Efficient and Near-Optimal Algorithm for Contextual Dueling Bandits with Offline Regression Oracles

Aadirupa Saha (UIC), Robert E Schapire (MSR)

Corresponding: aadirupa.saha@gmail.com

39th Annual Conference on Neural Information Processing Systems

Learning from Preference (Relative) Feedback

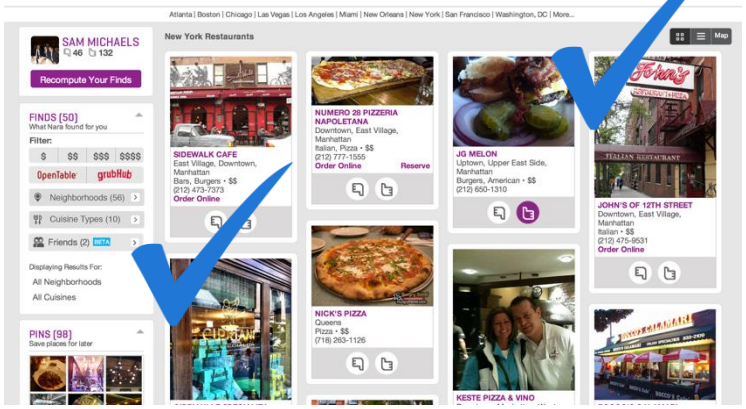
--- *How much you score it out of 5?*

Rankings (Relative)

--- *Do you like movie A over B?*

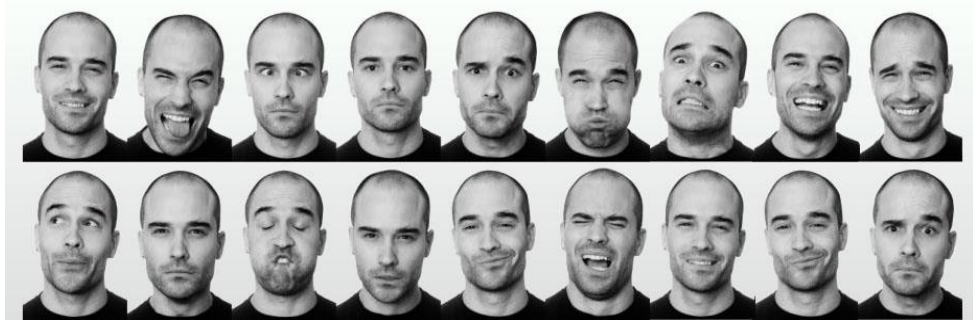
Multitude of Applications

Learning from Preferences



Recommender Systems

Google
YAHOO!
Ask.com
msn
bing
About

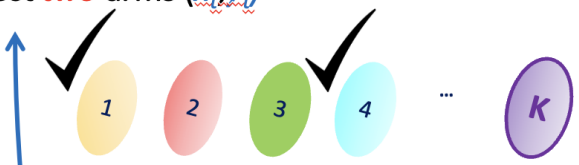


Performance Measure?

Problem setup: Dueling Bandits

At round $t = 1,2,3 \ldots T$

Select *two* arms (x_t, y_t)



repeat

Observe (noisy) *comparison* $o_t \in \{0,1\}$ ~ P:=

Yue et al. The K-armed dueling bandits problem. Jour. Of CSS 2012.

Yue and Joachims. Beat the mean bandit. ICML 2011.

Objective: **Find the BEST arm(s)**
[[Regret minimization (or)
PAC best-arm identification]]

Regret :=

$$\sum_{t=1}^{T} \frac{\left\{\left[P(i^*, x_t) - \frac{1}{2}\right] + \left[P(i^*, y_t) - \frac{1}{2}\right]\right\}}{2}$$

Condorcet winner (i^*)

P: Preference Matrix (K = 5)

	1	2	3	4	5
1	0.5	0.53	0.54	0.56	0.6
2	0.47	0.5	0.53	0.58	0.61
3	0.46		0.5	0.54	0.57
4	0.44	.42		0.5	0.51
5	0.4	0.39	0.43	0.49	0.5

Limitations of the Existing Framework

Regret :=

$$\sum_{t=1}^{T} \frac{\left\{\left[P(i^*, x_t) - \frac{1}{2}\right] + \left[P(i^*, y_t) - \frac{1}{2}\right]\right\}}{2}$$

1. Fixed benchmark (i^*) over T
2. Stationary Preferences over T
3. No contextualization!!

Saha & Gupta, Optimal and Efficient Dynamic Regret Algorithms for Non-Stationary Dueling Bandits. ICML'22

Best-Response Regret!

