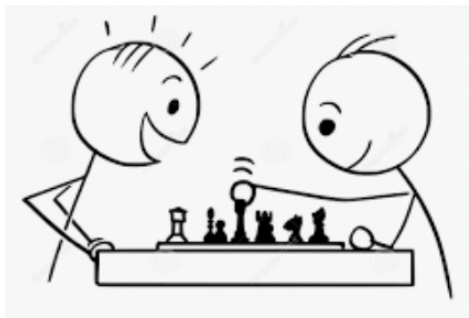
Tracking Significant Changes in Bandits

Samory Kpotufe
Columbia University, Statistics

Based on works with **Joe Suk***

Long Term Motivation:

Sequential decisions under noisy, partial feedback


(Contextual) Bandits, RL, ...

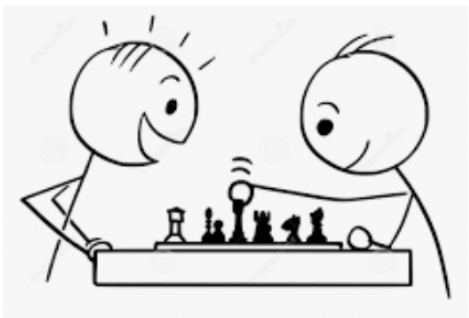
We may learn good policies if the environment remains consistent

...

However, environmental changes are frequent in practice ☹

Long Term Motivation:

Sequential decisions under noisy, partial feedback


(Contextual) Bandits, RL, ...

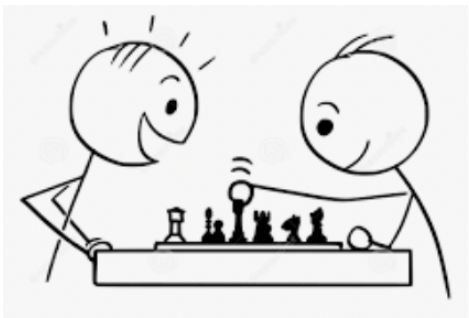
We may learn good policies if the environment remains consistent

...

However, environmental changes are frequent in practice ☹

Long Term Motivation:

Sequential decisions under noisy, partial feedback

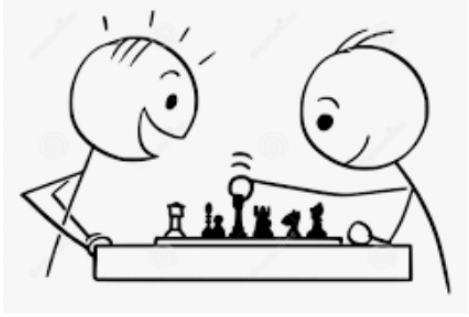

(Contextual) Bandits, RL, ...

We may learn good policies if the environment remains consistent ...

However, environmental changes are frequent in practice ☹

Long Term Motivation:

Sequential decisions under noisy, partial feedback


(Contextual) Bandits, RL, ...

We may learn good policies if the environment remains consistent

...

However, environmental changes are frequent in practice ☹

Environmental changes are frequent in practice ☹

Many solutions so far (in ML):

Detect changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle

Mostly Open Questions:

- Which changes are actually **severe**?
- Can we **adapt** to such severe changes?

We'll focus for now on (contextual) bandits ...

Environmental changes are frequent in practice ☹

Many solutions so far (in ML):

Detect changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle

Mostly Open Questions:

- Which changes are actually **severe**?
- Can we **adapt** to such severe changes?

We'll focus for now on (contextual) bandits ...

Environmental changes are frequent in practice ☹

Many solutions so far (in ML):

Detect changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle

Mostly Open Questions:

- Which changes are actually **severe**?
- Can we **adapt** to such severe changes?

We'll focus for now on (contextual) bandits ...

Environmental changes are frequent in practice ☹

Many solutions so far (in ML):

Detect changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle

Mostly Open Questions:

- Which changes are actually **severe**?
- Can we **adapt** to such severe changes?

We'll focus for now on (contextual) bandits ...

Environmental changes are frequent in practice ☹

Many solutions so far (in ML):

Detect changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle

Mostly Open Questions:

- Which changes are actually **severe**?
- Can we **adapt** to such severe changes?

We'll focus for now on (contextual) bandits ...

Environmental changes are frequent in practice ☹

Many solutions so far (in ML):

Detect changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle

Mostly Open Questions:

- Which changes are actually **severe**?
- Can we **adapt** to such severe changes?

We'll focus for now on (contextual) bandits ...

Environmental changes are frequent in practice ☹

Many solutions so far (in ML):

Detect changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle

Mostly Open Questions:

- Which changes are actually **severe**?
- Can we **adapt** to such severe changes?

We'll focus for now on (contextual) bandits ...

(Contextual) Bandits:

At time t:

- Observe context X_t (Patient profile)
- Pick action $a \in [K]$ (Treatment)
- Observe reward $Y_t(a) \mid X_t$ (Outcome)

Contexts and or Reward distributions may change frequently ...

When are these changes severe and can we adapt?

Partial Answer: for now, consider Context vs Rewards separately ...

(Contextual) Bandits:

At time t:

- Observe context X_t (Patient profile)
- Pick action $a \in [K]$ (Treatment)
- Observe reward $Y_t(a) \mid X_t$ (Outcome)

Contexts and or Reward distributions may change frequently ...

When are these changes severe and can we adapt?

Partial Answer: for now, consider Context vs Rewards separately ...

(Contextual) Bandits:

At time t:

- Observe context X_t (Patient profile)
- Pick action $a \in [K]$ (Treatment)
- Observe reward $Y_t(a) \mid X_t$ (Outcome)

Contexts and or Reward distributions may change frequently ...

When are these changes severe and can we adapt?

Partial Answer: for now, consider Context vs Rewards separately ...

(Contextual) Bandits:

At time t:

- Observe context X_t (Patient profile)
- Pick action $a \in [K]$ (Treatment)
- Observe reward $Y_t(a) \mid X_t$ (Outcome)

Contexts and or Reward distributions may change frequently ...

When are these changes severe and can we adapt?

Partial Answer: for now, consider Context vs Rewards separately ...

(Contextual) Bandits:

At time t:

- Observe context X_t (Patient profile)
- Pick action $a \in [K]$ (Treatment)
- Observe reward $Y_t(a) \mid X_t$ (Outcome)

Contexts and or Reward distributions may change frequently ...

When are these changes severe and can we adapt?

Partial Answer: for now, consider Context vs Rewards separately ...

Outline:

When are these changes severe and can we adapt?

- **Changes in Reward Y_t distribution**
  In fact, we'll simply consider regular bandits (COLT 2022)
- Changes in Context X_t distribution
  as in Covariate Shift (ALT 2021)
- Bringing it all together ... (ongoing work)

Outline:

When are these changes severe and can we adapt?

- **Changes in Reward Y_t distribution**
  In fact, we'll simply consider regular bandits (COLT 2022)
- Changes in Context X_t distribution
  as in Covariate Shift (ALT 2021)
- Bringing it all together ... (ongoing work)

Outline:

When are these changes severe and can we adapt?

- **Changes in Reward Y_t distribution**
  In fact, we'll simply consider regular bandits (COLT 2022)
- Changes in Context X_t distribution
  as in Covariate Shift (ALT 2021)
- Bringing it all together ... (ongoing work)

Outline:

When are these changes severe and can we adapt?

- **Changes in Reward Y_t distribution**
  In fact, we'll simply consider regular bandits (COLT 2022)
- Changes in Context X_t distribution
  as in Covariate Shift (ALT 2021)
- Bringing it all together ... (ongoing work)

Outline:

When are these changes severe and can we adapt?

- **Changes in Reward Y_t distribution**
  In fact, we'll simply consider regular bandits (COLT 2022)
- Changes in Context X_t distribution
  as in Covariate Shift (ALT 2021)
- Bringing it all together ... (ongoing work)

Changes in Reward Y_t distribution

Non-Stationary Bandit

- **At time t**: select $a \in [K]$, observe $Y_t(a)$, with mean $\mu_t(a)$.
- **Dynamic Regret**: $\mathrm{R_T} \doteq \sum_{t=1}^{T} \underbrace{\mu_t^\star - \mu_t(a_t)}_{\delta_t(a_t)}$.

What was known:

Parameters	Best Oracle Rate	Adaptive Rates
L changes in $\mu_t(a)$	$\sqrt{LT}$ [Garivier, Moulines. 11]	**Yes** [Auer et al. 19]
S best-arm switches	$\sqrt{ST} \ll \sqrt{LT}$ [Auer. 02]	**OPEN**
Total-Variation V	$V^{1/3}T^{2/3}$ [Besbes et al. 14]	**Yes** [Chen et al. 19]

What we show:

A much weaker notion of change admits adaptivity ...

Changes in Reward Y_t distribution

Non-Stationary Bandit

- **At time t**: select $a \in [K]$, observe $Y_t(a)$, with mean $\mu_t(a)$.
- **Dynamic Regret**: $\mathbf{R_T} \doteq \sum_{t=1}^{T} \underbrace{\mu_t^\star - \mu_t(a_t)}_{\delta_t(a_t)}$.

What was known:

Parameters	Best Oracle Rate	Adaptive Rates
L changes in $\mu_t(a)$	$\sqrt{LT}$ [Garivier, Moulines. 11]	**Yes** [Auer et al. 19]
S best-arm switches	$\sqrt{ST} \ll \sqrt{LT}$ [Auer. 02]	**OPEN**
Total-Variation V	$V^{1/3}T^{2/3}$ [Besbes et al. 14]	**Yes** [Chen et al. 19]

What we show:

A much weaker notion of change admits adaptivity ...

