ANACONDA: An Improved Dynamic Regret Algorithm for Adaptive Non-Stationary Dueling Bandits

Thomas Kleine Buening [1]   Aadirupa Saha [2]

[1]University of Oslo   [2]TTIC / Apple ML Research

Non-Stationary Dueling Bandits

- Sequence of preference matrices $\mathbf{P}_1, \dots, \mathbf{P}_T \in [0,1]^{K\times K}$ with $\mathbf{P}_t(a,b) = 1 - \mathbf{P}_t(b,a)$ and $\mathbf{P}_t(a,a) = 1/2$.
- The stochastic (i.e., stationary) dueling bandit problem is recovered in the special case where $\mathbf{P}_1 = \dots = \mathbf{P}_T$.
- Given a preference matrix $\mathbf{P}_t$, arm $a_t^* \in [K]$ is called the **Condorcet Winner** of $\mathbf{P}_t$ if $\mathbf{P}_t(a_t^*, b) > 1/2$ for all $b \neq a_t^*$.
- Every round $t \in [T]$:
  - select a pair of actions $(a_t, b_t) \in [K] \times [K]$
  - observe preference feedback $o_t(a_t, b_t) \sim \mathrm{Ber}(\mathbf{P}_t(a_t, b_t))$
- Preference-strength of arm a over arm b in round t:

$$\delta_t(a,b) := \mathbf{P}_t(a,b) - 1/2.$$

Regret Objective: Dynamic Regret

$$\mathrm{DR}(T) := \sum_{t=1}^{T} \frac{\delta_t(a_t^*, a_t) + \delta_t(a_t^*, b_t)}{2}.$$

Transitivity Assumptions:

Every $\mathbf{P}_t$ satisfies for $a \succ_t b \succ_t c$:

- Strong Stochastic Transitivity (**SST**): $\delta_t(a,c) \geq \delta_t(a,b) \vee \delta_t(b,c)$.
- Stochastic Triangle Inequality (**STI**): $\delta_t(a,c) \leq \delta_t(a,b) + \delta_t(b,c)$.

Prior Work

Main limitations of prior work [1, 2]:

- pessimistic notions of non-stationarity.
- non-adaptive parameter tuning, i.e., require knowledge of the number of preference changes in advance.

Research Questions

Q1. Can we guarantee low dynamic regret for **meaningful notions of non-stationarity**?

Q2. Can we achieve near-optimal regret **adaptively**, without prior knowledge of the underlying non-stationarity?

Notions of Non-Stationarity

① Preference Switches (weak)

$$S^{\mathrm{P}} := \sum_{t=2}^{T} \mathbf{1}\{\mathbf{P}_t \neq \mathbf{P}_{t-1}\}$$

② Condorcet Winner Switches (strong)

$$S^{\mathtt{CW}} := \sum_{t=2}^{T} \mathbf{1}\{a_t^* \neq a_{t-1}^*\}$$

③ Significant Condorcet Winner Switches (stronger)

Let $\nu_0 := 1$ and define ν_{i+1} recursively as the first round in $[\nu_i, T)$ such that for all arms $a \in [K]$ there exist rounds $\nu_i \leq s_1 < s_2 < \nu_{i+1}$ such that $\sum_{t=s_1}^{s_2} \delta_t(a_t^*, a) \geq \sqrt{K(s_2 - s_1)}$. Let $\tilde{S}^{\mathtt{CW}}$ denote the number of such rounds $\nu_1, \dots, \nu_{\tilde{S}^{\mathtt{CW}}}$.

④ Total Variation (weak)

$$V := \sum_{t=2}^{T} \max_{a,b\in[K]} |\mathbf{P}_t(a,b) - \mathbf{P}_{t-1}(a,b)|$$

⑤ Condorcet Winner Variation (strong)

$$\tilde{V} := \sum_{t=2}^{T} \max_{a\in[K]} |\mathbf{P}_t(a_t^*,a) - \mathbf{P}_{t-1}(a_t^*,a)|$$

Observation: $\tilde{S}^{\mathtt{CW}} \leq S^{\mathtt{CW}} \leq S^{\mathrm{P}}$ and $\tilde{V} \leq V$.