A Strong Notion of Regret!

Standard Regret := $\max_{i^* \in [K]} \sum_{t=1}^{T} \frac{\left\{\left[P_t(i^*, x_t) - \frac{1}{2}\right] + \left[P_t(i^*, y_t) - \frac{1}{2}\right]\right\}}{2}$

Fixed benchmark over T

Strongest opponent

Best-response Regret (BR-Reg$_T$) := $\sum_{t=1}^{T} \max_{i_t^* \in [K]} \left[\frac{\left\{\left[P_t(i_t^*, x_t) - \frac{1}{2}\right] + \left[P_t(i_t^*, y_t) - \frac{1}{2}\right]\right\}}{2}\right]$

$\tilde{P} = 2P - 1$

	1	2	3	4	5
1	0.5	0.53	0.54	0.56	0.6
2	0.47	0.5	0.53	0.58	0.61
3	0.46	0.47	0.5	0.54	0.57
4	0.44	0.42	0.46	0.5	0.51
5	0.4	0.39	0.43	0.49	0.5

$P \in [0,1]^{K \times K}$

(0.06,-0.06)

	1	2	3	4	5
1	0	0.06	0.08	0.12	0.20
2	-0.06	0	0.06	0.16	0.22
3	-0.08	-0.06	0	0.08	0.04
4	-0.12	-0.16	-0.08	0	0.02
5	-0.20	-0.22	-0.07	-0.02	0

$\tilde{P} \in [-1,1]^{K \times K}$

Preference matrix as a 2 player Zero-Sum Game

However...

Rock Paper Scissor

$$\boldsymbol{P} := \begin{pmatrix} 0 & 1 & -1 \\ -1 & 0 & 1 \\ 1 & -1 & 0 \end{pmatrix}$$

Leanrer's choice (x_t, y_t)	Adversary's choice (i^*)
(1,2)	(1)
(2,3)	(2)
(3,1)	(3)

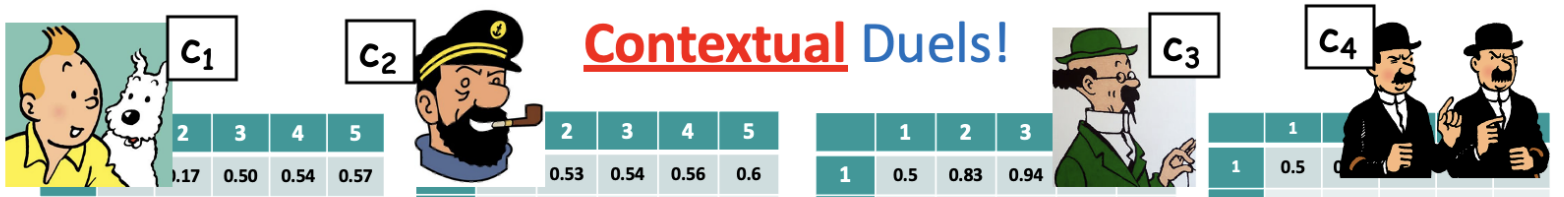
Learner is always forced to suffer a loss of ½ $\rightarrow \mathrm{Reg_T} = O(T)$!
(too bad)

... hence the Learner must randomize!