Changes in Reward Y_t distribution

Non-Stationary Bandit

- **At time t**: select $a \in [K]$, observe $Y_t(a)$, with mean $\mu_t(a)$.
- **Dynamic Regret**: $\mathbf{R_T} \doteq \sum_{t=1}^{T} \underbrace{\mu_t^\star - \mu_t(a_t)}_{\delta_t(a_t)}$.

What was known:

Parameters	Best Oracle Rate	Adaptive Rates
L changes in $\mu_t(a)$	$\sqrt{LT}$ [Garivier, Moulines. 11]	**Yes** [Auer et al. 19]
S best-arm switches	$\sqrt{ST} \ll \sqrt{LT}$ [Auer. 02]	**OPEN**
Total-Variation V	$V^{1/3}T^{2/3}$ [Besbes et al. 14]	**Yes** [Chen et al. 19]

What we show:

A much weaker notion of change admits adaptivity ...

Changes in Reward Y_t distribution

Non-Stationary Bandit

- **At time t**: select $a \in [K]$, observe $Y_t(a)$, with mean $\mu_t(a)$.
- **Dynamic Regret**: $\mathbf{R_T} \doteq \sum_{t=1}^{T} \underbrace{\mu_t^\star - \mu_t(a_t)}_{\delta_t(a_t)}$.

What was known:

Parameters	Best Oracle Rate	Adaptive Rates
L changes in $\mu_t(a)$	$\sqrt{LT}$ [Garivier, Moulines. 11]	**Yes** [Auer et al. 19]
S best-arm switches	$\sqrt{ST} \ll \sqrt{LT}$ [Auer. 02]	**OPEN**
Total-Variation V	$V^{1/3}T^{2/3}$ [Besbes et al. 14]	**Yes** [Chen et al. 19]

What we show:

A much weaker notion of change admits adaptivity ...

Key Contributions:

A new notion of Significant Shift (only most *severe* changes)

Best-arm-switches, or even large TV, *can be ignored*.

Adaptive Rates $(unknown\ parameters)$

$$\mathbb{E}\ R_T \lesssim \sum_{\text{Sig. Phases } \mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$$

Always faster than $\sqrt{ST} \ll \sqrt{LT}$, and faster than $V^{1/3}T^{2/3}$.

Key Contributions:

A new notion of Significant Shift (only most *severe* changes)

Best-arm-switches, or even large TV, *can be ignored*.

Adaptive Rates *(unknown parameters)*

$$\mathbb{E}\, R_T \lesssim \sum_{\text{Sig. Phases } \mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$$

Always faster than $\sqrt{ST} \ll \sqrt{LT}$, and faster than $V^{1/3}T^{2/3}$.

Key Contributions:

A new notion of Significant Shift (only most *severe* changes)

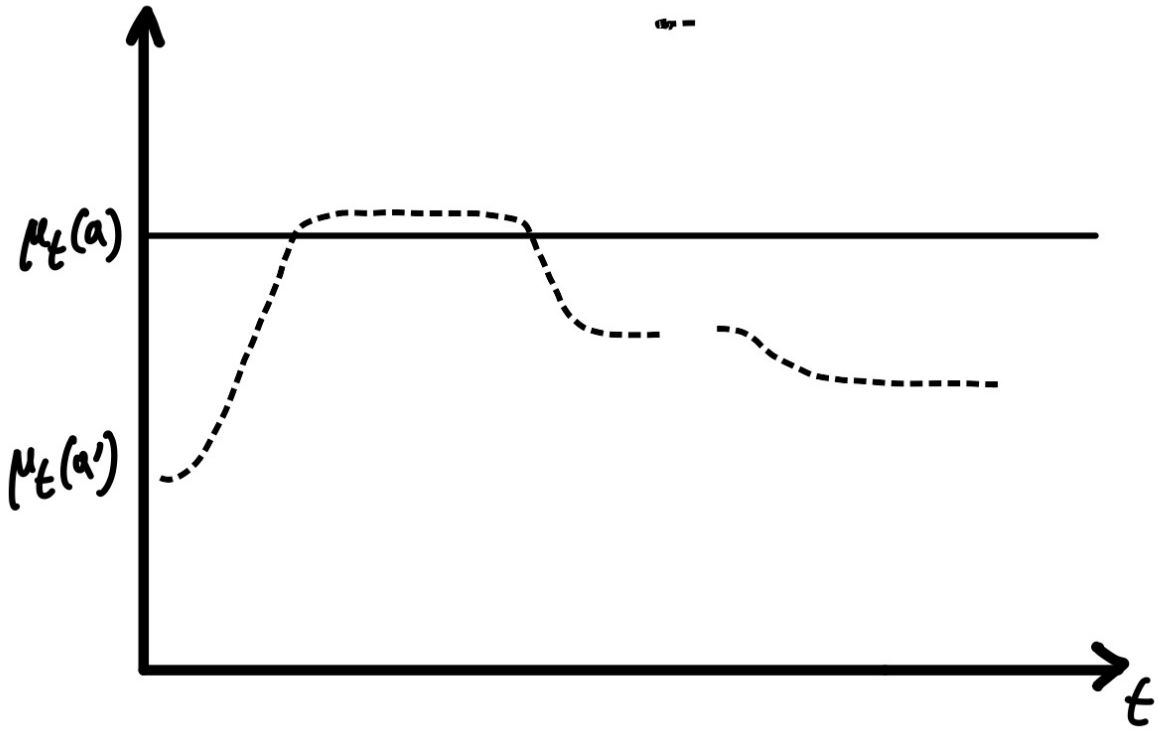
Best-arm-switches, or even large TV, *can be ignored*.

Adaptive Rates *(unknown parameters)*

$$\mathbb{E}\ R_T \lesssim \sum_{\text{Sig. Phases } \mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$$

Always faster than $\sqrt{ST} \ll \sqrt{LT}$, and faster than $V^{1/3}T^{2/3}$.

Intuition: various changes can safely be ignored

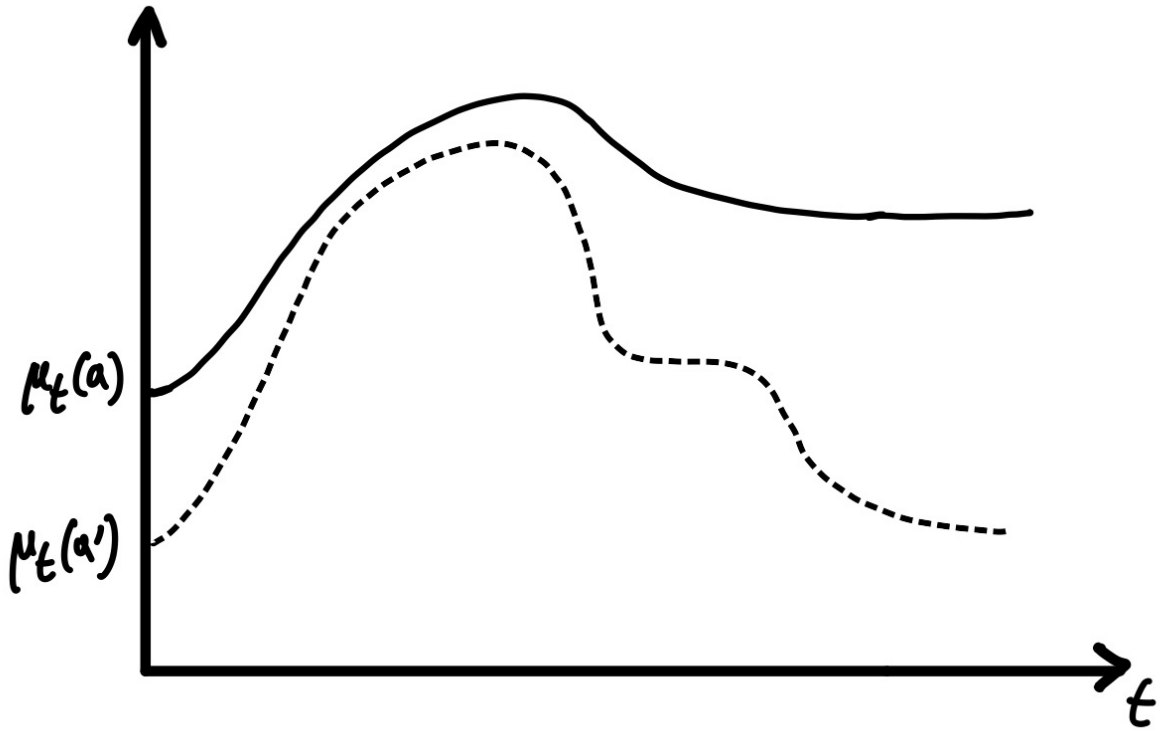


Best arm changes may not be "significant".