Overview of Results

Algorithm	DR(T)	Notion	Adaptive?	SST&STI?
ANACONDA	$\tilde{O}(K\sqrt{S^{\mathtt{CW}}T})$	②	yes	no
ANACONDA	$\tilde{O}(K\sqrt{\tilde{S}^{\mathtt{CW}}T})$	③	yes	yes
ANACONDA	$\tilde{O}(\tilde{V}^{1/3}(KT)^{2/3})$	⑤	yes	yes
[3]	$\tilde{O}(\sqrt{K\tilde{S}^{\mathtt{CW}}T})$	③	yes	yes
[2]	$\tilde{O}(\sqrt{KS^{\mathrm{P}}T})$	①	no	no
[2]	$\tilde{O}((KV)^{1/3}T^{2/3})$	④	no	no
[1]	$\tilde{O}(K\sqrt{S^{\mathrm{P}}T})$	①	no	no

Lower Bounds: $\Omega(\sqrt{KS^{\mathtt{CW}}T})$ and $\Omega((K\tilde{V})^{1/3}T^{2/3})$. Recently, [3] showed that SST and STI are necessary conditions in order to achieve $O(\sqrt{K\tilde{S}^{\mathtt{CW}}T})$ regret. In fact, there exists a family of problem instances such that $\tilde{S}^{\mathtt{CW}} = 0$, but no algorithm can achieve $o(T)$ regret.

Algorithm

Gap Estimates: Importance weighted estimates of $\delta_t(a,b)$:

$$\hat{\delta}_t(a,b) = |\mathcal{A}_t|^2 \mathbf{1}_{\{a_t=a, b_t=b\}} o_t(a,b) - 1/2. \qquad (1)$$

Elimination Rule: Eliminate an arm $a \in [K]$ in episode ℓ and round t if there exist rounds $t_\ell \leq s_1 < s_2 \leq t$ such that

$$\max_{a'\in[K]} \sum_{t=s_1}^{s_2} \hat{\delta}_t(a', a) > C \log(T) K \sqrt{(s_2 - s_1) \vee K^2}. \qquad (2)$$

Algorithm 1 ANACONDA: Adaptive Non-stationAry CONdorcet Dueling Algorithm

input: horizon T
$t \leftarrow 1$
while $t \leq T$ **do**
  $t_\ell \leftarrow t$
  $\mathcal{A}_{\text{good}} \leftarrow [K]$
  for $m \in \{2, \dots, 2^{\lceil \log(T) \rceil}\}$ and $s \in \{t_\ell + 1, \dots, T\}$ **do**
    Sample $B_{s,m} \sim \mathrm{Bern}\left(\frac{1}{\sqrt{m(s-t_\ell)}}\right)$
  Run CondaLet$(t_\ell, T + 1 - t_\ell)$

Algorithm 2 CondaLet(t_0, m_0)

input: scheduled time t_0, duration m_0, replay schedule $\{B_{s,m}\}_{s,m}$
initialize: $t \leftarrow t_0$, $\mathcal{A}_t \leftarrow [K]$
while $t \leq T$ and $t \leq t_0 + m_0$ and $\mathcal{A}_{\text{good}} \neq \emptyset$ **do**
  Play arm-pair $(a_t, b_t) \in \mathcal{A}_t$ with each arm being selected with probability $1/|\mathcal{A}_t|$
  $\mathcal{A}_{\text{good}} \leftarrow \mathcal{A}_{\text{good}} \setminus \{a \in [K] \colon \exists [s_1, s_2] \subseteq [t_\ell, t)$ s.t. (2) holds$\}$
  $\mathcal{A}_{\text{local}} \leftarrow \mathcal{A}_t$
  $t \leftarrow t + 1$
  if $\exists m$ such that $B_{t,m} = 1$ **then**
    Run $\mathtt{CondaLet}(t, m)$ with $m = \max\{m \in \{2, \dots, 2^{\lceil \log(T) \rceil}\} \colon B_{t,m} = 1\}$
  $\mathcal{A}_t \leftarrow \mathcal{A}_{\text{local}} \setminus \{a \in [K] \colon \exists [s_1, s_2] \subseteq [t_0, t)$ s.t. (2) holds$\}$

Challenges under Preference-Based Feedback

- We generally cannot decompose regret, as is done in non-stationary MAB, due to the lack of transitivity.
- It is more difficult to detect a bad arm, since an arm can beat every arm except the best arm (by a large margin).

Future Work

- Other solution concepts, e.g., Borda scores, Copeland winner, von Neumann winner

References

[1] P. Kolpaczki, V. Bengs, and E. Hüllermeier. Non-stationary dueling bandits. *arXiv preprint arXiv:2202.00935*, 2022.

[2] A. Saha and S. Gupta. Optimal and efficient dynamic regret algorithms for non-stationary dueling bandits. In *International Conference on Machine Learning*, pages 19027–19049. PMLR, 2022.

[3] J. Suk and A. Agarwal. When can we track significant preference shifts in dueling bandits? *arXiv preprint:2302.06595*, 2023.