

SUPPLEMENTARY MATERIALS : RECURRENT EXPLORATION NETWORKS FOR RECOMMENDER SYSTEMS

Anonymous authors

Paper under double-blind review

1 PROOFS IN THE MAIN PAPER

In this section, we provide the detailed proofs for the lemmas and main theorem in the paper.

Assumption 1.1. Assume there exists an optimal θ^* , with $\|\theta^*\| \leq 1$ and $\mathbf{x}_{t,k}^*$ such that $\mathbf{E}[r_{t,k}] = \mathbf{x}_{t,k}^{*\top} \theta^*$. Further assume that there is an effective distribution $\mathcal{N}(\mu_{t,k}, \Sigma_{t,k})$ such that $\mathbf{x}_{t,k}^* \sim \mathcal{N}(\mu_{t,k}, \Sigma_{t,k})$ where $\Sigma_{t,k} = \text{diag}(\sigma_{t,k}^2)$. Thus, the true underlying context is unavailable, but we are aided with the knowledge that it is generated with a multivariate normal whose parameters are known.

Algorithm 1: BaseREN: Basic REN Inference at Step t

- 1 **Input:** $\alpha, \Psi_t \subseteq \{1, 2, \dots, t-1\}$.
 - 2 Obtain item embeddings from REN: $\mu_{\tau, k_\tau} \leftarrow f_e(\mathbf{e}_{\tau, k_\tau})$ for all $\tau \in \Psi_t$.
 - 3 Obtain the current user embedding from REN: $\theta_t \leftarrow R(\mathbf{D}_t)$.
 - 4 $\mathbf{A}_t \leftarrow \mathbf{I}_d + \sum_{\tau \in \Psi_t} \mu_{\tau, k_\tau}^\top \mu_{\tau, k_\tau}$.
 - 5 Obtain candidate items' embeddings from REN: $\mu_{t,k} \leftarrow f_e(\mathbf{e}_{t,k})$, where $k \in [K]$.
 - 6 Obtain candidate items' uncertainty estimates $\sigma_{t,k}$, where $k \in [K]$.
 - 7 **for** $a \in [K]$ **do**
 - 8 $w_{t,k} \leftarrow (\alpha + 1)s_{t,k} + (4\sqrt{d} + 2\sqrt{\ln \frac{TK}{\delta}})\|\sigma_{t,k}\|_\infty$.
 - 9 $\hat{r}_{t,k} \leftarrow \theta_t^\top \mu_{t,k}$.
 - 10 **end**
 - 11 Recommend item $k \leftarrow \text{argmax}_k \hat{r}_{t,k} + w_{t,k}$.
-

1.1 UPPER CONFIDENCE BOUND FOR UNCERTAIN EMBEDDINGS

For simplicity we follow the notation from Chu et al. (2011) and denote the item embedding (context) as $\mathbf{x}_{t,k}$, where t indexes the rounds and k indexes the items. We define:

$$\begin{aligned} s_{t,k} &= \sqrt{\mu_{t,k}^\top \mathbf{A}_t^{-1} \mu_{t,k}} \in \mathbb{R}_+, \quad \mathbf{D}_t = [\mu_{\tau, k_\tau}]_{\tau \in \Psi_t} \in \mathbb{R}^{|\Psi_t| \times d}, \\ \mathbf{y}_t &= [r_{\tau, k_\tau}]_{\tau \in \Psi_t} \in \mathbb{R}^{|\Psi_t| \times 1}, \quad \mathbf{A}_t = \mathbf{I}_d + \mathbf{D}_t^\top \mathbf{D}_t, \\ \mathbf{b}_t &= \mathbf{D}_t^\top \mathbf{y}_t, \quad \hat{r}_{t,k} = \mu_{t,k}^\top \hat{\theta} = \mu_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{b}_t, \end{aligned}$$

where \mathbf{y}_t is the collected user feedback. Lemma 1.1 below shows that with $\lambda_d = 1 + \alpha = 1 + \sqrt{\frac{1}{2} \ln \frac{2TK}{\delta}}$ and $\lambda_u = 4\sqrt{d} + 2\sqrt{\ln \frac{TK}{\delta}}$, the main equation in the paper is the upper confidence bound with high probability, meaning that it upper bounds the true reward with high probability, which makes it a reasonable score for recommendations.