(e.g., when of small magnitude or duration)

We may have $R_T \lesssim \sqrt{T}$ while $\sqrt{ST} \approx T$

Intuition: various changes can safely be ignored

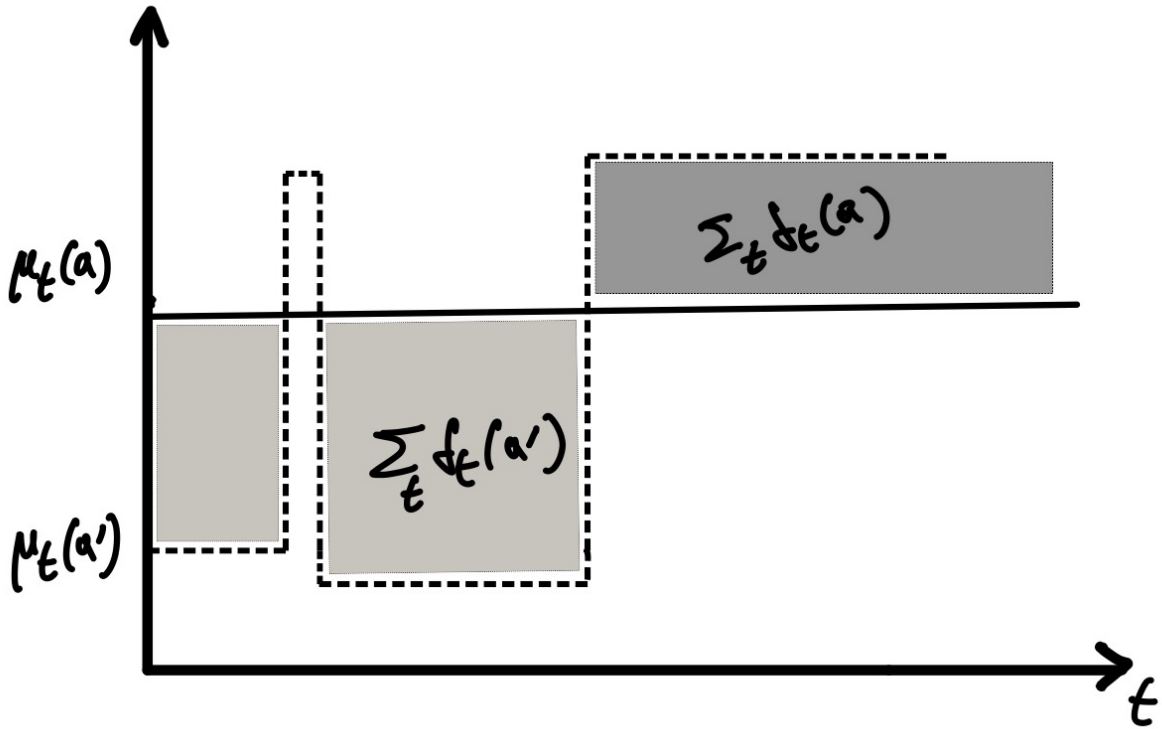


Large $V \doteq \sum_t \max_a |\mu_{t+1}(a) - \mu_t(a)|$ may not be "significant".

(e.g., if mean rewards remain close)

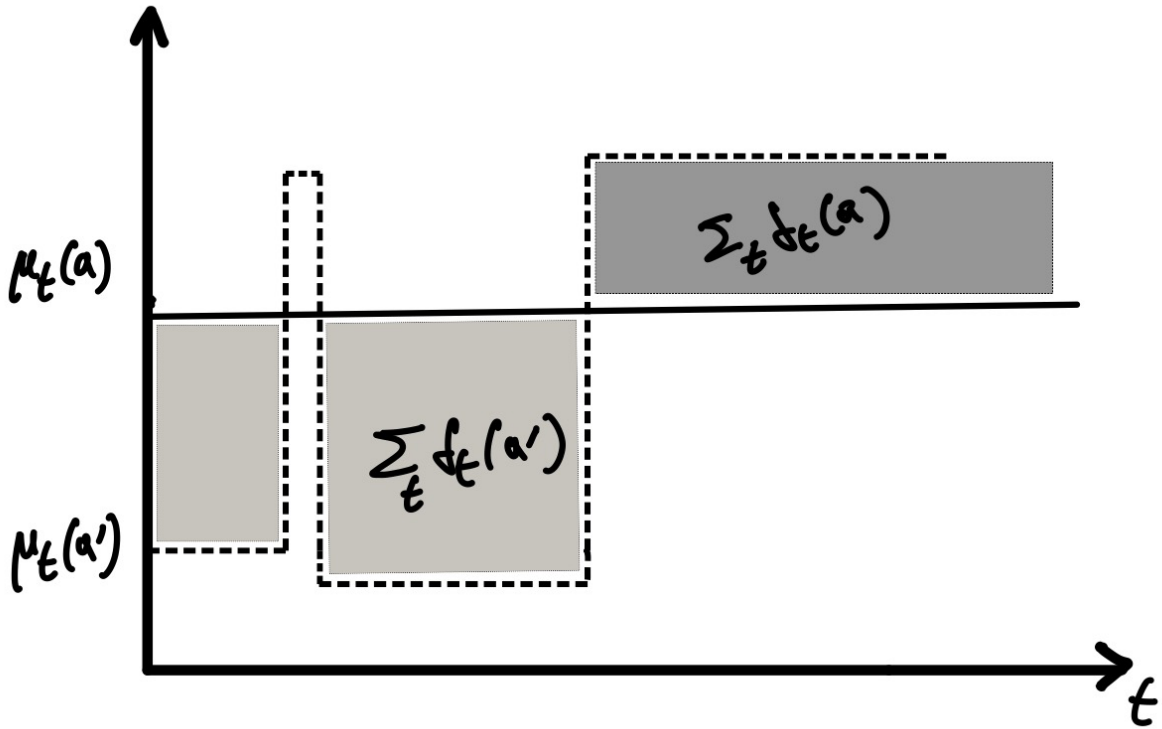
We may have $R_T \lesssim \sqrt{T}$ while $V^{1/3}T^{2/3} \approx T$

Hard to Ignore:



All arms became unsafe to play ... a' first, then a!

Hard to Ignore:



All arms became unsafe to play ... a' first, then a!

Definition (Significant Phases $\mathcal{P}_i$)

A Significant Phase ends only when no safe arm left to play

$$\forall a \in [K], \exists \text{ an interval } I, \quad \sum_{t \in I} \delta_t(a) \gtrsim \sqrt{|I|}.$$

$\Longrightarrow$ best arm switches, and large TV, but not the other way

Prop. (Sanity Check) An Oracle achieves $\mathbb{E}\ R_T \lesssim \sum_{\mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$

Key: $\mathcal{P}_i$ admits *last safe arm* $a^\sharp$, s.t. $\sum_{t \in \mathcal{P}_i} \delta_t(a^\sharp) \lesssim \sqrt{|\mathcal{P}_i|}$.

The same rate can be achieved adaptively by tracking $a^\sharp$ ...

Definition (Significant Phases $\mathcal{P}_i$)

A Significant Phase ends only when no safe arm left to play

$$\forall a \in [K], \exists \text{ an interval } I, \quad \sum_{t \in I} \delta_t(a) \gtrsim \sqrt{|I|}.$$

$\Longrightarrow$ best arm switches, and large TV, but not the other way

Prop. (Sanity Check) An Oracle achieves $\mathbb{E}\ R_T \lesssim \sum_{\mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$

Key: $\mathcal{P}_i$ admits *last safe arm* $a^\sharp$, s.t. $\sum_{t \in \mathcal{P}_i} \delta_t(a^\sharp) \lesssim \sqrt{|\mathcal{P}_i|}$.

The same rate can be achieved adaptively by tracking $a^\sharp$ ...

Definition (Significant Phases $\mathcal{P}_i$)

A Significant Phase ends only when no safe arm left to play

$$\forall a \in [K], \exists \text{ an interval } I, \quad \sum_{t \in I} \delta_t(a) \gtrsim \sqrt{|I|}.$$

$\Longrightarrow$ best arm switches, and large TV, but not the other way

Prop. (Sanity Check) An Oracle achieves $\mathbb{E}\ R_T \lesssim \sum_{\mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$

Key: $\mathcal{P}_i$ admits *last safe arm* $a^\sharp$, s.t. $\sum_{t \in \mathcal{P}_i} \delta_t(a^\sharp) \lesssim \sqrt{|\mathcal{P}_i|}$.

The same rate can be achieved adaptively by tracking $a^\sharp$ ...

Definition (Significant Phases $\mathcal{P}_i$)

A Significant Phase ends only when no safe arm left to play

$$\forall a \in [K], \exists \text{ an interval } I, \quad \sum_{t \in I} \delta_t(a) \gtrsim \sqrt{|I|}.$$

$\Longrightarrow$ best arm switches, and large TV, but not the other way

Prop. (Sanity Check) An Oracle achieves $\mathbb{E}\ R_T \lesssim \sum_{\mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$

Key: $\mathcal{P}_i$ admits *last safe arm* $a^\sharp$, s.t. $\sum_{t \in \mathcal{P}_i} \delta_t(a^\sharp) \lesssim \sqrt{|\mathcal{P}_i|}$.

The same rate can be achieved adaptively by tracking $a^\sharp$ ...

Definition (Significant Phases $\mathcal{P}_i$)

A Significant Phase ends only when no safe arm left to play

$$\forall a \in [K], \exists \text{ an interval } I, \quad \sum_{t \in I} \delta_t(a) \gtrsim \sqrt{|I|}.$$

$\Longrightarrow$ best arm switches, and large TV, but not the other way

Prop. (Sanity Check) An Oracle achieves $\mathbb{E}\ R_T \lesssim \sum_{\mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$

Key: $\mathcal{P}_i$ admits *last safe arm* $a^\sharp$, s.t. $\sum_{t \in \mathcal{P}_i} \delta_t(a^\sharp) \lesssim \sqrt{|\mathcal{P}_i|}$.

The same rate can be achieved adaptively by tracking $a^\sharp$ ...

Adaptive Procedure:

Key insights:

Track changes in $\sum_t \underbrace{\delta_t(a)}_{\mu_t^* - \mu_t(a)}$ rather than in $\mu_t(a)$ ...

Since $\sum_t \delta_t(a^\sharp)$ is small, it will suffice to track $\sum_t \delta_t(a', a)$ ...

Proceed in episodes:

- **Evict** arm a whenever $\sum_{t \in I} \hat{\delta}_t(a', a) \gtrsim \sqrt{|I|}$
- **Replay** evicted arms a' often to detect changes
- **Start new episode** when last arm $\hat{a}_\sharp$ gets evicted

2 Important Objects: $a^\sharp$ (Phase), and $\hat{a}_\sharp$ (Episode)

Adaptive Procedure:

Key insights:

Track changes in $\sum_t \underbrace{\delta_t(a)}_{\mu_t^* - \mu_t(a)}$ rather than in $\mu_t(a)$ ...

