Non-stationary Dueling Bandits for Online Learning to Rank

 S^{11} L¹, $-M^2$, Pⁱ₁ $\frac{2}{3}$, $-H^2$, $-H^3$, $\frac{1}{2}$

 1 Na $\hat{\blacktriangle}$ a K $\sqrt{\text{Lap}}$ a $\ell \sqrt{\ell}$ Novel \mathbb{S}_ℓ of \mathbb{R} of \mathbb{R} of \mathbb{R} of \mathbb{R} \mathbb{R} of $\mathbb{R$ Na_{nj}in, 210023, C_{ri}ina {lusy,zhanglj}@lamda.nju.edu.cn ² A baba G_r, Hangzhou 311121, C_ra {miaoyuan.my,yangping.yangping,yaoohu}@alibaba-inc.com

 Ab^{\bullet} ac. We sd_{ree} in a it a part (OL2R), where a parameters a parameters and \bullet eterized ranking model is optimized based on sequential feedback from \mathcal{S} . SA a aad paa pac pOL2R is papa. a madd milading problem, where are a manged sp a ar, i., ari µd a scicaa q. a**t** Will due bandits and its and its application to \mathbf{A} $e^{\frac{i}{2}\int \mathbf{y} - \mathbf{S} \, \mathbf{d} \hat{\mathbf{r}} \, \mathbf{d} \hat{\mathbf{r}}}$ in the literature, existing \mathbf{A} $\begin{array}{ccc} \bullet\quad & -\circ\bullet\end{array}$ ronments where $\begin{array}{ccc} \bullet\quad & \bullet\end{array}$ μ a μ H_{ow}ever, μ sasses in real-world in real-world OL2R a prapplications as user preference that the and so does a so does a dopped to μ **the ard The added** μ **signal rankers** μ problem, we problem, we problem, dueling bandits where the preference order over rankers is modeled by a time-variant function. We develop an efficient and adaptive method $f \mapsto \mathcal{F}$ ationary dueling bandits \mathcal{F} is the strong theoretical guarantees. The main idea of our method is to reduce the multiple of the multiple dueling banding sadir ent different different step sizes in a and different step sizes in parallel and different step sizes in parallel and e_{∞} a aa μ algorithm to dynamically combine the DBGD algo- $\dot{\mathcal{A}}$ sacquiding to the time performance. With straightforward $\mathcal{A} \rightarrow \mathcal{S}$ **c** and can also a set of a set of \mathcal{S} algorithm algorithm algorithm \mathcal{S} algorithm algorithm \mathcal{S} algorithm \mathcal{S} algorithm \mathcal{S} algorithm \mathcal{S} algorithm \mathcal{S} algorithm \mathcal{S} \mathbb{R} \mathbb{R}^s

Ke di: Online a \vec{A} a \vec{A} a \vec{A} a \vec{B} a \vec{C} badd \vec{C} s N_f \rightarrow S

1 Introduction

168 S. Luis a.

2 Problem Setup

$$
\begin{array}{ccccccccccc}\n\mathbf{w} & \mathbf{i} & \mathbf{j} & \mathbf{k} & \mathbf{k
$$

3 Method

 $I \rightarrow I$, we first review the dueling bandits gradient descent descriptions of $I \rightarrow I$ algorithm and derive its dynamic regret bound, then present our method as well as well as well as well as well as its theoretical guarantee, and finally discuss the extensions of our method to extensions of our method to DBGD-

3.1
\n
$$
A \rightarrow A
$$
\n
$$
A \rightarrow A
$$
\n
$$
A \rightarrow A
$$
\n
$$
B \rightarrow B
$$
\n
$$
B \triangleq \{x \in A : ||x||_2 = 1\} - \{x \in A \mid x \in B
$$
\n
$$
B \triangleq \{x \in A : ||x||_2 = 1\} - \{x \in A \mid x \in B \
$$

$$
\mathbf{w}' = \varPi_{\mathcal{W}}[\mathbf{w} + \delta \mathbf{u}] \tag{4}
$$

1. Let C be the path length of the optimal rankers over T rounds, *defined as*

$$
C = \sum_{i=2}^n \|\mathbf{w}^* - \mathbf{w}^*_{-1}\|_2.
$$
 (5)