Lemma 1.1 (Confidence Bound). With probability at least $1 - 2\delta/T$, we have for all $k \in [K]$ that

$$|\hat{r}_{t,k} - \mathbf{x}_{t,k}^{*\top} \theta^*| \leq (\alpha + 1)s_{t,k} + (4\sqrt{d} + 2\sqrt{\ln \frac{TK}{\delta}})\|\sigma_{t,k}\|_\infty,$$

where $\|\sigma_{t,k}\|_\infty = \max_i |\sigma_{t,k}^{(i)}|$ is the L_∞ norm.

Algorithm 2: SupREN

```

1 Input: Number of rounds  $T$ .
2  $S \leftarrow \ln T$  and  $\Psi_t^{(s)} \leftarrow \emptyset$  for all  $s \in [T]$ .
3 for  $t = 1, 2, \dots, T$  do
4    $s \leftarrow 1$  and  $\hat{A}_1 \leftarrow [K]$ .
5   repeat
6     Use BaseREN with  $\Psi_t^{(s)}$  to calculate the width,  $w_{t,k}^{(s)}$ , and the upper confidence bound,
        $\hat{r}_{t,k}^{(s)} + w_{t,k}^{(s)}$ , for all  $k \in \hat{A}_s$ .
7     if  $w_{t,k}^{(s)} \leq \frac{1}{\sqrt{T}}$  for all  $k \in \hat{A}_s$  then
8       Choose  $k_t = \operatorname{argmax}_{k \in \hat{A}_s} (\hat{r}_{t,k}^{(s)} + w_{t,k}^{(s)})$  and update:  $\Psi_{t+1}^{(s')} \leftarrow \Psi_t^{(s')}$  for all  $s' \in [S]$ .
9     else if  $w_{t,k}^{(s)} \leq 2^{-s}$  for all  $k \in \hat{A}_s$  then
10       $\hat{A}_{s+1} \leftarrow \{k \in \hat{A}_s \mid \hat{r}_{t,k}^{(s)} + w_{t,k}^{(s)} \geq \max_{k' \in \hat{A}_s} (\hat{r}_{t,k'}^{(s)} + w_{t,k'}^{(s)}) - 2^{1-s}\}$ ,  $s \leftarrow s + 1$ .
11    else
12      Choose  $k_t \in \hat{A}_s$  such that  $w_{t,k_t}^{(s)} > 2^{-s}$  and update:  $\Psi_{t+1}^{(s)} \leftarrow \Psi_t^{(s)} \cup \{t\}$ ,
         $\Psi_{t+1}^{(s')} \leftarrow \Psi_t^{(s')}$  for  $s' \neq s$ .
13    end
14  until an item  $k_t$  is found;
15  Update the REN model  $R(\cdot)$  and  $f_e(\cdot)$  using collected user feedbacks.
16 end