Since $\sum_t \delta_t(a^\sharp)$ is small, it will suffice to track $\sum_t \delta_t(a', a)$ ...

Proceed in episodes:

- **Evict** arm a whenever $\sum_{t \in I} \hat{\delta}_t(a', a) \gtrsim \sqrt{|I|}$
- **Replay** evicted arms a' often to detect changes
- **Start new episode** when last arm $\hat{a}_\sharp$ gets evicted

2 Important Objects: $a^\sharp$ (Phase), and $\hat{a}_\sharp$ (Episode)

Adaptive Procedure:

Key insights:

Track changes in $\sum_t \underbrace{\delta_t(a)}_{\mu_t^* - \mu_t(a)}$ rather than in $\mu_t(a)$ ...

Since $\sum_t \delta_t(a^\sharp)$ is small, it will suffice to track $\sum_t \delta_t(a', a)$ ...

... can estimate via $\hat{\mu}_t(a) \doteq Y_t(a) \cdot \dfrac{\mathbb{1}\{a_t = a\}}{\mathbb{P}(a_t = a)}$

Proceed in episodes:

- **Evict** arm a whenever $\sum_{t \in I} \hat{\delta}_t(a', a) \gtrsim \sqrt{|I|}$
- **Replay** evicted arms a' often to detect changes
- **Start new episode** when last arm $\hat{a}_\sharp$ gets evicted

Adaptive Procedure:

Key insights:

Track changes in $\sum_t \underbrace{\delta_t(a)}_{\mu_t^* - \mu_t(a)}$ rather than in $\mu_t(a)$ ...

Since $\sum_t \delta_t(a^\sharp)$ is small, it will suffice to track $\sum_t \delta_t(a', a)$ ...

Proceed in episodes:

- **Evict** arm a whenever $\sum_{t \in I} \hat{\delta}_t(a', a) \gtrsim \sqrt{|I|}$
- **Replay** evicted arms a' often to detect changes
- **Start new episode** when last arm $\hat{a}_\sharp$ gets evicted

2 **Important Objects:** $a^\sharp$ (Phase), and $\hat{a}_\sharp$ (Episode)

Adaptive Procedure:

Key insights:

Track changes in $\sum_t \underbrace{\delta_t(a)}_{\mu_t^* - \mu_t(a)}$ rather than in $\mu_t(a)$ ...

Since $\sum_t \delta_t(a^\sharp)$ is small, it will suffice to track $\sum_t \delta_t(a', a)$ ...

Proceed in episodes:

- **Evict** arm a whenever $\sum_{t \in I} \hat{\delta}_t(a', a) \gtrsim \sqrt{|I|}$
- **Replay** evicted arms a' often to detect changes
- **Start new episode** when last arm $\hat{a}_\sharp$ gets evicted

2 Important Objects: $a^\sharp$ (Phase), and $\hat{a}_\sharp$ (Episode)

Adaptive Procedure:

Key insights:

Track changes in $\sum_t \underbrace{\delta_t(a)}_{\mu_t^* - \mu_t(a)}$ rather than in $\mu_t(a)$ ...

Since $\sum_t \delta_t(a^\sharp)$ is small, it will suffice to track $\sum_t \delta_t(a', a)$ ...

Proceed in episodes:

- **Evict** arm a whenever $\sum_{t \in I} \hat{\delta}_t(a', a) \gtrsim \sqrt{|I|}$
- **Replay** evicted arms a' often to detect changes
- **Start new episode** when last arm $\hat{a}_\sharp$ gets evicted

2 Important Objects: $a^\sharp$ (Phase), and $\hat{a}_\sharp$ (Episode)

Adaptive Procedure:

Key insights:

Track changes in $\sum_t \underbrace{\delta_t(a)}_{\mu_t^* - \mu_t(a)}$ rather than in $\mu_t(a)$ ...

Since $\sum_t \delta_t(a^\sharp)$ is small, it will suffice to track $\sum_t \delta_t(a', a)$ ...

Proceed in episodes:

- **Evict** arm a whenever $\sum_{t \in I} \hat{\delta}_t(a', a) \gtrsim \sqrt{|I|}$
- **Replay** evicted arms a' often to detect changes
- **Start new episode** when last arm $\hat{a}_\sharp$ gets evicted

2 Important Objects: $a^\sharp$ (Phase), and $\hat{a}_\sharp$ (Episode)

Designing Random Replays:

1.) Consider Critical Segments:

Smallest Intervals I s.t. $\sum_{t \in I} \delta_t(a^\sharp, a) \gtrsim \sqrt{|I|}$

If we detect any of these, we evict a in time ...

$\therefore$ **Ideal Replay**:

Starts close to I and plays $a^\sharp$ for at least $|I|$ rounds ...

At time t, start Replay of duration $m = 2, 4, \ldots$, w' prob. $\delta_{t,m}$

. . .

Prop. For large $\delta_{t,m}$, few such I's before Ideal Replay

$\Longrightarrow$ **Since $\hat{a}_\sharp$ wasn't evicted, $\sum_t \delta_t(a^\sharp, \hat{a}_\sharp)$ must be small.**

Designing Random Replays:

1.) Consider Critical Segments:

Smallest Intervals I s.t. $\sum_{t \in I} \delta_t(a^\sharp, a) \gtrsim \sqrt{|I|}$

If we detect any of these, we evict a in time ...

$\therefore$ **Ideal Replay**:

Starts close to I and plays $a^\sharp$ for at least $|I|$ rounds ...

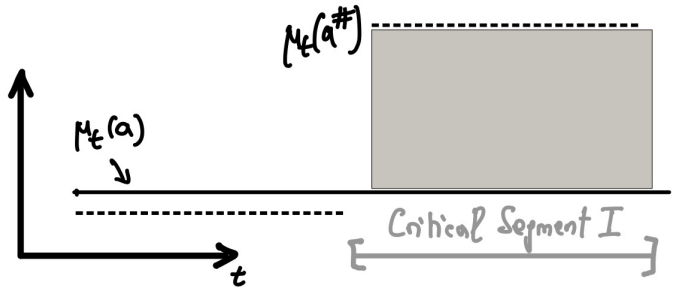
At time t, start Replay of duration $m = 2, 4, \ldots$, w' prob. $\delta_{t,m}$

...

Prop. For large $\delta_{t,m}$, few such I's before Ideal Replay

$\Longrightarrow$ **Since $\hat{a}_\sharp$ wasn't evicted, $\sum_t \delta_t(a^\sharp, \hat{a}_\sharp)$ must be small.**

Designing Random Replays:



1.) Consider Critical Segments:

Smallest Intervals I s.t. $\sum_{t\in I} \delta_t(a^\sharp, a) \gtrsim \sqrt{|I|}$

If we detect any of these, we evict a in time ...

$\therefore$ **Ideal Replay**:

Starts close to I and plays $a^\sharp$ for at least $|I|$ rounds ...

At time t, start Replay of duration $m = 2, 4, \ldots$, w' prob. $\delta_{t,m}$

. . .

Prop. For large $\delta_{t,m}$, few such I's before Ideal Replay

Designing Random Replays:

1.) Consider Critical Segments:

Smallest Intervals I s.t. $\sum_{t \in I} \delta_t(a^\sharp, a) \gtrsim \sqrt{|I|}$

If we detect any of these, we evict a in time ...

$\therefore$ Ideal Replay:

Starts close to I and plays $a^\sharp$ for at least $|I|$ rounds ...

At time t, start Replay of duration $m = 2, 4, \ldots$, w' prob. $\delta_{t,m}$

$\ldots$

Prop. For large $\delta_{t,m}$, few such I's before Ideal Replay

$\Longrightarrow$ **Since $\hat{a}_\sharp$ wasn't evicted, $\sum_t \delta_t(a^\sharp, \hat{a}_\sharp)$ must be small.**

Designing Random Replays:

1.) Consider Critical Segments:

Smallest Intervals I s.t. $\sum_{t \in I} \delta_t(a^\sharp, a) \gtrsim \sqrt{|I|}$

If we detect any of these, we evict a in time ...

$\therefore$ **Ideal Replay**:

Starts close to I and plays $a^\sharp$ for at least $|I|$ rounds ...

At time t, start Replay of duration $m = 2, 4, \ldots$, w' prob. $\delta_{t,m}$

$\ldots$

Prop. For large $\delta_{t,m}$, few such I's before Ideal Replay

$\Longrightarrow$ **Since $\hat{a}_\sharp$ wasn't evicted, $\sum_t \delta_t(a^\sharp, \hat{a}_\sharp)$ must be small.**

Designing Random Replays:

1.) Consider Critical Segments:

Smallest Intervals I s.t. $\sum_{t \in I} \delta_t(a^\sharp, a) \gtrsim \sqrt{|I|}$

If we detect any of these, we evict a in time ...

$\therefore$ **Ideal Replay**:

Starts close to I and plays $a^\sharp$ for at least $|I|$ rounds ...

At time t, start Replay of duration $m = 2, 4, \ldots$, w' prob. $\delta_{t,m}$

. . .

Prop. For large $\delta_{t,m}$, few such I's before Ideal Replay

$\Longrightarrow$ **Since $\hat{a}_\sharp$ wasn't evicted, $\sum_t \delta_t(a^\sharp, \hat{a}_\sharp)$ must be small.**

Designing Random Replays:

1.) Consider Critical Segments:

Smallest Intervals I s.t. $\sum_{t \in I} \delta_t(a^\sharp, a) \gtrsim \sqrt{|I|}$

If we detect any of these, we evict a in time ...

$\therefore$ **Ideal Replay**:

Starts close to I and plays $a^\sharp$ for at least $|I|$ rounds ...

