Sequential Learning Under Distribution Shifts

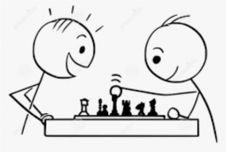
 . . .  . . . 

Joe Suk
NYU

joint with Samory Kpotufe (Columbia), Arpit Agarwal (IIT Bombay), Jung-hun Kim (CREST/ENSAE)

Long Term Motivation:

Sequential decisions under noisy, partial feedback

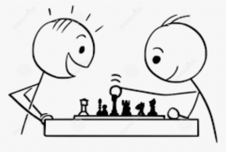
(Contextual) Bandits, Reinforcement Learning, . . .

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

However, the Environment changes frequently in practice ☹

Long Term Motivation:

Sequential decisions under noisy, partial feedback



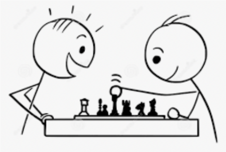
(Contextual) Bandits, Reinforcement Learning, . . .

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

However, the Environment changes frequently in practice ☹

Long Term Motivation:

Sequential decisions under noisy, partial feedback

(Contextual) Bandits, Reinforcement Learning, . . .

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

However, the Environment changes frequently in practice ☹

The Environment changes frequently in practice ☹

Many solutions so far (in machine learning):

Detect (unknown) changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle who knows changes

Suppose we had an algorithm `alg` which works well in stationary environments.

Restarting Oracle Strategy:

Main Question: can we mimic restarting oracle without knowledge of changes?

The Environment changes frequently in practice ☹

Many solutions so far (in machine learning):

Detect (unknown) changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle who knows changes

Suppose we had an algorithm `alg` which works well in stationary environments.

Restarting Oracle Strategy:

Main Question: can we mimic restarting oracle without knowledge of changes?

The Environment changes frequently in practice ☹

Many solutions so far (in machine learning):

Detect (unknown) changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle who knows changes

Suppose we had an algorithm `alg` which works well in stationary environments.

Restarting Oracle Strategy:

Main Question: can we mimic restarting oracle without knowledge of changes?

The Environment changes frequently in practice ☹

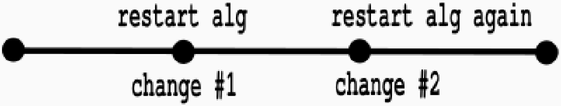
Many solutions so far (in machine learning):

Detect (unknown) changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle who knows changes

Suppose we had an algorithm `alg` which works well in stationary environments.

Restarting Oracle Strategy:



Main Question: can we mimic restarting oracle without knowledge of changes?

The Environment changes frequently in practice ☹

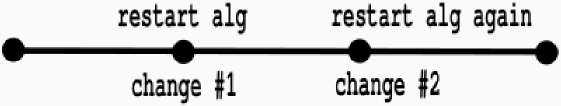
Many solutions so far (in machine learning):

Detect (unknown) changes quickly, and **restart** learning process ...

Usual Guarantees: Compete with *Restarting* Oracle who knows changes

Suppose we had an algorithm `alg` which works well in stationary environments.

Restarting Oracle Strategy:



Main Question: can we mimic restarting oracle without knowledge of changes?

What was known:

Theoretical and Adaptive State-of-the-art (Wei & Luo, '21)

There's a blackbox procedure to convert a good stationary algorithm into a good non-stationary algorithm which matches guarantees of restarting oracle *without oracle knowledge of changepoints*.

Mostly Open Questions:

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

What was known:

Theoretical and Adaptive State-of-the-art (Wei & Luo, '21)

There's a blackbox procedure to convert a good stationary algorithm into a good non-stationary algorithm which matches guarantees of restarting oracle *without oracle knowledge of changepoints*.

Mostly Open Questions:

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

Changes in Reward Y_t distribution

Non-Stationary Bandits

- **At time t**: select $a \in [K]$, observe random reward $Y_t(a)$, 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 $V \doteq \sum_t \max_a \lvert\mu_{t+1}(a) - \mu_t(a)\rvert$	$V^{1/3}T^{2/3}$ [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 Bandits

- **At time t**: select $a \in [K]$, observe random reward $Y_t(a)$, 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 $V \doteq \sum_t \max_a \lvert\mu_{t+1}(a) - \mu_t(a)\rvert$	$V^{1/3}T^{2/3}$ [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 Bandits

- **At time t**: select $a \in [K]$, observe random reward $Y_t(a)$, 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 $V \doteq \sum_t \max_a \lvert\mu_{t+1}(a) - \mu_t(a)\rvert$	$V^{1/3}T^{2/3}$ [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 Bandits

- **At time t**: select $a \in [K]$, observe random reward $Y_t(a)$, 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 $V \doteq \sum_t \max_a \lvert\mu_{t+1}(a) - \mu_t(a)\rvert$	$V^{1/3}T^{2/3}$ [Besbes et al. 14]	**Yes** [Chen et al. 19]

What we show:
A much weaker notion of change admits adaptivity ...

Main Results for Non-Stationary Bandits:

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

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

Adaptive Rates (unknown parameters)

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

- Always faster than $\sqrt{ST} \ll \sqrt{LT}$, and faster than $V^{1/3}T^{2/3}$.
- Adaptive: no algorithm knowledge of non-stationarity required.

Main Results for Non-Stationary Bandits:

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

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

Adaptive Rates (unknown parameters)

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

- Always faster than $\sqrt{ST} \ll \sqrt{LT}$, and faster than $V^{1/3}T^{2/3}$.
- Adaptive: no algorithm knowledge of non-stationarity required.

