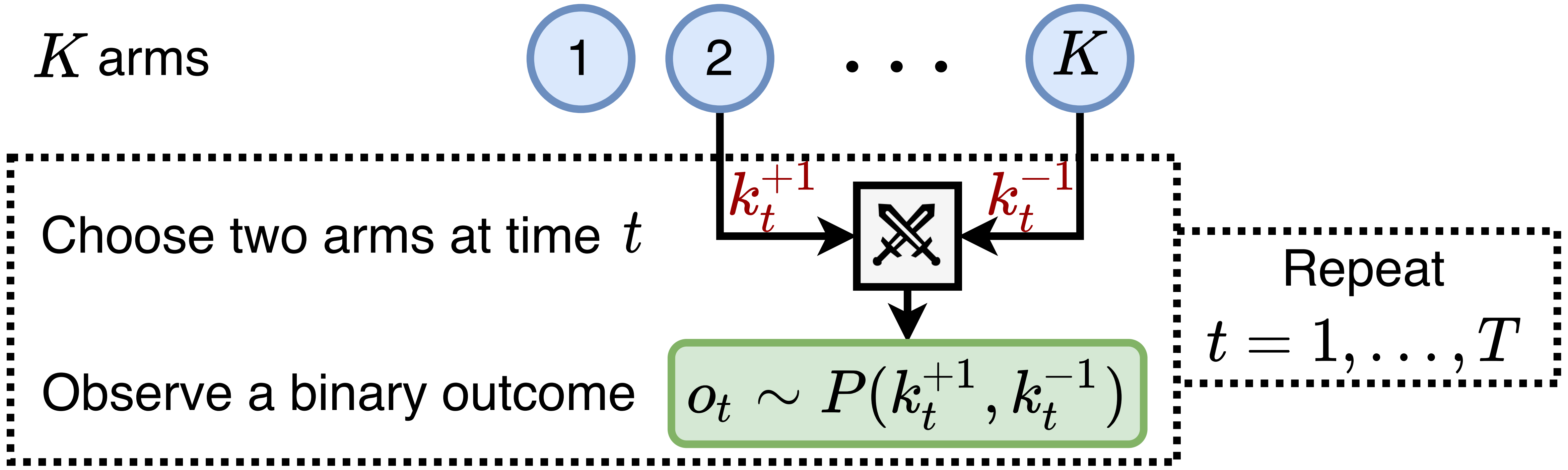
Optimal and Efficient Dynamic Regret Algorithms for Non-Stationary Dueling Bandits

Aadirupa Saha[1] and Shubham Gupta[2]
(Equal contribution)

[1]Toyota Technological Institute at Chicago
aadirupa@ttic.edu

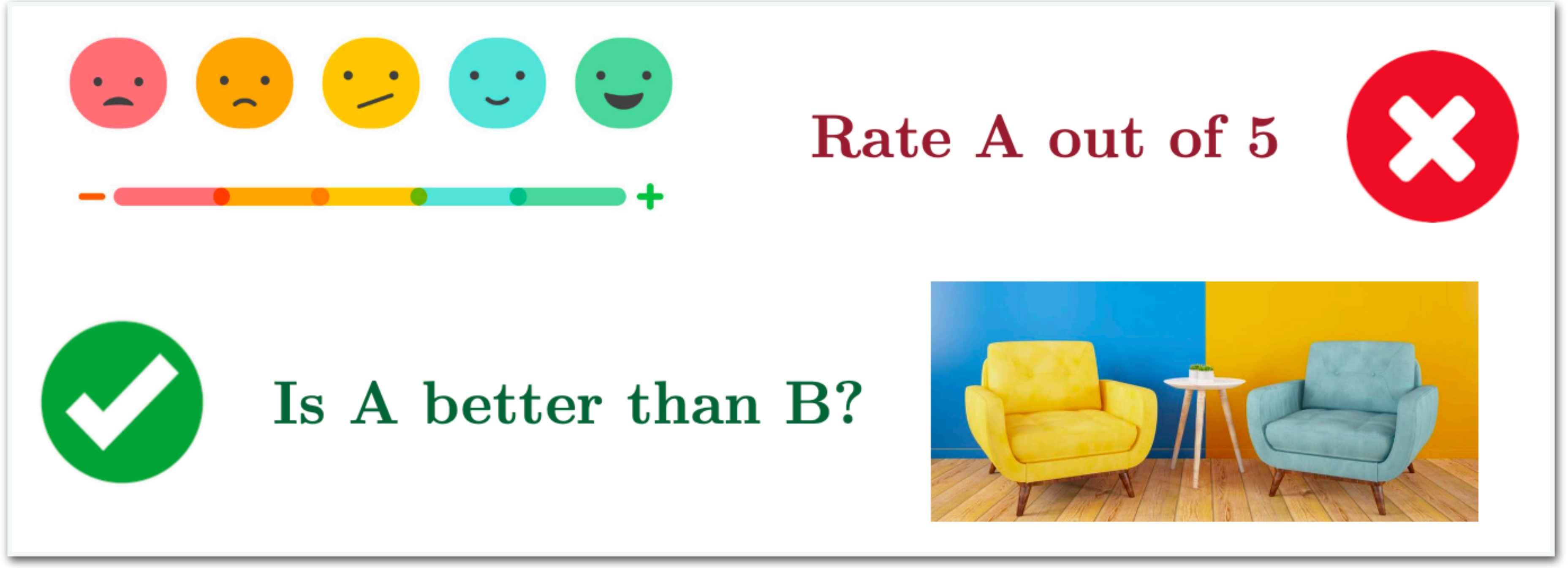
[2]IBM Research, Paris-Saclay, France
shubham.gupta1@ibm.com