At time t, start Replay of duration $m = 2, 4, \ldots$, w' prob. $\delta_{t,m}$

. . .

Prop. For large $\delta_{t,m}$, few such I's before Ideal Replay

$\Longrightarrow$ **Since $\hat{a}_\sharp$ wasn't evicted, $\sum_t \delta_t(a^\sharp, \hat{a}_\sharp)$ must be small.**

Designing Random Replays:

1.) Consider Critical Segments:

Smallest Intervals I s.t. $\sum_{t \in I} \delta_t(a^\sharp, a) \gtrsim \sqrt{|I|}$

If we detect any of these, we evict a in time ...

$\therefore$ **Ideal Replay**:

Starts close to I and plays $a^\sharp$ for at least $|I|$ rounds ...

At time t, start Replay of duration $m = 2, 4, \ldots$, w' prob. $\delta_{t,m}$

. . .

Prop. For large $\delta_{t,m}$, few such I's before Ideal Replay

$\Longrightarrow$ **Since $\hat{a}_\sharp$ wasn't evicted, $\sum_t \delta_t(a^\sharp, \hat{a}_\sharp)$ must be small.**

Designing Random Replays:

2.) Consider Cost of Replaying Bad a's:

We don't want too many replays $\Longrightarrow$ Tradeoff on $\delta_{t,m}$

In fact dynamic regret $\sum_t \delta_t(a_t)$ can get large during replay ...

However, a Sufficient Fact:

Replay of duration m contributes $\sum_t \delta_t(\hat{a}^\sharp, a_t) \lesssim \sqrt{m}$

$\Longrightarrow$ **Schedule $\delta_{t,m} \approx \sqrt{\frac{1}{t \cdot m}}$ ensures small Replay regret ...**

Schedule similar to [Auer et al. 19], [Chen et al. 19], [Wei and Luo 21]

Designing Random Replays:

2.) Consider Cost of Replaying Bad a's:

We don't want too many replays $\Longrightarrow$ Tradeoff on $\delta_{t,m}$

In fact dynamic regret $\sum_t \delta_t(a_t)$ can get large during replay ...

However, a Sufficient Fact:

Replay of duration m contributes $\sum_t \delta_t(\hat{a}^\sharp, a_t) \lesssim \sqrt{m}$

$\Longrightarrow$ **Schedule $\delta_{t,m} \approx \sqrt{\frac{1}{t \cdot m}}$ ensures small Replay regret ...**

Schedule similar to [Auer et al. 19], [Chen et al. 19], [Wei and Luo 21]

Designing Random Replays:

2.) Consider Cost of Replaying Bad a's:

We don't want too many replays $\Longrightarrow$ Tradeoff on $\delta_{t,m}$

In fact dynamic regret $\sum_t \delta_t(a_t)$ can get large during replay ...

However, a Sufficient Fact:

Replay of duration m contributes $\sum_t \delta_t(\hat{a}^\sharp, a_t) \lesssim \sqrt{m}$

$\Longrightarrow$ **Schedule $\delta_{t,m} \approx \sqrt{\frac{1}{t \cdot m}}$ ensures small Replay regret ...**

Schedule similar to [Auer et al. 19], [Chen et al. 19], [Wei and Luo 21]

Designing Random Replays:

2.) Consider Cost of Replaying Bad a's:

We don't want too many replays $\Longrightarrow$ Tradeoff on $\delta_{t,m}$

In fact dynamic regret $\sum_t \delta_t(a_t)$ can get large during replay ...

However, a Sufficient Fact:

Replay of duration m contributes $\sum_t \delta_t(\hat{a}^\sharp, a_t) \lesssim \sqrt{m}$

$\Longrightarrow$ **Schedule $\delta_{t,m} \approx \sqrt{\frac{1}{t \cdot m}}$ ensures small Replay regret ...**

Schedule similar to [Auer et al. 19], [Chen et al. 19], [Wei and Luo 21]

Designing Random Replays:

2.) Consider Cost of Replaying Bad a's:

We don't want too many replays $\Longrightarrow$ Tradeoff on $\delta_{t,m}$

In fact dynamic regret $\sum_t \delta_t(a_t)$ can get large during replay ...

However, a Sufficient Fact:

Replay of duration m contributes $\sum_t \delta_t(\hat{a}^\sharp, a_t) \lesssim \sqrt{m}$

$\Longrightarrow$ **Schedule $\delta_{t,m} \approx \sqrt{\frac{1}{t \cdot m}}$ ensures small Replay regret ...**

Schedule similar to [Auer et al. 19], [Chen et al. 19], [Wei and Luo 21]

Designing Random Replays:

2.) Consider Cost of Replaying Bad a's:

We don't want too many replays $\implies$ Tradeoff on $\delta_{t,m}$

In fact dynamic regret $\sum_t \delta_t(a_t)$ can get large during replay ...

However, a Sufficient Fact:

Replay of duration m contributes $\sum_t \delta_t(\hat{a}^\sharp, a_t) \lesssim \sqrt{m}$

$\implies$ **Schedule $\delta_{t,m} \approx \sqrt{\frac{1}{t \cdot m}}$ ensures small Replay regret ...**

Schedule similar to [Auer et al. 19], [Chen et al. 19], [Wei and Luo 21]

... as a result:

Theo. *The following rates are achieved adaptively:*

$$\mathbb{E}\, R_T \lesssim \sum_{\text{Sig. Phases } \mathcal{P}_i} \sqrt{K \cdot |\mathcal{P}_i|}$$

Always faster than $\sqrt{ST} \ll \sqrt{LT}$, and faster than $V^{1/3}T^{2/3}$.

Remark: Recent arXiv of [Yadkori et al. 22] independently gets $\sqrt{ST}$ ...

Open: it's unclear whether our notion may be further weakened ...

... as a result:

Theo. *The following rates are achieved adaptively:*

$$\mathbb{E}\ R_T \lesssim \sum_{\text{Sig. Phases } \mathcal{P}_i} \sqrt{K \cdot |\mathcal{P}_i|}$$

Always faster than $\sqrt{ST} \ll \sqrt{LT}$, and faster than $V^{1/3}T^{2/3}$.

Remark: Recent arXiv of [Yadkori et al. 22] independently gets $\sqrt{ST}$ ...

Open: it's unclear whether our notion may be further weakened ...

... as a result:

Theo. *The following rates are achieved adaptively:*

$$\mathbb{E}\, R_T \lesssim \sum_{\text{Sig. Phases } \mathcal{P}_i} \sqrt{K \cdot |\mathcal{P}_i|}$$

Always faster than $\sqrt{ST} \ll \sqrt{LT}$, and faster than $V^{1/3}T^{2/3}$.

Remark: Recent arXiv of [Yadkori et al. 22] independently gets $\sqrt{ST}$ ...

Open: it's unclear whether our notion may be further weakened ...

Outline:

When are these changes severe and can we adapt?

- Changes in Reward Y_t distribution
  In fact, we'll simply consider regular bandits (COLT 2022)
- **Changes in Context X_t distribution**
  as in Covariate Shift (ALT 2021)
- Bringing it all together ... (ongoing work)

Changes in Context X_t distribution

Covariate Shift

- **At time t**: Observe X_t, select $a \in [K]$, observe $Y_t(a) \mid X_t$.
- **Covariate-Shift**: dist(X_t) changes, but $\mu(a \mid X_t)$ fixed.
- **Interval Regret**: Suppose $X_t \sim Q_X$ for $t \in I_Q$,

$$\mathbf{R}(Q_X) \doteq \sum_{t \in I_Q} \mathop{\mathbb{E}}_{X_t \sim Q_X} \mu^\star(X_t) - \mu(a_t \mid X_t)$$

How is this affected by past distributions $P_{1,X}, P_{2,X}, \ldots Q_X$?

What we show:

No need to restart $\implies$ **keep learning up to dist**$(P_{i,X} \to Q_X)$

However, usual (stationary) strategies can be suboptimal ...

Changes in Context X_t distribution

Covariate Shift

- **At time t**: Observe X_t, select $a \in [K]$, observe $Y_t(a) \mid X_t$.
- **Covariate-Shift**: dist(X_t) changes, but $\mu(a \mid X_t)$ fixed.
- **Interval Regret**: Suppose $X_t \sim Q_X$ for $t \in I_Q$,

$$\mathbf{R}(Q_X) \doteq \sum_{t \in I_Q} \mathop{\mathbb{E}}_{X_t \sim Q_X} \mu^\star(X_t) - \mu(a_t \mid X_t)$$

How is this affected by past distributions $P_{1,X}, P_{2,X}, \ldots Q_X$?

What we show:

No need to restart $\implies$ **keep learning up to dist**$(P_{i,X} \to Q_X)$

However, usual (stationary) strategies can be suboptimal ...

Changes in Context X_t distribution

Covariate Shift

- **At time t**: Observe X_t, select $a \in [K]$, observe $Y_t(a) \mid X_t$.
- **Covariate-Shift**: dist(X_t) changes, but $\mu(a \mid X_t)$ fixed.
- **Interval Regret**: Suppose $X_t \sim Q_X$ for $t \in I_Q$,

$$\mathbf{R}(Q_X) \doteq \sum_{t \in I_Q} \mathop{\mathbb{E}}_{X_t \sim Q_X} \mu^\star(X_t) - \mu(a_t \mid X_t)$$

How is this affected by past distributions $P_{1,X}, P_{2,X}, \ldots Q_X$?

What we show:

No need to restart $\implies$ **keep learning up to dist**$(P_{i,X} \to Q_X)$

However, usual (stationary) strategies can be suboptimal ...