```

Proof. Using the notation defined above, we have

$$\begin{aligned}
& |\hat{r}_{t,k} - \mathbf{x}_{t,k}^*{}^\top \boldsymbol{\theta}^*| \\
&= |\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{b}_t - \mathbf{x}_{t,k}^*{}^\top \mathbf{A}_t^{-1} (\mathbf{I}_d + \mathbf{D}_t^\top \mathbf{D}_t) \boldsymbol{\theta}^*| \\
&= |\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{y}_t - \mathbf{x}_{t,k}^*{}^\top \mathbf{A}_t^{-1} (\boldsymbol{\theta}^* + \mathbf{D}_t^\top \mathbf{D}_t \boldsymbol{\theta}^*)| \\
&= |\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{y}_t - \mathbf{x}_{t,k}^*{}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{D}_t \boldsymbol{\theta}^* - \mathbf{x}_{t,k}^*{}^\top \mathbf{A}_t^{-1} \boldsymbol{\theta}^*| \\
&= |(\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{y}_t - \boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{D}_t \boldsymbol{\theta}^*) + \boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{D}_t \boldsymbol{\theta}^* - \mathbf{x}_{t,k}^*{}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{D}_t \boldsymbol{\theta}^* - \mathbf{x}_{t,k}^*{}^\top \mathbf{A}_t^{-1} \boldsymbol{\theta}^*| \\
&= |\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top (\mathbf{y}_t - \mathbf{D}_t \boldsymbol{\theta}^*) + (\boldsymbol{\mu}_{t,k} - \mathbf{x}_{t,k}^*)^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{D}_t \boldsymbol{\theta}^* - \mathbf{x}_{t,k}^*{}^\top \mathbf{A}_t^{-1} \boldsymbol{\theta}^*| \\
&= |\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top (\mathbf{y}_t - \mathbf{D}_t \boldsymbol{\theta}^*) - (\boldsymbol{\Sigma}_{t,k}^{1/2} \boldsymbol{\epsilon})^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{D}_t \boldsymbol{\theta}^* - (\boldsymbol{\mu}_{t,k} + \boldsymbol{\Sigma}_{t,k}^{1/2} \boldsymbol{\epsilon})^\top \mathbf{A}_t^{-1} \boldsymbol{\theta}^*| \\
&= |\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top (\mathbf{y}_t - \mathbf{D}_t \boldsymbol{\theta}^*) - (\boldsymbol{\Sigma}_{t,k}^{1/2} \boldsymbol{\epsilon})^\top \boldsymbol{\theta}^* - (\boldsymbol{\mu}_{t,k})^\top \mathbf{A}_t^{-1} \boldsymbol{\theta}^*| \\
&\leq |\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top (\mathbf{y}_t - \mathbf{D}_t \boldsymbol{\theta}^*)| + \|\boldsymbol{\Sigma}_{t,k}^{1/2} \boldsymbol{\epsilon}\| + s_{t,k}.
\end{aligned} \tag{1}$$

$$\leq |\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top (\mathbf{y}_t - \mathbf{D}_t \boldsymbol{\theta}^*)| + \|\boldsymbol{\Sigma}_{t,k}^{1/2} \boldsymbol{\epsilon}\| + s_{t,k}. \tag{2}$$

To see Eqn. 1 is true, note that $\mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{D}_t + \mathbf{A}_t^{-1} = \mathbf{A}_t^{-1} (\mathbf{D}_t^\top \mathbf{D}_t + \mathbf{I}_d) = \mathbf{I}_d$. To see Eqn. 2 is true, note that since $\|\boldsymbol{\theta}^*\| \leq 1$, we have $|(\boldsymbol{\Sigma}_{t,k}^{1/2} \boldsymbol{\epsilon})^\top \boldsymbol{\theta}^*| \leq \|\boldsymbol{\Sigma}_{t,k}^{1/2} \boldsymbol{\epsilon}\|$. Similarly for the last term in Eqn. 2, observe that

$$\begin{aligned}
& \|\mathbf{A}_t^{-1} \boldsymbol{\mu}_{t,k}\| \\
&= \sqrt{\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{I}_d \mathbf{A}_t^{-1} \boldsymbol{\mu}_{t,k}} \\
&\leq \sqrt{\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} (\mathbf{I}_d + \mathbf{D}_t^\top \mathbf{D}_t) \mathbf{A}_t^{-1} \boldsymbol{\mu}_{t,k}} \\
&= \sqrt{\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \boldsymbol{\mu}_{t,k}} \\
&= s_{t,k}.
\end{aligned} \tag{3}$$