Dueling Bandits

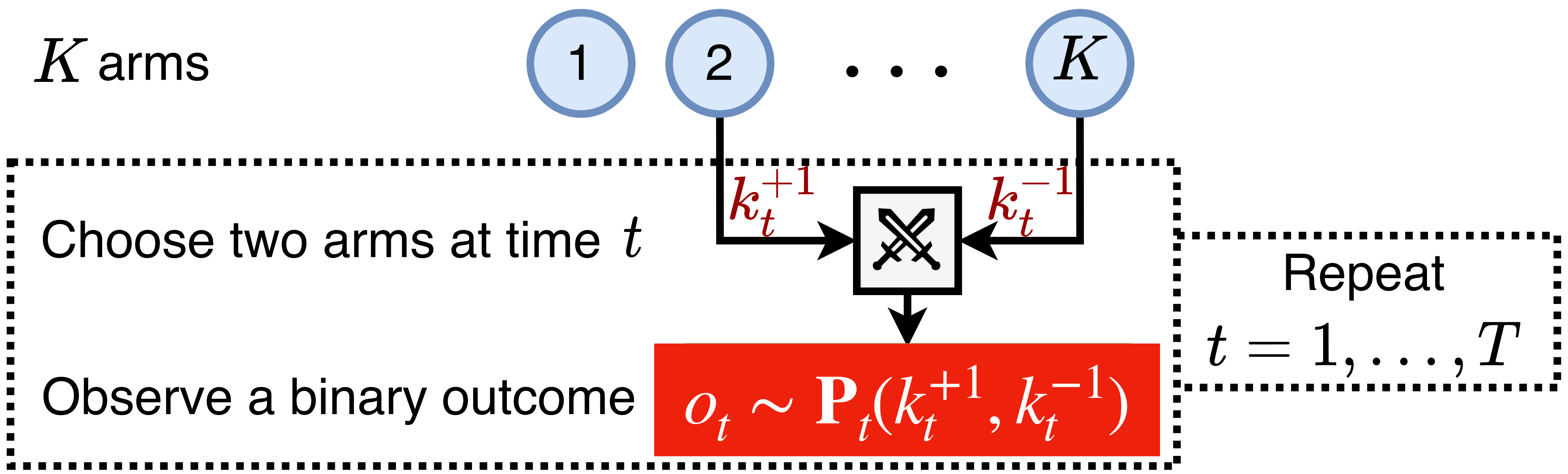


Preference Matrix $\mathbf{P}$

	1	2	$\dots$	K
1	0.5	0.3	$\dots$	0.8
2	0.7	0.5	$\dots$	0.9
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
K	0.2	0.1	$\dots$	0.5