Best-Response Regret

$$\sum_{t=1}^{T} \max_{i_t^* \in [K]} E_{(x_t, y_t) \sim \boldsymbol{p}_t} \left[\frac{\{[\tilde{P}_t(i_t^*, x_t)] + [\tilde{P}_t(i_t^*, y_t)]\}}{2}\right]$$

Personalization: Contextual DB



	1	2	3	4	5
1	0.5	0.17	0.50	0.54	0.57
2	0.83	0.5	0.93	0.98	0.91
3	0.50	0.07	0.5	0.60	0.53
4	0.44	0.02	0.40	0.5	0.59
5	0.43	0.09	0.47	0.41	0.5

$f_1 = f(c_1) = P_1$ (t = 1)

	1	2	3	4	5
1	0.47	0.53	0.54	0.56	0.6
2	0.47	0.5	0.93	0.98	0.91
3	0.46	0.07	0.5	0.54	0.57
4	0.44	0.02	0.46	0.5	0.51
5	0.4	0.09	0.43	0.49	0.5

$f_2 = f(c_2) = P_2$ (t = 2)

	1	2	3	4	5
1	0.5	0.83	0.94		
2	0.17	0.5	0.43	0.18	0.71
3	0.06	0.57	0.5	0.14	0.07
4	0.84	0.02	0.86	0.5	0.51
5	0.94	0.29	0.93	0.49	0.5

$f_3 = f(c_3) = P_3$ (t = 3)

	1	2	3	4	5
1	0.5				
2	0.21	0.5	0.43	0.08	0.27
3	0.02	0.57	0.5	0.14	0.07
4	0.84	0.92	0.86	0.5	0.51
5	0.84	0.73	0.93	0.49	0.5

$f_4 = f(c_4) = P_4$ (t = 4) ....

Assume: f: Context ↦ Preference matrix; $f \in \mathcal{M}$ (f: Unknown, $\mathcal{M}$: Known)

(Given $\mathcal{C}$: Context Class)

π: Context ↦ Arm

"Contextual" Regret:

Policy Regret

$$\max_{\{\pi^* \in \Pi\}} \sum_{t=1}^{T} E_{(x_t, y_t) \sim \boldsymbol{p_t}} \left[\frac{\{[f_t(\pi^*(c_t), x_t)] + [f_t(\pi^*(c_t), y_t)]\}}{2} \right]$$

Contextual Best-Response DB

Best Response Regret

$$\sum_{t=1}^{T} \max_{i_t^* \in [K]} E_{(x_t, y_t) \sim \boldsymbol{p_t}} \left[\frac{\{[f_t(i_t^*, x_t)] + [f_t(i_t^*, y_t)]\}}{2} \right]$$

Strongest opponent

where $\boldsymbol{p_t} \in \Delta_{K \times K}$

Dudik et al. (2015)

- Suboptimal
- (or) Runtime in-efficient
- But not BOTH

Dudik, Hofmann, Schapire, Slivkins. Contextual Dueling Bandits. COLT'15

Saha, Krishnamurthy. Efficient and Optimal Algorithms for Contextual Dueling Bandits under Realizability. ALT'22

Offline vs Online Regression Oracle

Offline Regression Oracle

Given a function class $\mathcal{F} \subseteq \{f \mid f : \mathcal{Z} \to \mathcal{Y}\}$ and i.i.d. data
$D_n = \{(z_i, y_i)\}_{i=1}^n,\ (z_i, y_i) \sim \mathcal{D}$,
the offline oracle outputs
$\hat{f}_n \leftarrow \mathrm{OffReg}_{\mathcal{F}}(D_n)$.
For any $\delta \in (0, 1]$, with probability at least $1 - \delta$:

$$\mathbb{E}_{(z,y)\sim\mathcal{D}}\left[(\hat{f}_n(z) - y)^2 - \inf_{f\in\mathcal{F}}(f(z) - y)^2\right] \le \mathcal{E}_{\mathrm{off},\mathcal{F}}(\delta, n).$$

Online Regression Oracle

For the same class
$\mathcal{F} \subseteq \{f \mid f : \mathcal{Z} \to \mathcal{Y}\}$,
the online oracle receives an arbitrary sequence
$\{(z_t, y_t)\}_{t=1}^n$
and guarantees:

$$\frac{1}{n}\left[\sum_{t=1}^n(\hat{y}_t(z_t) - y_t)^2 - \inf_{f\in\mathcal{F}}\sum_{t=1}^n(f(z_t) - y_t)^2\right] \le \mathcal{E}_{\mathrm{on},\mathcal{F}}(n),$$

where $\mathcal{E}_{\mathrm{on},\mathcal{F}}(n) = o(1)$.