By setting $\delta = \sqrt{\frac{2}{(11+2 + \sqrt{2})^2}}$ *and* $\gamma = \sqrt{\frac{5^2+2\sqrt{2}}{2}}$ *, the dynamic regret of DBGD satisfies*

$$
[DR(T)] \le \sqrt{2(11+2\lambda)\lambda dL} \left(1 + \sqrt{5R^2 + 2RC}\right) T^{\frac{3}{4}}.
$$

3.2 DBGD Meets Meta Learning While DBGD can achieve a sub-linear dynamic regret bound for ^C = ^o(√ T), it requires the value of the path-length C for tuning the step size γ, which is clearly impossible in practice since C depends on the unknown optimal rankers

 \overline{A} 1. DBGD $\overline{\text{Re}}$ \sqrt{r} e: S \overline{A} \overline{S} \overline{r} \overline{r} 1: I**nitialize a** army w₁ ∈ W aboniani_y 2: $f \, \tilde{\,} t = 1, 2, \ldots, T \, d$ 3: Draw a vector **u***^t* uniformly at random from S 4: C a a $\oint a \oint y$ ar $w'_t = T_w[w_t + \delta u_t]$ 5: $\sqrt{\ }$ a W*t* ad W_t b_y plabilistic in a interleaving interleaving intervalse intervalse in a intervalse intervalse in a intervalse in a intervalse intervalse in a intervalse intervalse in a intervalse intervalse in 6: $\mathbf{v} \cdot \mathbf{f} \cdot \mathbf{w}_t' \succ \mathbf{w}_t' \cdot \mathbf{e}$ 7: S $\mathbf{W}_{t+1} = \Pi_{\mathcal{W}}[\mathbf{W}_t + \gamma \mathbf{U}_t]$
8: **e e** e e 9: S, $w_{t+1} = w_t$ 10: e d $\mathbf{d} \cdot \mathbf{f}$ 11: e d f \tilde{C}

$$
\frac{1}{\pi} \sum_{i=1}^{k} \frac{t_i}{\pi} + \frac{1}{\pi} \sum_{i=1}^{k} \frac{1}{\pi} \sum_{j=1}^{k} \frac{1}{\pi} \sum_{j=1}^{k} \frac{t_j}{\pi} \sum_{j
$$

Non-sa $\dot{\phi}$ a $_{\mathbf{y}}$ D $\dot{\phi}$ Band $\dot{\phi}$ -s_p O $\dot{\phi}$ Learning to Rank 171

$$
\sum_{i=1}^{n} \pi_{i} \mathbf{w}_{i} \cdot \mathbf{T} \rightarrow \mathbf{w}_{i} \quad \mathbf{w}' = H_{\mathcal{W}}[\mathbf{w} + \delta \mathbf{u}]_{i} \qquad \mathbf{w}'_{i} \sim \mathbf{w}_{i} \quad \mathbf{w}'_{i}
$$
\n
$$
\pi_{+1} = \frac{\pi \exp(-\alpha \ell(\mathbf{w}))}{\sum_{i=1}^{n} \pi \exp(-\alpha \ell(\mathbf{w}))}, \quad i = 1, ..., N \qquad (9)
$$
\n
$$
\ell(\mathbf{w}) \rightarrow \cdots \qquad \mathbf{w}' \rightarrow \mathbf{w}' \quad \mathbf{w}' \rightarrow \mathbf{w}' \quad (9)
$$
\n
$$
\ell(\mathbf{w}) = -\frac{d}{\delta} \langle \{w'_{i} \succ w_{i}\} \mathbf{u}, \mathbf{w} - \mathbf{w} \rangle
$$
\n
$$
\mathbf{w} = \frac{1}{\delta} \mathbf{w}'_{i} \sim \mathbf{w}' \quad \mathbf{w}' \rightarrow \mathbf{w}' \quad \mathbf{w}' \rightarrow \mathbf{w}' \quad \mathbf{w}' \rightarrow \mathbf{w}' \quad \mathbf{w}' \rightarrow \mathbf{w} \quad \mathbf{w} \rightarrow \mathbf{w} \quad \mathbf{w}' \rightarrow \mathbf{w} \quad \mathbf{w} \rightarrow \mathbf{w
$$