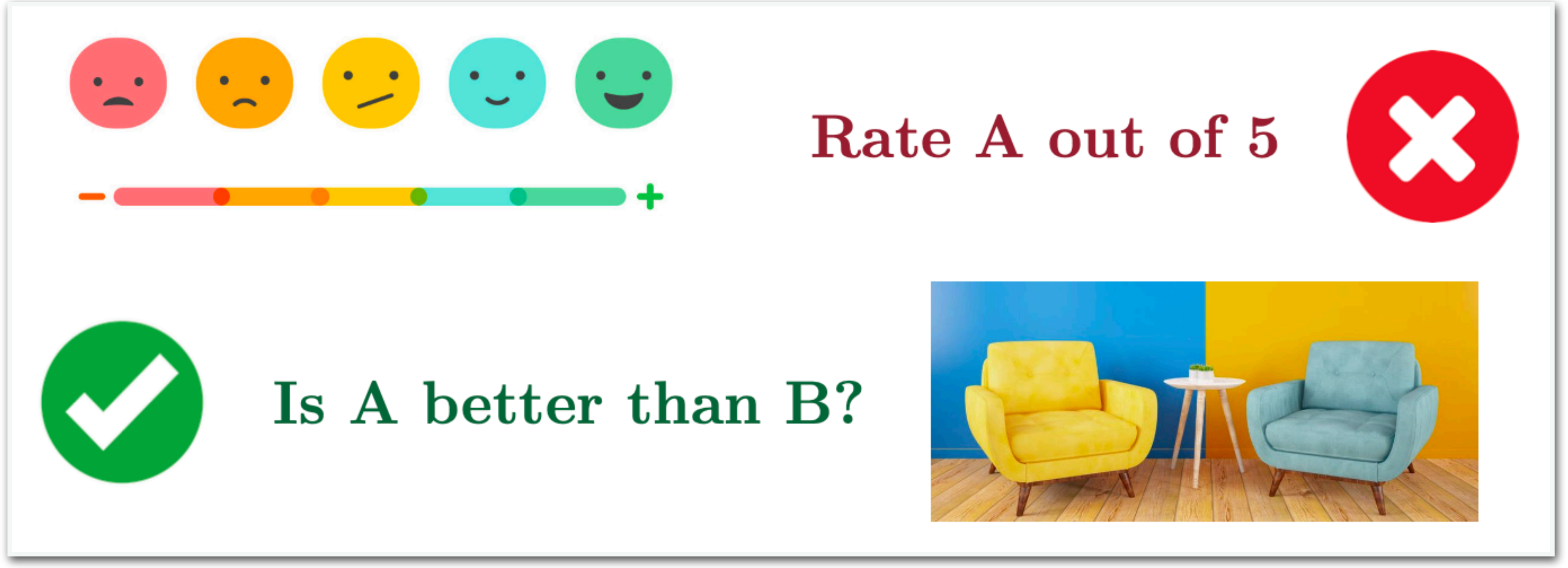


Non-Stationary Dueling Bandits

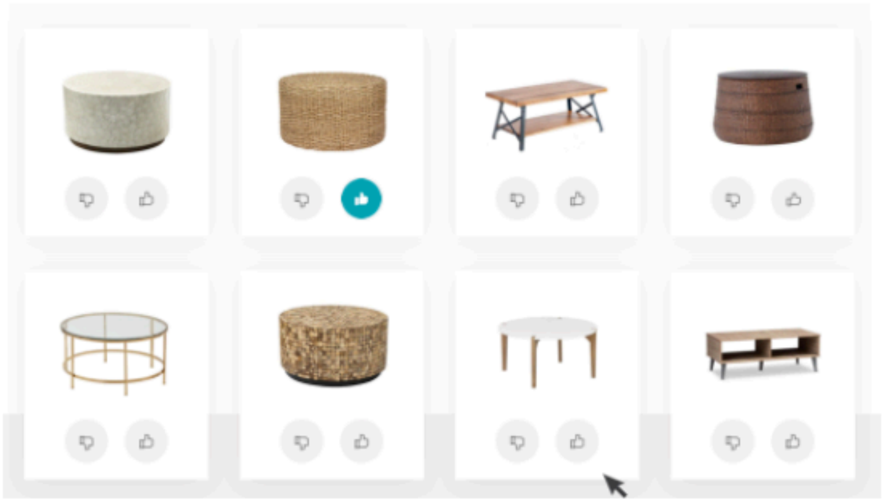


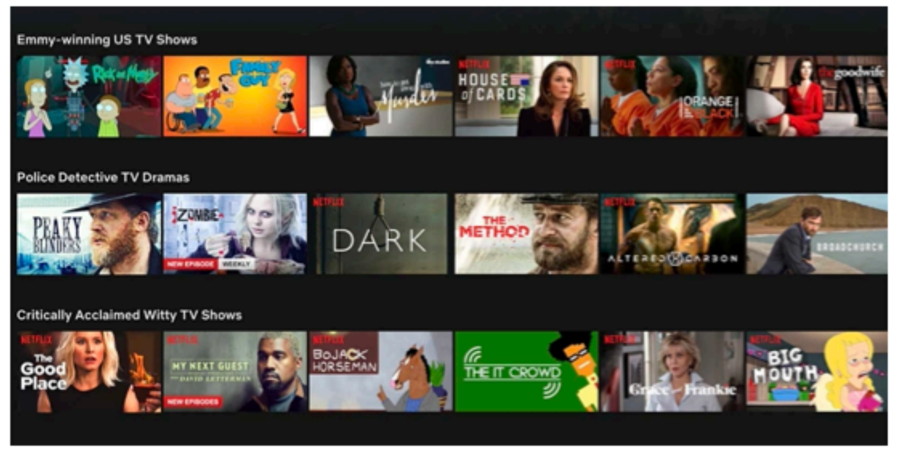
Preference Matrix $\mathbf{P}_t$

	1	2	. . .	K
1	0.5	0.3	. . .	0.8
2	0.7	0.5	. . .	0.9
⋮	⋮	⋮	⋱	⋮
K	0.2	0.1	. . .	0.5

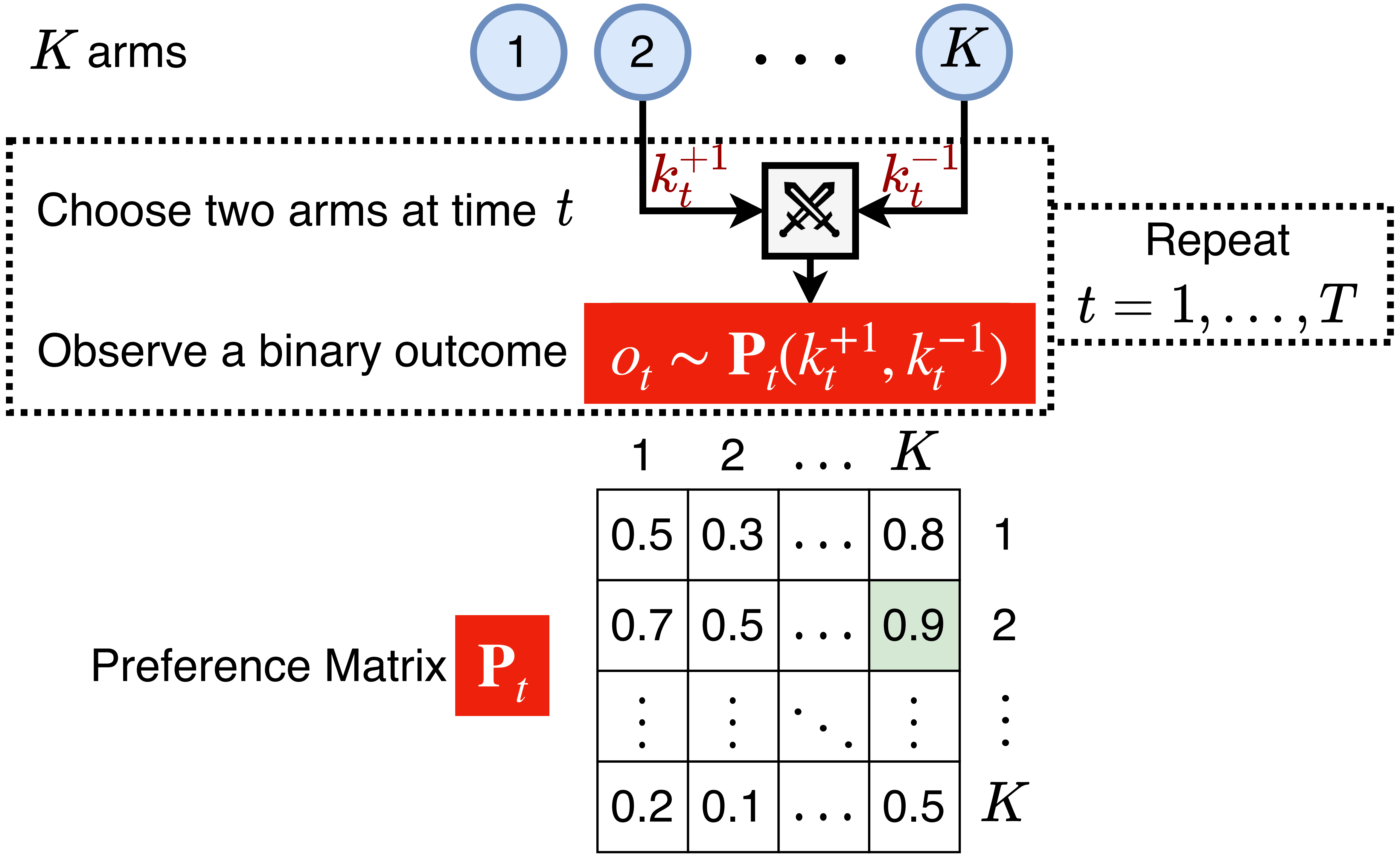


Multitude of applications





Non-Stationary Dueling Bandits



Measures of non-stationarity

- **Switching variation**

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