Offline Oracle ($\mathrm{OffReg}_{\mathcal{F}}$)	Online Oracle ($\mathrm{OnReg}_{\mathcal{F}}$)
✓ Batch processing	✓ Real-time & adaptive
✓ Higher accuracy potential	✓ Stronger guarantee (adversarial-robust)
✓ Works only when test ≈ train	✓ Distribution-free (no i.i.d.)
✓ Easier to implement	✓ Can convert to offline (not vice-versa)
	✕ Harder to construct

Result (Finite Actions)

Algorithm 1 **`Double-Monster`**: Efficient Contextual Dueling Bandit with Offline Regressors

input Epoch schedule $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$. Confidence parameter δ. Tuning parameter $\gamma_1, \gamma_2, \ldots$.
Arm set: $[K]$. An instance of `OffReg` for function class $\mathcal{M}$
for epoch $m = 1, 2, \ldots$ **do**
  Let $\gamma_m = \left(\frac{K}{\sqrt{\mathcal{E}_{\text{off},\mathcal{M}}(\delta/(2m^2), \tau_{m-1} - \tau_{m-2})}} \right)$ (for epoch 1, $\gamma_1 = 1$).
  Compute $\widehat{\mathbf{M}}_m \leftarrow \texttt{OffReg}\left(\{(c_\tau, x_\tau, y_\tau), o_\tau\}_{\tau=\tau_{m-2}+1}^{\tau_{m-1}} \right)$, i.e.
  $\widehat{\mathbf{M}}_m = \arg\min_{f \in \mathcal{M}} \sum_{\tau=\tau_{m-2}+1}^{\tau_{m-1}} (f(c_\tau)[x_\tau, y_\tau] - \tilde{o}_\tau)^2$ via the **offline least squares oracle**, where $\tilde{o}_\tau = 2o_\tau - 1$, for any $\tau \in [\tau_{m-2}+1, \tau_{m-1}]$.
  for round $t = \tau_{m-1} + 1, \cdots, \tau_m$ **do**
    Observe context $c_t \in \mathcal{C}$.
    Compute $\mathbf{p}_t \sim \texttt{Cvx-Constraint-Solver}(c_t, \widehat{\mathbf{M}}_m, \gamma_m)$ s.t. $\mathbf{p}_t \in \Delta_{K \times K}$ satisfies:

$$\forall i, j \in [K] : \mathbf{E}_{x \sim \mathbf{p}_t^\ell}\left[\widehat{\mathbf{M}}_m(c_t)[i, x]\right] + \mathbf{E}_{y \sim \mathbf{p}_t^r}\left[\widehat{\mathbf{M}}_m(c_t)[j, y]\right] + \frac{2}{\gamma_m} \frac{1}{p_t(i, j)} \leq \frac{5K^2}{\gamma_m}.$$

    Sample $(x_t, y_t) \sim \mathbf{p}_t$ and observe preference feedback $o_t \sim \text{Ber}\left(\frac{f^\star(c_t)[x_t, y_t]+1}{2} \right)$.
  end for
end for