For the first term in Eqn. 2, since $\mathbf{E}[\mathbf{y}_t - \mathbf{D}_t \boldsymbol{\theta}^*] = 0$, and $\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top \mathbf{y}_t$ is a random variable bounded by $\|\mathbf{D}_t \mathbf{A}_t^{-1} \boldsymbol{\mu}_{t,k}\|$, by Azuma-Hoeffding inequality, we have

$$\begin{aligned} & \Pr(|\boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \mathbf{D}_t^\top (\mathbf{y}_t - \mathbf{D}_t \boldsymbol{\theta}^*)| > \alpha s_{t,k}) \\ & \leq 2 \exp\left(-\frac{2\alpha^2 s_{t,k}^2}{\|\mathbf{D}_t \mathbf{A}_t^{-1} \boldsymbol{\mu}_{t,k}\|^2}\right) \\ & \leq 2 \exp(-2\alpha^2) \\ & = \frac{\delta}{TK}, \end{aligned} \tag{4}$$

where Eqn. 4 is due to

$$\begin{aligned} s_{t,k}^2 &= \boldsymbol{\mu}_{t,k}^\top \mathbf{A}_t^{-1} \boldsymbol{\mu}_{t,k} \\ &= \boldsymbol{\mu}_{t,k}^\top \mathbf{A}^{-1} (\mathbf{I}_d + \mathbf{D}_t^\top \mathbf{D}_t) \mathbf{A}^{-1} \boldsymbol{\mu}_{t,k} \\ &\geq \boldsymbol{\mu}_{t,k}^\top \mathbf{A}^{-1} \mathbf{D}_t^\top \mathbf{D}_t \mathbf{A}^{-1} \boldsymbol{\mu}_{t,k} \\ &= \|\mathbf{D}_t \mathbf{A}_t^{-1} \boldsymbol{\mu}_{t,k}\|^2. \end{aligned}$$

For the second term of Eqn. 2, $\|\boldsymbol{\epsilon}^\top \boldsymbol{\Sigma}_{t,k}^{1/2}\|$, since $\boldsymbol{\epsilon}^\top \boldsymbol{\Sigma}_{t,k}^{1/2} \sim \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_{t,k})$, we can guarantee that with probability at most $\frac{\delta}{TK}$,

$$\|\boldsymbol{\epsilon}^\top \boldsymbol{\Sigma}_{t,k}^{1/2}\| > 2\sqrt{\lambda_{\max}(\boldsymbol{\Sigma}_{t,k})(2\sqrt{d} + \sqrt{\ln \frac{TK}{\delta}})}, \tag{6}$$

where $\lambda_{\max}(\boldsymbol{\Sigma}_{t,k}) = \|\boldsymbol{\Sigma}_{t,k}\|_{op}$ is the operator norm of the matrix $\boldsymbol{\Sigma}_{t,k}$ corresponding to the L_2 vector norm.

Combining Eqn. 2, Eqn. 5, and Eqn. 6, with a union bound, we have that with probability at least $1 - \frac{2\delta}{T}$, for all actions $a \in [K]$,

$$\begin{aligned} |\hat{r}_{t,k} - \mathbf{x}_{t,k}^*{}^\top \boldsymbol{\theta}^*| &\leq (\alpha + 1)s_{t,k} + (4\sqrt{d} + 2\sqrt{\ln \frac{TK}{\delta}})\sqrt{\lambda_{\max}(\boldsymbol{\Sigma}_{t,k})}, \\ &= (\alpha + 1)s_{t,k} + (4\sqrt{d} + 2\sqrt{\ln \frac{TK}{\delta}})\|\boldsymbol{\sigma}_{t,k}\|_\infty, \end{aligned}$$

□

1.2 REGRET BOUND

Lemma 1.1 above provides a reasonable estimate of the reward's upper bound at time t . Based on this estimate, one natural next step is to analyze the regret after all T rounds. Formally, we define the regret of the algorithm after T rounds as

$$B(T) = \sum_{t=1}^T r_{t,k_t^*} - \sum_{t=1}^T r_{t,k_t}, \tag{7}$$

where k_t^* is the optimal item (action) k at round t that maximizes $\mathbf{E}[r_{t,k}] = \mathbf{x}_{t,k}^T \boldsymbol{\theta}^*$, and k_t is the action chose by the algorithm at round t . In a similar fashion as in Chu et al. (2011), SupREN calls BaseREN as a sub-routine. In this subsection, we derive the regret bound for SupREN with uncertain item embeddings.