- **Continuous variation**

$$V_T \coloneqq \sum_{t=2}^{T} \max_{i,j} |P_t(i,j) - P_{t-1}(i,j)|$$

Static and Dynamic Regret

Static regret: Regret with respect to a fixed arm

$$\mathrm{SR}_T = \max_{i \in [K]} \sum_{t=1}^{T} \underbrace{\frac{\left[P_t(i, k_t^{+1}) - 0.5\right] + \left[P_t(i, k_t^{-1}) - 0.5\right]}{2}}_{\text{Avg. strength of arm } i \text{ w.r.t arms } k_t^{+1} \text{ and } k_t^{-1}}$$

$t = 1,2,\ldots,T$

$t = 1$

$t = 2$

$\cdots$

$t = T$

Dynamic regret: Regret with respect to **ANY** sequence of arms $i^T = (i_1, \ldots, i_T)$

$$\mathrm{DR}_T(i^T) = \sum_{t=1}^{T} \frac{\left[P_t(i_t, k_t^{+1}) - 0.5\right] + \left[P_t(i_t, k_t^{-1}) - 0.5\right]}{2}$$

Prior Works

Stochastic preferences

Yue et al. (2009)
Yue and Joachims (2011)
Zoghi et al. (2014)
Jamieson et al. (2015)
Komiyama et al. (2015)
Wu and Liu (2016)
Kumagai (2017)
Saha and Gopalan (2020);
. . .
[Survey] Bengs et al. (2021)