Experiments

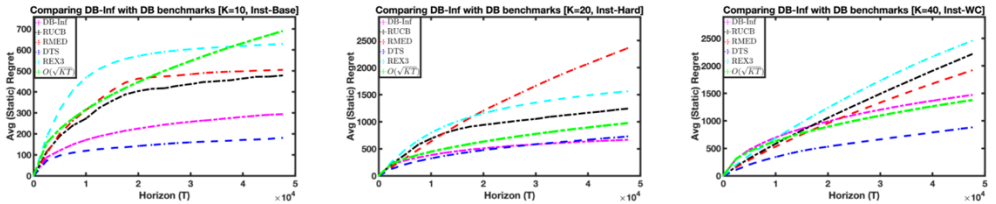


Figure 2: Comparing with Finite-armed Non-Contextual Baselines

Theorem (Regret Analysis of Algorithm 1): With an epoch schedule $\tau_1, \ldots, \tau_m$ such that $\tau_m \geq 2^m$ for $m \leq \log T$, with probability at least $(1-\delta)$ for any $\delta \in (0, 1)$ the best-response regret (BR-Reg_T) of **`Double-Monster`** after T rounds is bounded by:

$$O\left(K \sum_{m=1}^{m(T)} \sqrt{\mathcal{E}_{\text{off},\mathcal{M}}(\delta/(2m^2), \tau_{m-1} - \tau_{m-2})}(\tau_m - \tau_{m-1}) \right)$$

$$\approx O(K\sqrt{T \log(|\mathcal{M}| T \log T)})$$

for special classes!

Continuous Action Spaces

- **Setup:** Continuous decision (action) space $\mathcal{K} \subset \mathbb{R}^d$
- **Context Embedding:** $\phi^* : \mathcal{C} \to \mathbb{R}^d,\ \phi^* \in \Phi$ [Given Φ: **Embedding Class**]
- **(Expected) Action Scores:** $\mathbb{E}[s_t \mid c_t = c, \mathbf{x}_t = \mathbf{x}] = \langle \mathbf{x}, \phi^*(c) \rangle,\ \mathbf{x} \in \mathcal{K}, c \in \mathcal{C}$
- **Preference Model:** $P_t(\mathbf{x}, \mathbf{y}) = \sigma\big(\langle \phi^*(c_t),\ \mathbf{x} - \mathbf{y} \rangle\big),\ \sigma(z) = \frac{1}{1+e^{-z}},\ z \in \mathbb{R}$
  Learner observes feedback $o_t \sim \mathrm{Ber}(P_t(\mathbf{x}_t, \mathbf{y}_t))$.
- **Best-Response Regret:**

$$\text{BR-Reg}_T^{\text{cont}} = \sum_{t=1}^{T} \mathbb{E}_{x_t \sim \mathcal{D}} \left[\max_{\mathbf{a}^* \in \mathcal{K}} \langle \mathbf{x}^*, \phi^*(x_t) \rangle - \frac{1}{2} \langle \mathbf{x}_t + \mathbf{y}_t, \phi^*(c_t) \rangle \right]$$

Main Algorithm

Algorithm 2 `Double-Monster-Inf` (for Continuous Action Spaces)

input Epoch schedule $0 = \tau_0 < \tau_1 < \tau_2 < \cdots$. Confidence parameter δ. Tuning parameter $\gamma_1, \gamma_2, \ldots$.
Arm set: $\mathcal{K} \subset \mathbb{R}^d$. An instance of `OffReg` for function class Φ
for epoch $m = 1, 2, \ldots$ **do**
  Set $\gamma_m = \left(\frac{\sqrt{d}}{2\sqrt{2\mathcal{E}_{\text{off},\Phi}(\delta/(2m^2), \tau_{m-1} - \tau_{m-2})}} \right)$ (for epoch 1, $\gamma_1 = 1$).
  Compute $\phi_m \leftarrow \texttt{OffReg}\Big(\{(c_\tau, \mathbf{x}_\tau, \mathbf{y}_\tau), o_\tau\}_{\tau=\tau_{m-2}+1}^{\tau_{m-1}}\Big)$, i.e.
  $\phi_m = \arg\min_{\phi \in \Phi} \sum_{\tau=\tau_{m-2}+1}^{\tau_{m-1}} (\sigma(\langle \phi(c_\tau), \mathbf{x}_\tau - \mathbf{y}_\tau \rangle) - o_\tau)^2$ via the **offline least squares oracle**
  for round $t = \tau_{m-1} + 1, \cdots, \tau_m$ **do**
    Observe context $x_t \in \mathcal{X}$.
    Compute $\mathbf{p}_t \sim \texttt{Cont-CvxConstraint-Solver}(x_t, \phi_m, \gamma_m)$ i.e.:

$$\mathbf{p}_t \leftarrow \text{argmax}_{q \in \Delta_{\mathcal{K} \times \mathcal{K}}} \left[\langle \mathbf{ab_q}, \phi_m(c_t) \rangle + \tfrac{1}{\gamma_m} \log\det(\mathbf{H_q}) \right],$$

    where $\mathbf{ab_q} = \mathbf{E}_{(\mathbf{a},\mathbf{b}) \sim \mathbf{q}}[(\mathbf{a}+\mathbf{b})/2]$, $\mathbf{H_q} = \mathbf{E}_{(\mathbf{a},\mathbf{b}) \sim \mathbf{q}}[(\mathbf{a}-\mathbf{b})(\mathbf{a}-\mathbf{b})^\top] + c\mathbf{I}_d$.
    Sample $(\mathbf{x}_t, \mathbf{y}_t) \sim \mathbf{p}_t$ and observe preference feedback $o_t \sim \text{Ber}\Big(\sigma(\langle \phi^*(c_t), \mathbf{x}_t - \mathbf{y}_t \rangle)\Big)$.
  end for
end for

Results

Theorem (Regret Analysis of Algorithm 2): Assume $\forall c \in \mathcal{C}$, $\|\phi^*(x)\| \leq D$, for some $D > 0$. Then with an epoch schedule $\tau_1, \ldots, \tau_m$ such that $\tau_m \geq 2^m$ for $m \leq \log T$, with probability at least $(1-\delta)$ for any $\delta \in (0,1)$ the best-response regret in T rounds is bounded by:

$$O\left(\sum_{m=2}^{m(T)} \sqrt{d\,\mathcal{E}_{\text{off},\Phi}(\delta/(2m^2), \tau_{m-1}-\tau_{m-2})(\tau_m - \tau_{m-1})}\right).$$

$$\approx \tilde{O}(\sqrt{dT})$$

for special classes!

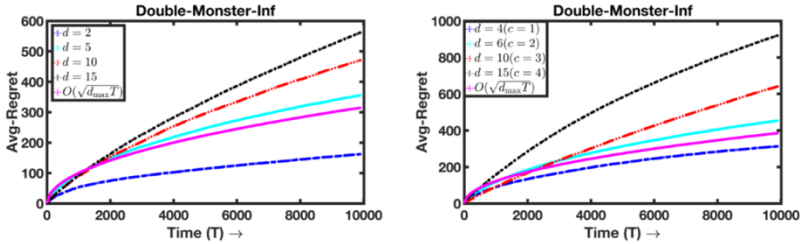


Figure 1: (Left) Inst-1 (linear transformation) and (Right) Inst-2 (quadratic transformation)

Special Classes

- **Finite class $\mathcal{M}$:** $\mathcal{E}_{\text{off},\mathcal{M}}(\delta, n) = O\left(\frac{\log(|\mathcal{M}|\, n)}{n\delta}\right)$
- **Nonparametric class (with empirical entropy $O(\varepsilon^{-p})$):**

$$\mathcal{E}_{\text{off},\mathcal{M}}(\delta, n) = O\left(n^{-2/(2+p)} \log(1/\delta)\right)$$

- **Linear predictors (d-dimensional):** $\mathcal{E}_{\text{off},\mathcal{M}}(\delta, n) = O\left(\frac{d\, \log(n/d)}{n\delta}\right)$
- **Deep MLP / ReLU nets (Sobolev smoothness β):**

$$\mathcal{E}_{\text{off},\mathcal{M}}(\delta, n) = \widetilde{O}\left(n^{-\beta/(\beta+d)}\right)$$

Paper

Poster + More

Thanks!

Follow-up/ questions?

Looking forward :)

aadirupa.saha@gmail.com

https://aadirupa.github.io/