Changes in Context X_t distribution

Covariate Shift

- **At time t**: Observe X_t, select $a \in [K]$, observe $Y_t(a) \mid X_t$.
- **Covariate-Shift**: dist(X_t) changes, but $\mu(a \mid X_t)$ fixed.
- **Interval Regret**: Suppose $X_t \sim Q_X$ for $t \in I_Q$,

$$\mathbf{R}(Q_X) \doteq \sum_{t \in I_Q} \mathop{\mathbb{E}}_{X_t \sim Q_X} \mu^\star(X_t) - \mu(a_t \mid X_t)$$

How is this affected by past distributions $P_{1,X}, P_{2,X}, \ldots Q_X$?

What we show:

No need to restart $\implies$ **keep learning up to dist**$(P_{i,X} \to Q_X)$

However, usual (stationary) strategies can be suboptimal ...

Usual (stationary) Strategies

Assumption: *similar rewards for similar X's*

(e.g., $\mu(a \mid x)$ is Lipschitz in x)

Partition X space into cells of size r:

- Observe context X_t (Patient profile)
- Play Bandit for $B_r(X_t)$ (Similar Treatments)

Best Choice of r depends on P_X ...

For a fixed P_X, and horizon T, you could choose $r = r(P_X; T)$

[Rigollet, Zeevi 10], [Slivkins 13], [Perchet, Rigollet, 13], [Reeve et al. 18] ...

Changing $\mathbf{P_X}$: choose r locally as $r(X_t)$ ...

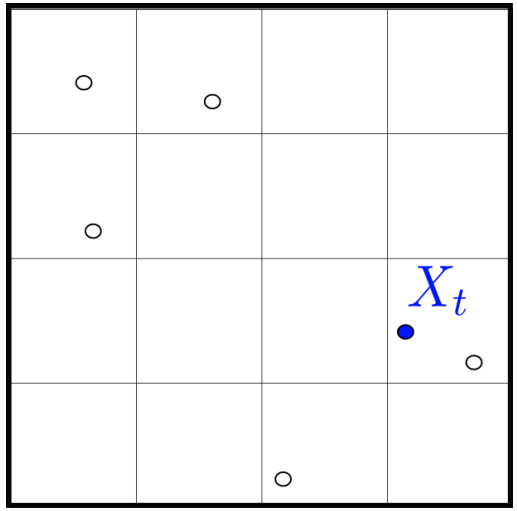
Usual (stationary) Strategies

Assumption: *similar rewards for similar X's*

(e.g., $\mu(a \mid x)$ is Lipschitz in x)

Partition X space into cells of size r:

- Observe context X_t (Patient profile)
- Play Bandit for $B_r(X_t)$ (Similar Treatments)



Best Choice of r depends on P_X ...

For a fixed P_X, and horizon T, you could choose $r = r(P_X; T)$

[Rigollet, Zeevi 10], [Slivkins 13], [Perchet, Rigollet, 13], [Reeve et al. 18] ...

Changing P_X: choose r locally as $r(X_t)$ ...

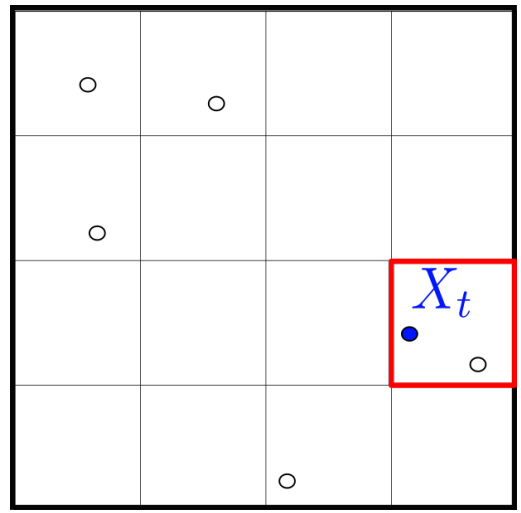
Usual (stationary) Strategies

Assumption: *similar rewards for similar X's*

(e.g., $\mu(a \mid x)$ is Lipschitz in x)

Partition X space into cells of size r:

- Observe context X_t (Patient profile)
- Play Bandit for $B_r(X_t)$ (Similar Treatments)



Best Choice of r depends on P_X ...

For a fixed P_X, and horizon T, you could choose $r = r(P_X; T)$

[Rigollet, Zeevi 10], [Slivkins 13], [Perchet, Rigollet, 13], [Reeve et al. 18] ...

Changing P_X: choose r locally as $r(X_t)$ ...

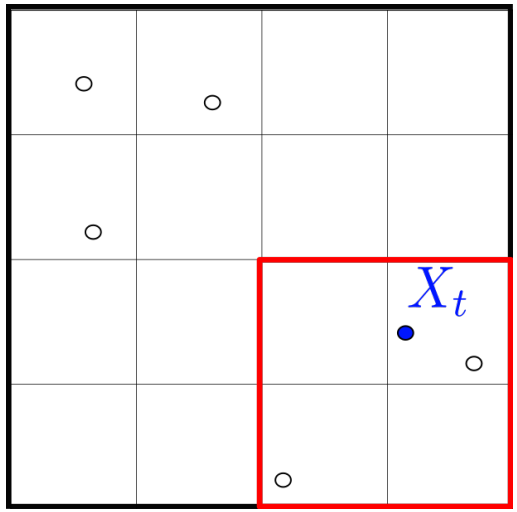
Usual (stationary) Strategies

Assumption: *similar rewards for similar X's*

(e.g., $\mu(a \mid x)$ is Lipschitz in x)

Partition X space into cells of size r:

- Observe context X_t (Patient profile)
- Play Bandit for $B_r(X_t)$ (Similar Treatments)



Best Choice of r depends on P_X ...

For a fixed P_X, and horizon T, you could choose $r = r(P_X; T)$

[Rigollet, Zeevi 10], [Slivkins 13], [Perchet, Rigollet, 13], [Reeve et al. 18] ...

Changing $\mathbf{P_X}$: choose r locally as $r(X_t)$ ...

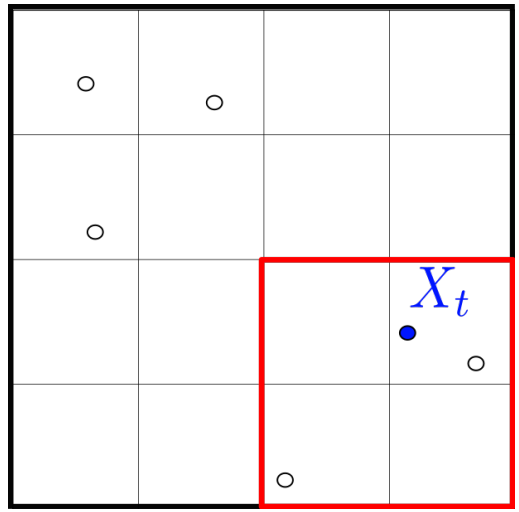
Usual (stationary) Strategies

Assumption: *similar rewards for similar X's*

(e.g., $\mu(a \mid x)$ is Lipschitz in x)

Partition X space into cells of size r:

- Observe context X_t (Patient profile)
- Play Bandit for $B_r(X_t)$ (Similar Treatments)



Best Choice of r depends on P_X ...

For a fixed P_X, and horizon T, you could choose $r = r(P_X; T)$

[Rigollet, Zeevi 10], [Slivkins 13], [Perchet, Rigollet, 13], [Reeve et al. 18] ...

Changing $\mathbf{P_X}$: choose r locally as $r(X_t)$ ...

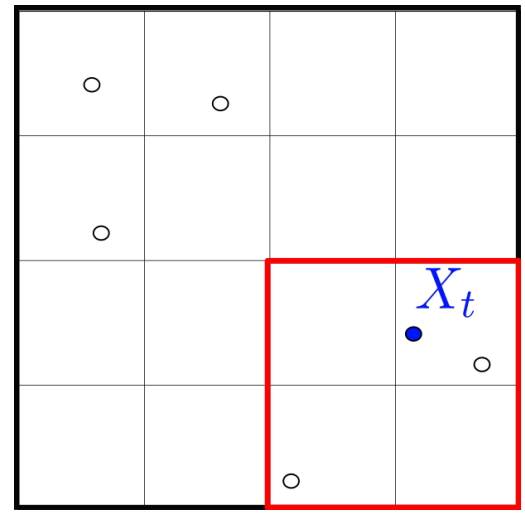
Usual (stationary) Strategies

Assumption: *similar rewards for similar X's*

(e.g., $\mu(a \mid x)$ is Lipschitz in x)

Partition X space into cells of size r:

- Observe context X_t (Patient profile)
- Play Bandit for $B_r(X_t)$ (Similar Treatments)



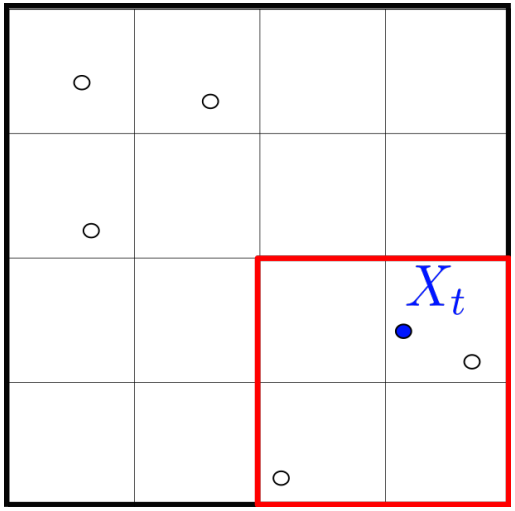
Best Choice of r depends on P_X ...