 $\{w'_t \succ w_t\}$ and $\{w'_t \succ w_t\}$ and $\{w'_t \succ w_t\}$ and $\{w'_t \succ w_t\}$

$$
\mathbf{w}_{+1} = \varPi_{\mathcal{W}}[\mathbf{w} + \gamma \, \mathbf{w}'_t \succ \mathbf{w}]
$$

 $\frac{1}{2}$ expert up to our rank expert updates its own rank expert updates its own rank expert updates its own rank

 \overline{A} 2. DM²L: M_A $\overline{\text{Re}}$, $\overline{\text{Re}}$ b μ $\overline{\text{SN}}$, s $\overline{\text{Re}}$ $\overline{\text{SN}}$, $\gamma_1, \ldots, \gamma_N$, a μ a α 1: I *j* A *j* \mathcal{A} **[3](#page-6-1)** *j* γ_i *j* ac $i \in [N]$ 2: I **j** \mathbf{A} **i** \mathbf{A} **c** \mathbf{A} **i** \mathbf{A} **f** \mathbf{A} **i** \mathbf 3: $f \, f \, f = 1, 2, \ldots, T \, d$ 4: $R \circ \mathbf{A}$ ar \mathbf{W}_t^i **f** ac \mathbf{W}_s^i i $\in [N]$ 5: A a a $\mathbf{a} \cdot \mathbf{a} = \sum_{i=1}^{N} \pi_i^i \mathbf{w}_i^i$ 6: Draw a vector **u***^t* uniformly at random from S 7: C a a \oint a_t \int a r $W'_t = H_W[\mathbf{w}_t + \delta \mathbf{u}_t]$ 8: C_f a W_t ad W'_t b_v **p** babilistic interval 9: U^{da} **the weight of each expert of** $\pi_t^i, i \in [N]$ **by [\(9\)](#page-5-0)** 10: S d_{*t*} $\mathbf{w}'_t \succ \mathbf{w}_t$ and \mathbf{u}_t **a** ach $i \in [N]$ 11: e d f \overline{f}

 \overline{A} 3. DM²L: E \overline{A} \overline{A}

```
\overline{\text{Re}} \sqrt{\text{e}}: \overline{s} \overline{a} \overline{y} \1: Initialize a ranker w<sup>i</sup> ∈ W a bitaliy
2: f(t = 1, 2, ..., T d)3: Sdar W^i_t2
 4: \operatorname{R} c \neq {\scriptstyle {\sqrt{\mathbf{w}}_t \succ \mathbf{w}_t}, and \operatorname{U}_t \neq {\scriptstyle {\log(\mathbf{w}_t)}}2
5: U da   a   vv+11
```
Ac edee^s. This was a partial was partially NSFC (61976112) and Jn SSF (BK20200064). We are anonymous $\mathbf{A} = \mathbf{A} + \mathbf{A} +$ $s = s$