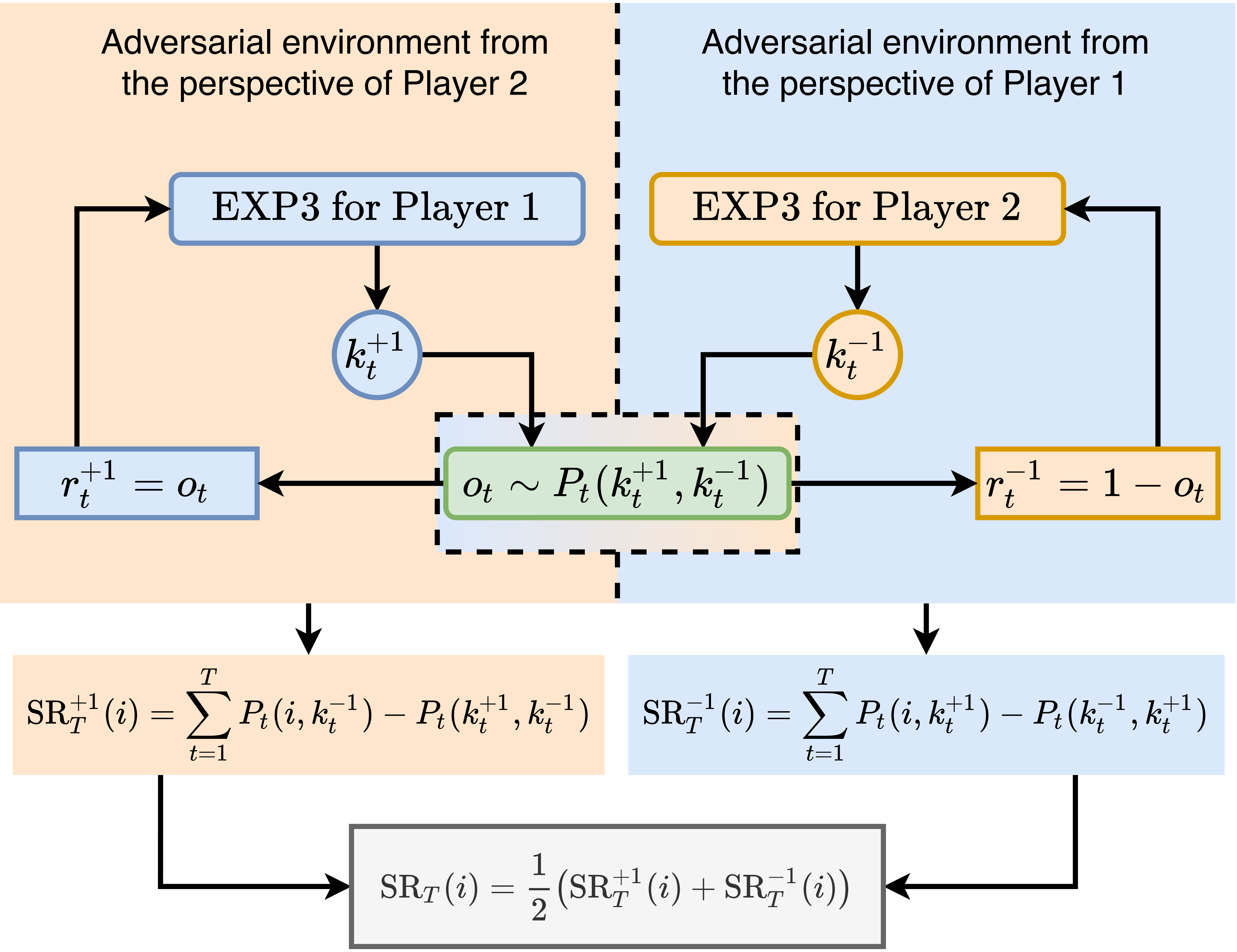
Adversarial preferences

Ailon et al. (2014)
Gajane et al. (2015)
Saha et al. (2021)

(Only static regret analysis)

No dynamic regret analysis for dueling bandits!

Key Idea : Regret Decomposition



$$\mathrm{SR}_T = \max_{i \in [K]} \sum_{t=1}^{T} \underbrace{\frac{\left[P_t(i, k_t^{+1}) - 0.5\right] + \left[P_t(i, k_t^{-1}) - 0.5\right]}{2}}_{\text{Avg. strength of arm } i \text{ w.r.t arms } k_t^{+1} \text{ and } k_t^{-1}}$$

Results

	Upper bounds		Lower bounds	
Static Regret: [Adversarial]	$O\left(\sqrt{KT}\ \ln\frac{K}{\delta}\right)$		$\Omega\left(\sqrt{KT}\right)$	[Gajane et al. 2015]
Dynamic Regret:	$O\left(\sqrt{SKT}\ \ln\frac{KT}{\delta}\right)$	Switching Variation	$\Omega\left(\sqrt{SKT}\right)$	
	$O\left((KV_T)^{1/3}T^{2/3}\ \ln\frac{KT}{\delta}\right)$	Continuous Variation	$\Omega\left((KV_T)^{1/3}T^{2/3}\right)$	

Additional Results in the Paper

- What happens when S is not known in advance?
- Dynamic regret analysis for Borda scores
- Numerical studies

Thank you!

Questions?

- ✓ Non-stationary dueling bandits problem
- ✓ Two measures of non-stationarity
- ✓ Regret decomposition idea
- ✓ Near optimal static (adversarial) and dynamic regret guarantees
- ✓ Lower bounds on dynamic regret