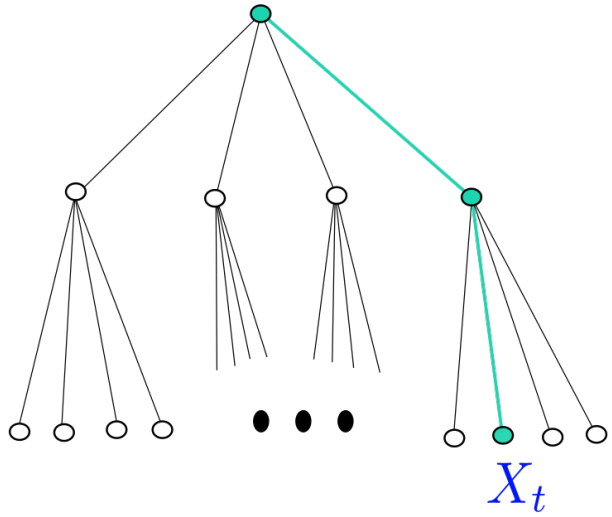
For a fixed P_X, and horizon T, you could choose $r = r(P_X; T)$

[Rigollet, Zeevi 10], [Slivkins 13], [Perchet, Rigollet, 13], [Reeve et al. 18] ...

Changing $\mathbf{P_X}$: choose r locally as $r(X_t)$ ...

Choose r locally as $r(X_t)$





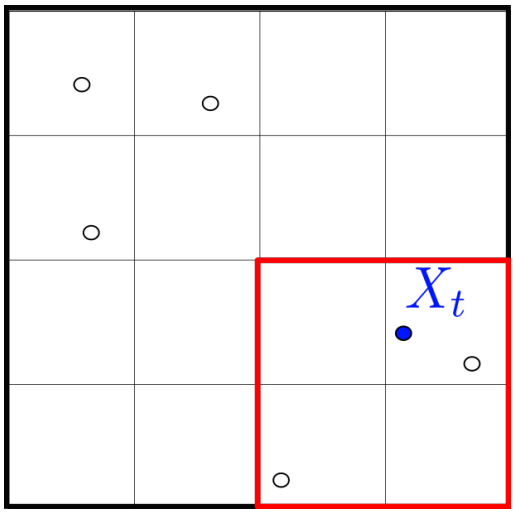
Maintain bandits at each level ...

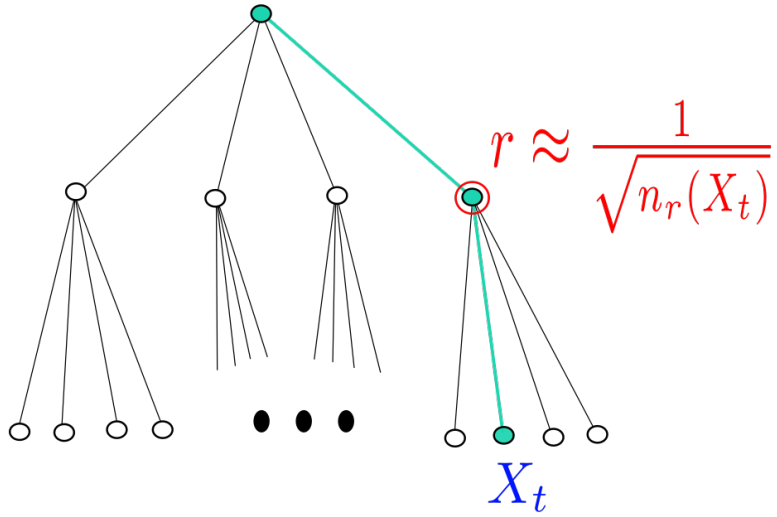
Keep Playing through Changes:

$n_r(X_t)$ automatically adapts to changes in distribution ...

For $X_t \sim Q_X$, large $n_r(X_t) \iff$ past $P_{i,X}$ yield similar points

Choose r locally as $r(X_t)$





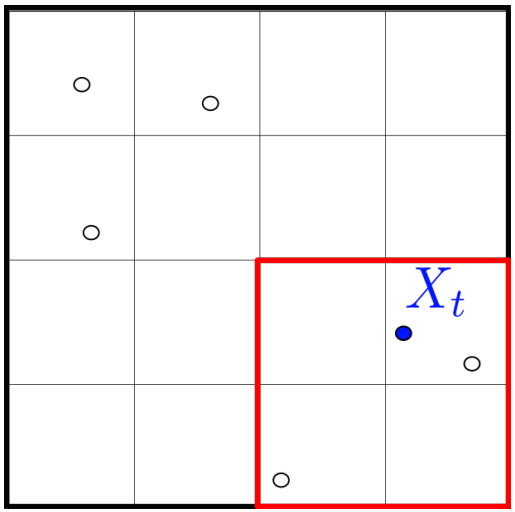
Maintain bandits at each level ...

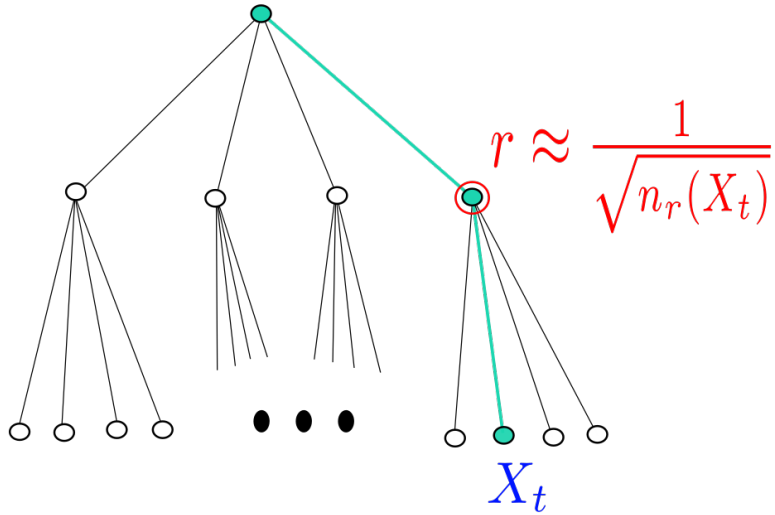
Keep Playing through Changes:

$n_r(X_t)$ automatically adapts to changes in distribution ...

For $X_t \sim Q_X$, large $n_r(X_t) \iff$ past $P_{i,X}$ yield similar points

Choose r locally as $r(X_t)$





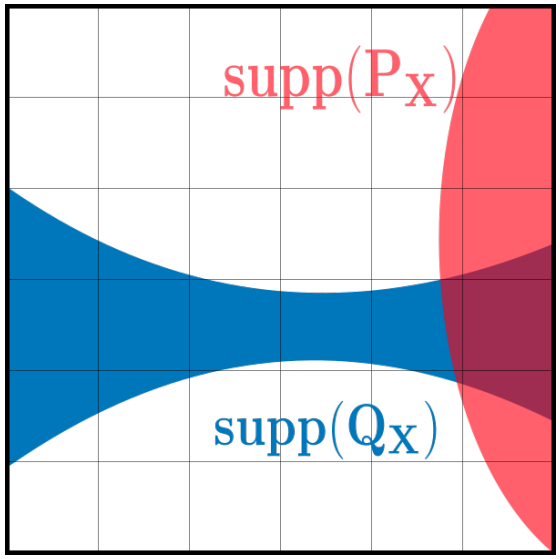
Maintain bandits at each level ...

Keep Playing through Changes:

$n_r(X_t)$ automatically adapts to changes in distribution ...

For $X_t \sim Q_X$, large $n_r(X_t) \iff$ past $P_{i,X}$ yield similar points

How informative is previous P_X for Q_X?



Nearly not Informative

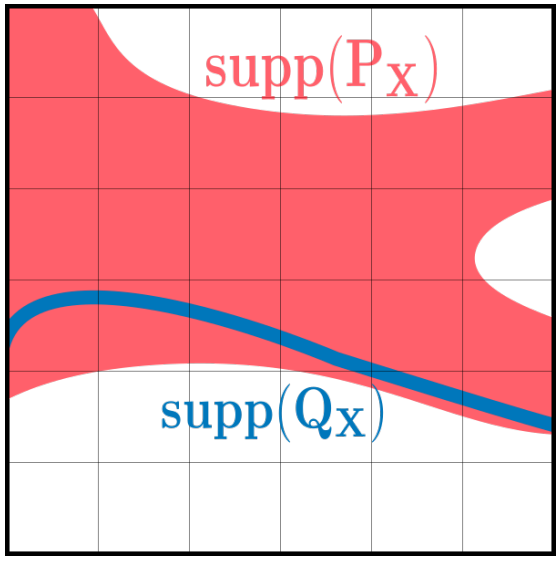
Transfer Exponent $\gamma \in [0, \infty]$:

$$\forall r \in [0,1], \quad P_X(B_r) \gtrsim r^\gamma \cdot Q_X(B_r)$$

P_X **is large where** Q_X **is large ...** [Kpo., Martinet, AoS 21]

Assymetry $\Longrightarrow$ metrics (e.g., TV, Wasserstein) are inadequate.

How informative is previous P_X for Q_X?



More informative

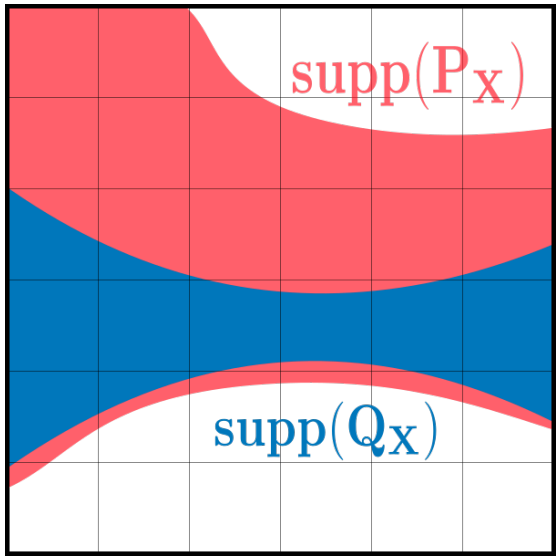
Transfer Exponent $\gamma \in [0, \infty]$:

$$\forall r \in [0,1], \quad P_X(B_r) \gtrsim r^\gamma \cdot Q_X(B_r)$$

P_X **is large where** Q_X **is large ...** [Kpo., Martinet, AoS 21]

Assymetry $\implies$ metrics (e.g., TV, Wasserstein) are inadequate.