References

- 1. C_ra-Bianchi, Lugosi, G.: Prediction, Learning, and Games. Cambridge Uni- \rightarrow P \rightarrow $s(2006)$
- 2. a. E., T., Ky.y., W.M.: Maad: Multiple lautes in a superior learning rates in \mathcal{A} . $\mathrm{I}:\mathrm{Ad}\ \mathfrak{a}$ \mathfrak{c} \sim Systems) N and I \mathfrak{a} \sim \mathfrak{a} \mathfrak{a} \sim \mathfrak{p} \sim \mathfrak{p} \sim \mathfrak{p} \sim \sim \sim \sim \sim 3666–3674 (2016)
- 3. Grotov, A., Rijke, M.: Online learning to rank for information retrieval. In: Proceedings of the 39th International ACM SIGIR Conference on Research and Devel- μ in Information Retrieval, pp. 1215–1218 (2016)
- $4. \; \rm{H}$ an, K.: Fasád and panielearning to rank for information retrieval. P. D. Dissertation (2013)
- 5. H_p a \overrightarrow{K} , Whiteson, S., Rijke, M.: A probabilistic method for inferrence preferring prefer $c = \epsilon$ from clicks. International Conference on the 20th ACM International Conference on $c = \epsilon$ I μ $a \not\mu$ ad K $_{\mu\bullet^2}$ d Maa $,$. 249–258 (2011)
- 6. Hofmann, K., Whiteson, S., Rijke, M.: Balancing exploration and exploitation in listwise and and pairwise online learning to rank for information retrieval. In the contract of \mathbf{a} R **A** \cdot 16(1), 63–90 (2013)
- $7. L$ i μ , $T.Y.:$ Learning to r information retrieval. For information r information μ information μ **3**(3), 225–331 (2009)
- 8. O_{pt}s \vec{A} , Rijke, M.: Differentiable unbiased on learning to rank. In: Pro c dings 27 ACM International Conference on International ACM $\frac{1}{\sqrt{2}}$ Maa, 1293–1302 (2018)
- 9. Radlinski, F., Kurup, M., Joachims, T.: How does clickthrough data reflect retrieval $a\lambda$, ℓ I: Prodings 17 ACM Conference on Information and Knowledge d Maa, $, .4352 (2008)$
- 10. Schuth, Oosterhuis, A., Oosterhuis, H., Win**i**a, S., Rajop, M.: Multila avadaja des^{cent} desc for fast online learning to rank. In: Proceedings of the 9th ACM International C_4 c μ W b S a c a d Da a Mining, 457–466 (2016)
- 11. Wang, H., Kim, S., McC_p d-S_{pp} , E., W., Q., Wang, H.: Variance reduction in \mathbf{K} addie pada pada and parti: Produce learning to 42 dI $a \triangle a$ ACM SIGIR C_b c $c \triangle a$ Research and D $c \triangle a$ in Information R \bm{a} , $.835844 (2019)$
- 12. Wang, H., Langley, R., Kim, S., McCord-Snow, E., Wang, H.: Efficient experiment explored to \mathbf{F} ration of gradient space for online learning to rank. In: Proceedings of the 41st International ACM SIGIR Conference on Research and Development in Informa**t** R **i** a, 145–154 (2018)
- 13. Y., Y., B., d., J., K. job., R., J., ac. j., S.T.: The karmed dueling banding ban μ b. J. C μ . S_ys. Sci. 78(5), 1538–1556 (2012)
- 14. Y., Y., Jacachims, T.: Interactively optimizing information retrieval systems as a system \mathbb{R}^2 dueling bandits problem. In: Proceedings of the 26th International Conference on Mac \blacksquare L a \blacksquare , 1201–1208 (2009)
- 15. Zang, Y., et al.: GISDCN: A a -based interpolation sequential recognition sequential recognition sequential recognition sequences in \mathbf{a} and \mathbf{a} and \mathbf{a} and \mathbf{a} and \mathbf{a} and \mathbf{a} and $\mathbf{a$ with deformable convolutional network. In: International Conference on Database $S_{\mathbf{v}}$ stems and Advanced Applications, pp. 289–297. Spinger, Cham (2022). $\frac{1}{\sqrt{2}}$ $d_{1.44}$ /10.1007/978-3-031-00126-0 21

 174 S. Lua.

- 16. Zhang, L., Lu, S., Zhou, Z.H.: Adaptive online learning in dynamic environments. $\mathrm{I}:\mathrm{Ad}\ \mathfrak{a}$ \mathfrak{c} \sim 5 i Neural i and P i \sim 5 i \sim 5 i i \sim 31, pp. 1323–1333 (2018)
- 17. Zay, P., Wa, G., Za, L., Zy, Z.H.: Badjoiq. في بِهْضُومَ الله عليه الصحيح الصحيح الصحيح الصحيح الصحيح الصح stationary environments. In: Proceedings of the 23rd International Conference on A $\overline{\text{A}}$ cone Intelligence and Statistics, pp. 1508–1518 (2020)