Lemma 1.2 (Azuma–Hoeffding Inequality). *Let X_1, \dots, X_m be random variables with $|X_\tau| \leq a_\tau$ for some $a_1, \dots, a_m > 0$. Then we have*

$$\Pr\left(\left|\sum_{\tau=1}^m X_\tau - \sum_{\tau=1}^m \mathbf{E}[X_\tau | X_1, \dots, X_{\tau-1}]\right| \geq B\right) \leq 2 \exp\left(-\frac{B^2}{2 \sum_{\tau=1}^m a_\tau^2}\right).$$

Lemma 1.3. *With probability $1 - 2\delta S$, for any $t \in [T]$ and any $s \in [S]$:*

1. $|\hat{r}_{t,k} - \mathbf{E}[r_{t,k}]| \leq w_{t,k}$ for any $k \in [K]$,
2. $k_t^* \in \hat{A}_s$, and
3. $\mathbf{E}[r_{t,k_t^*}] - \mathbf{E}[r_{t,k}] \leq 2^{(3-s)}$ for any $k \in \hat{A}_s$.

Proof. The proof is a simple modification of that in Auer (2002) (Lemma 15) to accommodate modification in Lemma 1.1. \square

Lemma 1.4. *In BaseREN, we have*

$$(1 + \alpha) \sum_{t \in \Psi_{T+1}} s_{t,k_t} \leq 5 \cdot (1 + \alpha^2) \sqrt{d|\Psi_{T+1}|}.$$

Proof. This is a direct result of Lemma 3 and Lemma 6 in Chu et al. (2011) as well as Lemma 16 in Auer (2002). \square

Lemma 1.5. *Assuming $\|\sigma_{1,k}\|_\infty = 1$ and $\|\sigma_{t,k}\|_\infty \leq \frac{1}{\sqrt{t}}$ for any k and t , then for any k ,*

$$\sum_{t \in \Psi_{T+1}} \|\sigma_{t,k}\|_\infty \leq \sqrt{|\Psi_{T+1}|}$$

Proof. Since the function $f(t) = \frac{1}{\sqrt{t}}$ is convex when $t > 0$, we have

$$\sum_{t=1}^{|\Psi_{T+1}|} \frac{1}{\sqrt{t}} \leq \int_0^{|\Psi_{T+1}|} \frac{1}{\sqrt{t}} = \sqrt{t} \Big|_0^{|\Psi_{T+1}|} = \sqrt{|\Psi_{T+1}|}$$

\square

Lemma 1.6. *For all $s \in [S]$,*

$$|\Psi_{T+1}^{(s)}| \leq 2^s \cdot \left(5(1 + \alpha^2) \sqrt{d|\Psi_{T+1}^{(s)}|} + 4\sqrt{dT} + 2\sqrt{T \ln \frac{TK}{\delta}} \right).$$

Proof. This is true by combining Lemma 1.4, Lemma 1.5, and Lemma 1.1 with a similar proving strategy as in Lemma 16 of Auer (2002).

$$\sum_{t \in \Psi_{T+1}^{(s)}} w_{t,k}^{(s)} = (1 + \alpha) \sum_{t \in \Psi_{T+1}} s_{t,k_t} + (4\sqrt{d} + 2\sqrt{\ln \frac{TK}{\delta}}) \sum_{t \in \Psi_{T+1}} \|\sigma_{t,k}\|_\infty \quad (8)$$

$$\leq 5 \cdot (1 + \alpha^2) \sqrt{d|\Psi_{T+1}|} + (4\sqrt{d} + 2\sqrt{\ln \frac{TK}{\delta}}) \sqrt{|\Psi_{T+1}|} \quad (9)$$

$$\leq 5 \cdot (1 + \alpha^2) \sqrt{d|\Psi_{T+1}|} + 4\sqrt{dT} + 2\sqrt{T \ln \frac{TK}{\delta}}, \quad (10)$$

where Eqn. 9 is due to Lemma 1.4 and Lemma 1.5. By Line 12 of Algorithm 2, we have

$$\sum_{t \in \Psi_{T+1}^{(s)}} w_{t,k}^{(s)} \geq 2^{-s} |\Psi_{T+1}^{(s)}|. \quad (11)$$