Main Results for Non-Stationary Bandits:

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

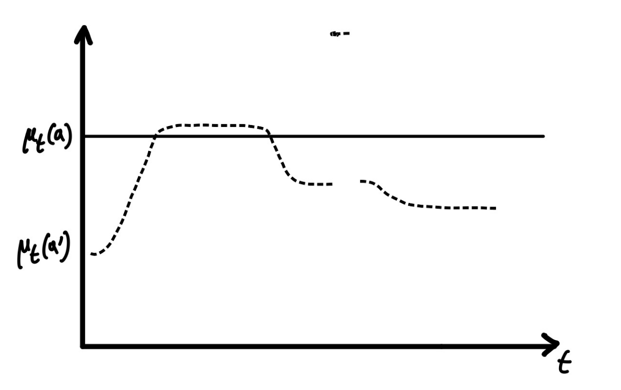
Best-arm-switches, or even large TV, *can be ignored*.

Adaptive Rates (unknown parameters)

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

- Always faster than $\sqrt{ST} \ll \sqrt{LT}$, and faster than $V^{1/3}T^{2/3}$.
- Adaptive: no algorithm knowledge of non-stationarity required.

Intuition: various changes can safely be ignored

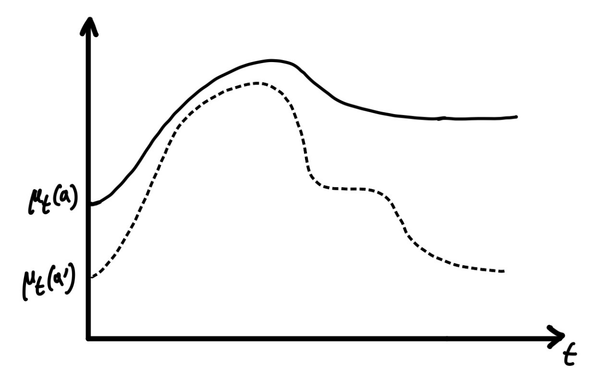


Best arm changes may not be “significant”.

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

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

Intuition: various changes can safely be ignored

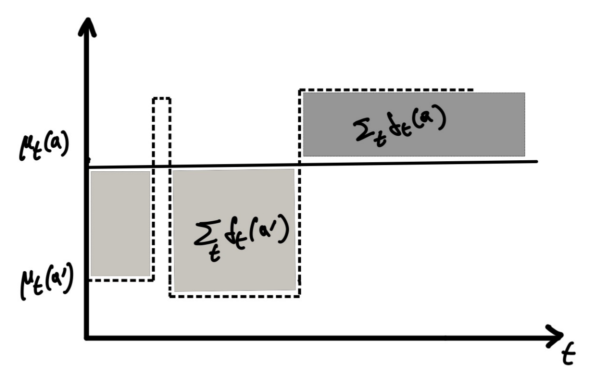


Large $V \doteq \sum_t \max_a |\mu_{t+1}(a) - \mu_t(a)|$ **may not be "significant".**

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

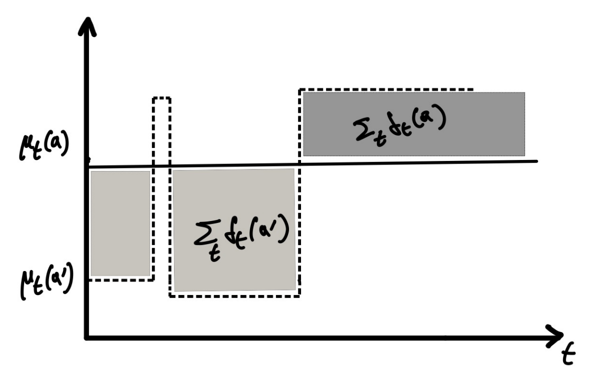
We may have $\mathbf{R_T} \lesssim \sqrt{T}$ while $V^{1/3}T^{2/3} \approx T$

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 when no safe arm is 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 $\mathbb{E}[\mathbf{R_T}] \lesssim \sum_{\mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$

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 when no safe arm is 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 $\mathbb{E}[\mathbf{R_T}] \lesssim \sum_{\mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

A Significant Phase ends when no safe arm is 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 $\mathbb{E}[\mathbf{R_T}] \lesssim \sum_{\mathcal{P}_i} \sqrt{|\mathcal{P}_i|}$

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

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

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

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

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

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

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

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

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

Broader Research Tracking Significant Changes in...

- Multi-armed bandits (w/ Samory Kpotufe, *COLT '22*).
- Non-parametric contextual bandits (w/ Samory Kpotufe, *NeurIPS '23*).
- Contextual bandits with covariate shift (w/ Samory Kpotufe, *ALT '21*).
- Dueling bandits with...
  - Condorcet winner (Buening & Saha, *AISTATS '23*)*
  - Condorcet winner + SST/STI (w/ Arpit Agarwal, *NeurIPS '23*).
  - Borda winner (w/ Arpit Agarwal, *TMLR '25*).
- Smoothly-varying non-stationary bandits (*SIMODS, '25*).
- Infinite-Arm Bandits with reservoir rewards (w/ Jung-hun Kim, *ICML '25*).
- Lipschitz Infinite-Arm Bandits (Nguyen et al., *NeurIPS '25*)*
- Online Outlier Detection (with Samory Kpotufe, *to be submitted*).
- Two-player games and multi-agent RL (*ongoing...*)

* Works of others

Summary

Main Design Element:

Track most severe changes, otherwise keep learning

Requires understanding the severity of changes (depends on problem nuances, not blackbox) ...

Thanks!

Summary

Main Design Element:

Track most severe changes, otherwise keep learning

Requires understanding the severity of changes (depends on problem nuances, not blackbox) ...

Thanks!

Summary

Main Design Element:

Track most severe changes, otherwise keep learning

Requires understanding the severity of changes (depends on problem nuances, not blackbox) ...

Thanks!