How informative is previous P_X for Q_X?



Quite informative

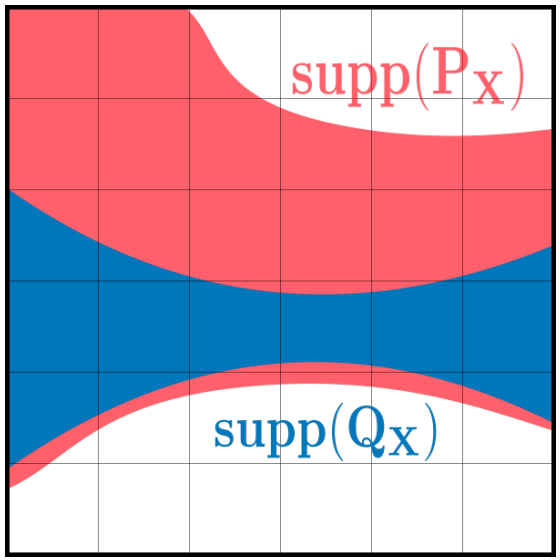
Transfer Exponent $\gamma \in [0, \infty]$:

$$\forall r \in [0,1], \quad P_X(B_r) \gtrsim r^\gamma \cdot Q_X(B_r)$$

P_X **is large where** Q_X **is large ...** [Kpo., Martinet, AoS 21]

Assymetry $\longrightarrow$ metrics (e.g., TV, Wasserstein) are inadequate.

How informative is previous P_X for Q_X?



Quite informative

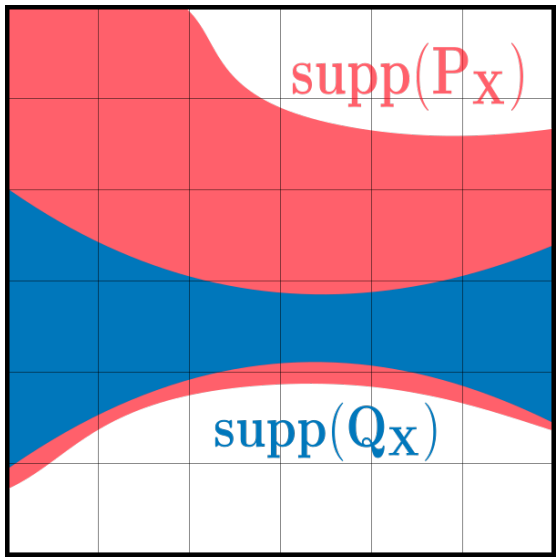
Transfer Exponent $\gamma \in [0, \infty]$:

$$\forall r \in [0,1], \quad P_X(B_r) \gtrsim r^\gamma \cdot Q_X(B_r)$$

P_X is large where Q_X is large ... [Kpo., Martinet, AoS 21]

Assymetry $\Longrightarrow$ metrics (e.g., TV, Wasserstein) are inadequate.

How informative is previous P_X for Q_X?



Quite informative

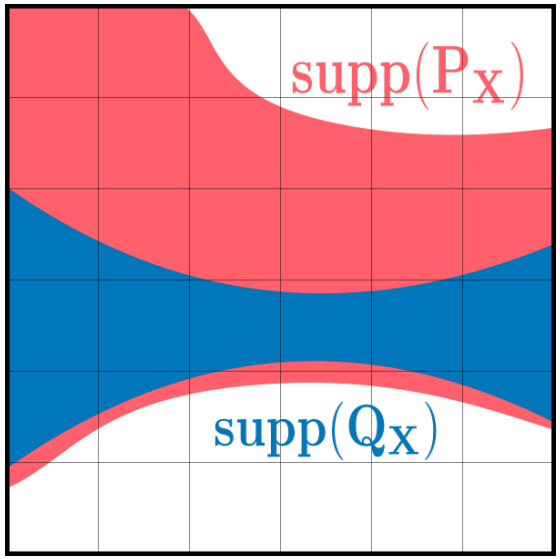
Transfer Exponent $\gamma \in [0, \infty]$:

$$\forall r \in [0,1], \quad P_X(B_r) \gtrsim r^\gamma \cdot Q_X(B_r)$$

P_X is large where Q_X is large ... [Kpo., Martinet, AoS 21]

Assymetry $\Longrightarrow$ metrics (e.g., TV, Wasserstein) are inadequate.

How informative is previous P_X for Q_X?



Quite informative

Transfer Exponent $\gamma \in [0, \infty]$:

$$\forall r \in [0,1], \quad P_X(B_r) \gtrsim r^\gamma \cdot Q_X(B_r)$$

P_X is large where Q_X is large ... [Kpo., Martinet, AoS 21]

Assymetry $\implies$ metrics (e.g., TV, Wasserstein) are inadequate.

Adaptive Guarantees on $\mathbf{R}(Q_X)$

Theo. Suppose n_P rounds before Q_X, then n_Q rounds:

$$\mathbb{E}\,\mathbf{R}(Q_X) \lesssim n_Q \cdot \left(n_P^{\varphi} + n_Q\right)^{-\frac{\alpha+1}{2+\alpha+d}},$$

for a function $\varphi = \varphi(\gamma) \searrow 0$, γ unknown.

α captures *margin* between arms.

The regret is tight, and integrates past information via n_P^{φ}.

Main Message:

Local strategy adapts to severity of changes

Adaptive Guarantees on $\mathbf{R}(Q_X)$

Theo. Suppose n_P rounds before Q_X, then n_Q rounds:

$$\mathbb{E}\,\mathbf{R}(Q_X) \lesssim n_Q \cdot \left(n_P^{\varphi} + n_Q\right)^{-\frac{\alpha+1}{2+\alpha+d}},$$

for a function $\varphi = \varphi(\gamma) \searrow 0$, γ unknown.

α captures *margin* between arms.

The regret is tight, and integrates past information via n_P^{φ}.

Main Message:

Local strategy adapts to severity of changes

Adaptive Guarantees on $\mathbf{R}(Q_X)$

Theo. Suppose n_P rounds before Q_X, then n_Q rounds:

$$\mathbb{E}\,\mathbf{R}(Q_X) \lesssim n_Q \cdot \left(n_P^{\varphi} + n_Q\right)^{-\frac{\alpha+1}{2+\alpha+d}},$$

for a function $\varphi = \varphi(\gamma) \searrow 0$, γ unknown.

α captures *margin* between arms.

The regret is tight, and integrates past information via n_P^{φ}.

Main Message:

Local strategy adapts to severity of changes

Adaptive Guarantees on $\mathbf{R}(Q_X)$

Theo. Suppose n_P rounds before Q_X, then n_Q rounds:

$$\mathbb{E}\,\mathbf{R}(Q_X) \lesssim n_Q \cdot \left(n_P^{\varphi} + n_Q\right)^{-\frac{\alpha+1}{2+\alpha+d}},$$

for a function $\varphi = \varphi(\gamma) \searrow 0$, γ unknown.

α captures *margin* between arms.

The regret is tight, and integrates past information via n_P^{φ}.

Main Message:

Local strategy adapts to severity of changes

Adaptive Guarantees on $\mathbf{R}(Q_X)$

Theo. Suppose n_P rounds before Q_X, then n_Q rounds:

$$\mathbb{E}\,\mathbf{R}(Q_X) \lesssim n_Q \cdot \left(n_P^{\varphi} + n_Q\right)^{-\frac{\alpha+1}{2+\alpha+d}},$$

for a function $\varphi = \varphi(\gamma) \searrow 0$, γ unknown.

α captures *margin* between arms.

The regret is tight, and integrates past information via n_P^{φ}.

Main Message:

Local strategy adapts to severity of changes

Adaptive Guarantees on $\mathrm{R}(Q_X)$

Theo. Suppose n_P rounds before Q_X, then n_Q rounds:

$$\mathbb{E}\,\mathbf{R}(Q_X) \lesssim n_Q \cdot \left(n_P^{\varphi} + n_Q\right)^{-\frac{\alpha+1}{2+\alpha+d}},$$

for a function $\varphi = \varphi(\gamma) \searrow 0$, γ unknown.

α captures *margin* between arms.

The regret is tight, and integrates past information via n_P^{φ}.

Main Message:

Local strategy adapts to severity of changes

Outline:

When are these changes severe and can we adapt?

- Changes in Reward Y_t distribution
  In fact, we'll simply consider regular bandits (COLT 2022)
- Changes in Context X_t distribution
  as in Covariate Shift (ALT 2021)
- **Bringing it all together ...** (ongoing work)
  A clean measure of *joint* severity remains elusive ...

Outline:

When are these changes severe and can we adapt?

- Changes in Reward Y_t distribution
  In fact, we'll simply consider regular bandits (COLT 2022)
- Changes in Context X_t distribution
  as in Covariate Shift (ALT 2021)
- **Bringing it all together ...** (ongoing work)
  A clean measure of *joint* severity remains elusive ...

Summary

Main Design Element:

Track most severe changes, otherwise keep learning

Requires understanding the severity of changes ...

Thanks!

Summary

Main Design Element:

Track most severe changes, otherwise keep learning

Requires understanding the severity of changes ...

Thanks!

Summary

Main Design Element:

Track most severe changes, otherwise keep learning

Requires understanding the severity of changes ...

Thanks!