Combine Eqn. 10 and Eqn. 11 yields this lemma. \square

Theorem 1.1. *If SupREN is run with $\alpha = \sqrt{\frac{1}{2} \ln \frac{2TK}{\delta}}$, with probability at least $1 - \delta$, the regret of the algorithm is*

$$O \left(\sqrt{Td \ln^3 \left(\frac{KT \ln(T)}{\delta} \right)} \right). \quad (12)$$

Proof. The proof is an extension of Theorem 6 in Auer (2002) to handle the uncertainty in item embeddings. We denote as Ψ_0 the set of trials for which an alternative is chosen in Line 8 of Algorithm 2. Note that $2^{-S} \leq \frac{1}{\sqrt{T}}$; therefore $\{1, \dots, T\} = \Psi_0 \cup \bigcup_s \Psi_{T+1}^{(s)}$. We have

$$\begin{aligned} E[B(T)] &= \sum_{t=1}^T [E[r_{t,k_t^*}] - E[r_{t,k_t}]] \\ &= \sum_{t \in \Psi_0} [E[r_{t,k_t^*}] - E[r_{t,k_t}]] + \sum_{s=1}^S \sum_{t \in \Psi_{T+1}^{(s)}} [E[r_{t,k_t^*}] - E[r_{t,k_t}]] \\ &\leq \frac{2}{\sqrt{T}} |\Psi_0| + \sum_{s=1}^S 8 \cdot 2^{-s} \cdot |\Psi_{T+1}^{(s)}| \end{aligned} \quad (13)$$

$$\leq \frac{2}{\sqrt{T}} |\Psi_0| + \sum_{s=1}^S 8 \cdot \left(5(1 + \alpha^2) \sqrt{d |\Psi_{T+1}^{(s)}|} + 4\sqrt{dT} + 2\sqrt{T \ln \frac{TK}{\delta}} \right) \quad (14)$$

$$\leq 2\sqrt{T} + 40(1 + \ln \frac{2TK}{\delta}) \sqrt{STd} + 32S\sqrt{dT} + 16S\sqrt{T \ln \frac{TK}{\delta}}, \quad (15)$$

with probability $1 - 2\delta S$. Eqn. 13 is by Lemma 1.3, and Eqn. 14 is by Lemma 1.6. By the Azuma–Hoeffding inequality (Lemma 1.2) with $B = 2\sqrt{2T} \sqrt{\ln \frac{2}{\delta}}$ and $a_\tau = 2$, we have

$$B(T) \leq 2\sqrt{T} + 44 \cdot (1 + \ln \frac{2TK}{\delta}) \sqrt{STd} + 32S\sqrt{dT} + 16S\sqrt{T \ln \frac{TK}{\delta}}, \quad (16)$$

with probability at least $1 - 2\delta(S + 1)$. To see this, note that $1 - 2\delta(S + 1) < 1 - 2\delta S - \delta$ and that

$$2\sqrt{2T} \sqrt{\ln \frac{2}{\delta}} \leq 4\sqrt{T} \sqrt{\ln \frac{2TK}{\delta}} \leq 4 \cdot (1 + \ln \frac{2TK}{\delta}) \sqrt{STd}.$$

