence Widening

Recall: Regret $\propto$ LCB $\times[\sum_{(s,a)}(c_t(s,a)+\eta)]$.

Confidence Widening

Recall: Regret $\propto$ LCB $\times[\sum_{(s,a)}(c_t(s,a)+\eta)]$.

Confidence Widening

Recall: Regret $\propto$ LCB $\times[\sum_{(s,a)}(c_t(s,a)+\eta)]$.

- If $1/\eta \leq \text{Diameter}(p_t)$, then LCB $\leq 1/\eta \leq \text{Diameter}(p_t)$.

Confidence Widening

Recall: Regret $\propto$ LCB $\times[\sum_{(s,a)}(c_t(s,a)+\eta)]$.

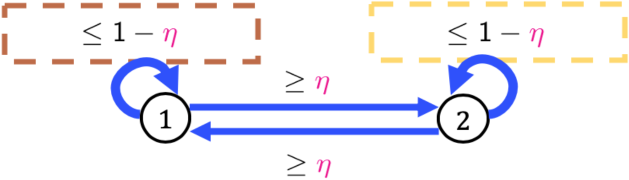
- If $1/\eta \leq$ Diameter(p_t), then LCB $\leq 1/\eta \leq$ Diameter(p_t).
- If $1/\eta \geq$ Diameter(p_t), then Pr(commuting) $\geq \eta$ for p_t :

Confidence Widening

Recall: Regret $\propto$ LCB $\times[\sum_{(s,a)}(c_t(s,a)+\eta)]$.

- If $1/\eta \leq$ Diameter(p_t), then LCB $\leq 1/\eta \leq$ Diameter(p_t).
- If $1/\eta \geq$ Diameter(p_t), then Pr(commuting) $\geq \eta$ for p_t :

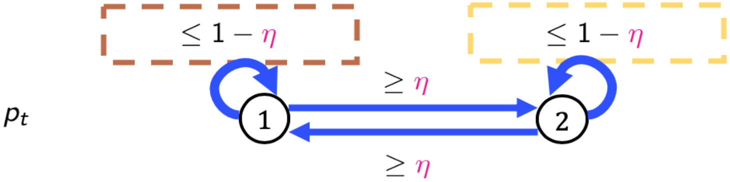
p_t



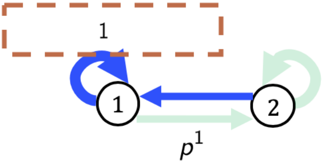
Confidence Widening

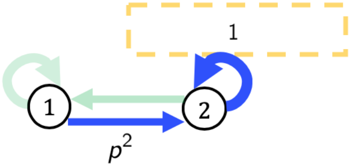
Recall: Regret $\propto$ LCB $\times[\sum_{(s,a)}(c_t(s,a)+\eta)]$.

- If $1/\eta \leq \text{Diameter}(p_t)$, then LCB $\leq 1/\eta \leq \text{Diameter}(p_t)$.
- If $1/\eta \geq \text{Diameter}(p_t)$, then Pr(commuting) $\geq \eta$ for p_t :



- Compare to p^1 and p^2 : a η variation is detected!





The Blessing of More Optimism

Confidence widening ensures either we enjoy reasonable upper bound for LCB or we consume η of variation budget.

The Blessing of More Optimism

Confidence widening ensures either we enjoy reasonable upper bound for LCB or we consume η of variation budget.

Theorem

If we choose the optimal W and η w.r.t. B_p, the dynamic regret bound for the SWUCB-CW algorithm is

$$\tilde{O}\left(D_{max} B_p^{\frac{1}{4}} T^{\frac{3}{4}}\right).$$

Conclusion

	Stationary	Non-stationary
MAB	OFU	OFU + Forgetting
RL	OFU	Extra optimism + Forgetting

- An unfavorable "phase transition" from MAB (1 state) to RL ($\geq$ 2 states) for SWUCB.
- **Blessing of more optimism:** Provably low dynamic regret for non-stationary RL.
- **Parameter-free:** Bandit-over-reinforcement learning *(Cheung et al. 20)*.

Thank You!

rzhu@mit.edu

isecwc@nus.edu.sg, dslevi@mit.edu