Replacing δ by $\frac{\delta}{2S+2}$ and S by $\ln T$ in Eqn. 16 along with simplification gives us

$$\begin{aligned} B(T) &\leq 2\sqrt{T} + 44 \cdot (1 + \ln \frac{2TK(2S+2)}{\delta}) \sqrt{T \ln T} \sqrt{d} + 32S\sqrt{dT} + 16S\sqrt{T \ln \frac{TK(2S+2)}{\delta}} \\ &\leq 2\sqrt{T} + 44 \cdot (1 + \ln \frac{2TK(2S+2)}{\delta}) (1 + \ln T)^{\frac{1}{2}} \sqrt{Td} + 32S\sqrt{dT} + 16 \ln T \sqrt{\ln \frac{TK(2S+2)}{\delta}} \sqrt{T} \\ &\leq 2\sqrt{T} + 44 \cdot (1 + \ln \frac{2TK(2 \ln T + 2)}{\delta})^{\frac{3}{2}} \sqrt{Td} \\ &\quad + 32 \cdot (1 + \ln \frac{2TK(2 \ln T + 2)}{\delta}) \sqrt{dT} + 16 \cdot (1 + \ln \frac{2TK(2 \ln T + 2)}{\delta})^{\frac{3}{2}} \sqrt{Td} \\ &\leq 2\sqrt{T} + 92 \cdot (1 + \ln \frac{2TK(2 \ln T + 2)}{\delta})^{\frac{3}{2}} \sqrt{Td}, \end{aligned}$$

with probability $1 - \delta$. Therefore we have

$$B(T) \leq 2\sqrt{T} + 92 \cdot (1 + \ln \frac{2TK(2 \ln T + 2)}{\delta})^{\frac{3}{2}} \sqrt{Td} = O(\sqrt{Td \ln^3(\frac{KT \ln(T)}{\delta})}),$$

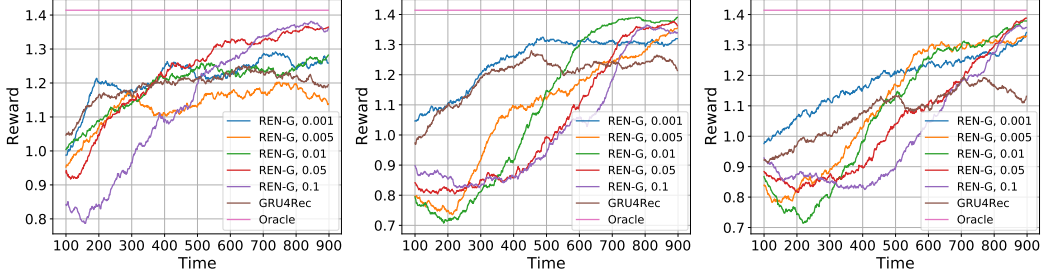
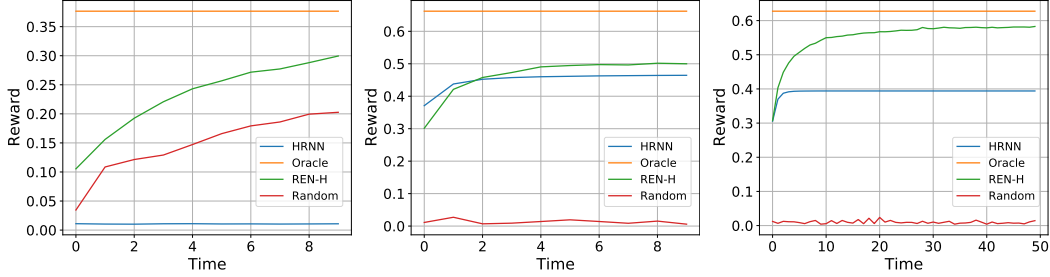
with probability $1 - \delta$. \square

Theorem 1.1 shows that even with the uncertainty in the item embeddings, our proposed REN can achieve the same rate-optimal sublinear regret bound as in Chu et al. (2011).

2 MORE DETAILS ON DATASETS

2.1 MovieLens-1M

We use *MovieLens-1M* (Harper & Konstan, 2016) containing 3,900 movies and 6,040 users. Specifically, we randomly select 1,000 users from *MovieLens-1M*, where each user has 120 interactions, and follow the joint learning and exploration procedure described in the main paper to evaluate all methods.

Figure 1: Hyperparameter sensitivity for λ_d in *SYN-S*, *SYN-M*, and *SYN-L*.Figure 2: Rewards over time on *Netflix*. One time step represents 100 recommendations to a user.

2.2 Trivago

Trivago is a hotel recommendation dataset with 730,803 users, 926,457 items, and 910,683 interactions. We use a subset with 57,778 users, 387,348 items, and 108,713 interactions and slice the data into $M = 48$ one-hour time intervals for the online experiment. Different from *MovieLens-1M*, *Trivago* has impression data available. Specifically, at each time step, besides which item is clicked by the user, we also know which 25 items are being shown to the user. Essentially the RecSys Challenge is a reranking problem with candidate sets of size 25.

2.3 Netflix

Our main conclusion with *Netflix* experiments is that REN-inference-only procedure collects more diverse data points about a user, which allows us to build a more generalizable user model, which leads to better long-term rewards. The main paper demonstrates better generalizability by comparing precision@100 reward on a holdout item set, where the items are inaccessible to the user - i.e., we never collect feedback on these holdout items in our simulations. Instead, recommendations are made by comparing the users' learned embeddings and the pretrained embeddings of the holdout items. Fig. 2(left) shows similar trends with recall@100 as the reward on the same holdout item set. This shows that the collected set contributes to building better user embedding models.

Fig. 2(middle) shows that the additional exploration power comes without significant harms to the user's immediate rewards on the exploration set, where the recommendations are served. In fact, we used a relatively large exploration coefficient, $\lambda_d = \lambda_u = 0.005$, which starts to affect recommendation results on the sixth position. By additional hyperparameter tuning, we realized that to achieve better rewards on the exploration set, we may choose smaller $\lambda_d = 0.0007$ and $\lambda_u = 0.0008$. Fig. 2(right) shows significantly higher recalls close to the oracle performance, where all of the users' histories are known and used as inputs to predict the top-100 personalized recommendations.¹ Note that, for fair presentation of the tuned results, we switched the exploration set and the holdout set and used a different test user group, consisting of 1543 users. We believe that the tuned results are generalizable with new users and items, but we also realize that the *Netflix* dataset still has a significant popularity bias and therefore we recommend using larger exploration coefficients with real online systems. The inference cost is 175 milliseconds to pick top-100 items

¹The gap between oracle and 100% recall lies in the model approximation errors.

from 8000 evaluation items. It includes 100 sequential linear function solutions with 50 embedding dimensions, which is further improvable by selecting multiple items at a time in slate generation.

3 HYPERPARAMETERS AND NEURAL NETWORK ARCHITECTURES

For the base models GRU4Rec, TCN, and HRNN, we use identical network architectures and hyperparameters whenever possible following (Hidasi et al., 2016; Bai et al., 2018; Ma et al., 2020). Each RNN consists of an encoding layer, a core RNN layer, and a decoding layer. We set the number of hidden neurons to 32 for all models including REN variants. Fig. 1 shows the REN-G’s performance for different λ_d (note that we fix $\lambda_u = \sqrt{10}\lambda_d$) in *SYN-S*, *SYN-M*, and *SYN-L*. We can observe stable REN performance across a wide range of λ_d . As expected, REN-G’s performance is closer to GRU4Rec when λ_d is small.

REFERENCES

- Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *JMLR*, 3:397–422, 2002.
- Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *CoRR*, abs/1803.01271, 2018.
- Wei Chu, Lihong Li, Lev Reyzin, and Robert Schapire. Contextual bandits with linear payoff functions. In *AISTATS*, pp. 208–214, 2011.
- F Maxwell Harper and Joseph A Konstan. The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19, 2016.
- Balázs Hidasi, Alexandros Karatzoglou, Linas Baltrunas, and Domonkos Tikk. Session-based recommendations with recurrent neural networks. In *ICLR*, 2016.
- Yifei Ma, Murali Balakrishnan Narayanaswamy, Haibin Lin, and Hao Ding. Temporal-contextual recommendation in real-time. In *KDD*, 2020.