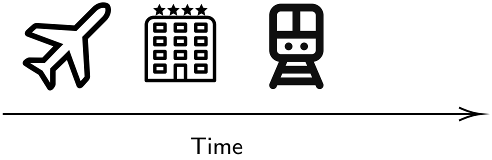
On the Unexpected Effectiveness of Reinforcement Learning for Sequential Recommendation

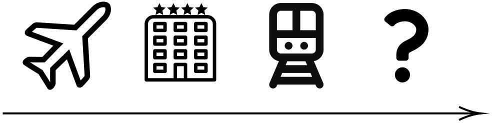
Álvaro Labarca Silva
Denis Parra
Rodrigo Toro

An Intriguing Result About Reinforcement Learning in Sequential RecSys

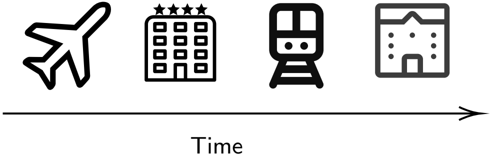
The Sequential Recommendation Problem



The Sequential Recommendation Problem

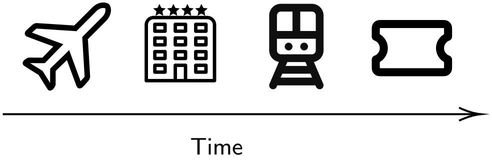


Problem

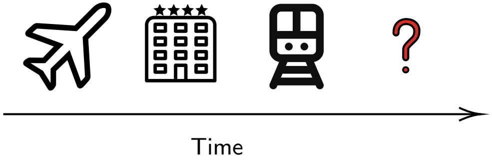


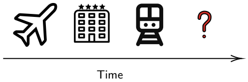
Collaborative Filtering Methods

The Sequential Recommendation Problem

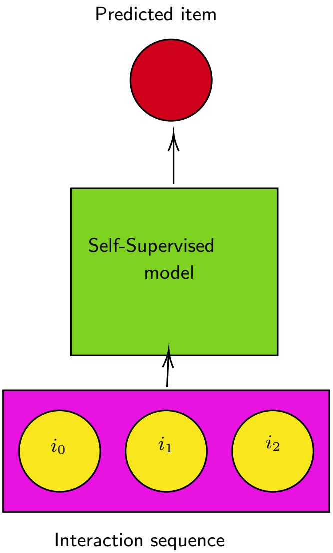


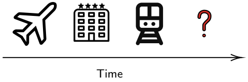
Sequential Models


Time

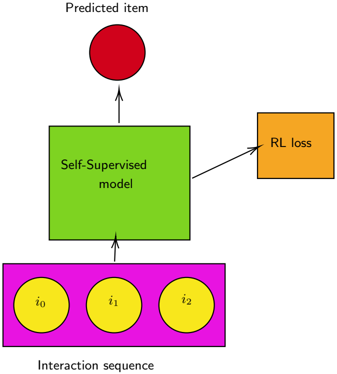


- Caser
- GRU4Rec
- NextItNet
- SASRec

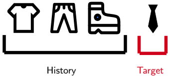


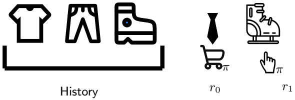


- Caser
- GRU4Rec
- NextItNet
- SASRec



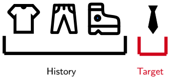
Reinforcement Learning for Sequential Recommendation

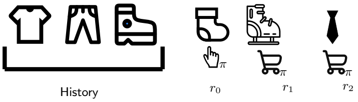




$$R = \sum_{t=0}^{\infty} \gamma^t r_t$$

Reinforcement Learning for Sequential Recommendation





$R = \sum_{t=0}^{\infty} \gamma^t r_t$

- Since RL optimizes performance in long-term, there is no reason to expect improvements on the short-term.

- Since RL optimizes performance in long-term, there is no reason to expect improvements on the short-term.

Theorem

For any consistent NIP metric $\mathcal{N}$ and discount factor $\gamma > 0$, the relative NIP performance of an optimal policy π_* can be arbitrarily worse than the performance of an optimal solution f_*, according to L_s.

Theorem

Let's consider any consistent NIP metric $\mathcal{N}$ and set of interaction sequences $\mathcal{D}$. Let f_* be an optimal solution to the cross-entropy loss L_s. Let π be any optimal policy with respect to $\mathcal{D}$. Then, $\mathcal{N}(\mathcal{D}, f_*) \geq \mathcal{N}(\mathcal{D}, \pi_*)$.

Research Question

Why do RL-based methods improve the performance under the NIP metric?

Intuition

Our hypothesis is that RL, as the process of learning an optimal policy from data, is not directly responsible for the performance improvements.

Intuition

Our hypothesis is that RL, as the process of learning an optimal policy from data, is not directly responsible for the performance improvements. We believe that a clever combination of reward signals and discount factors entails useful auxiliary losses

Intuition

Our hypothesis is that RL, as the process of learning an optimal policy from data, is not directly responsible for the performance improvements. We believe that a clever combination of reward signals and discount factors entails useful auxiliary losses that create embeddings containing information about the user or the sequence that the model can use to improve short-term recommendation.

Top Down Approach

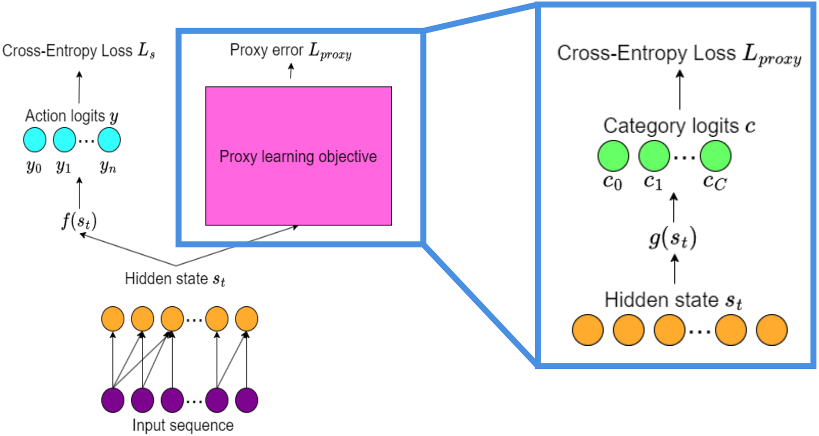
Interaction prediction

Purchase: 5 Click: 1 Discount Factor (γ): 0.5

i_0	i_1	i_2	i_3	i_4	i_5	i_6	i_7	i_8	i_9	i_{10}	i_{11}	i_{12}	i_{13}	i_{14}	i_{15}	i_{16}	i_{17}	i_{18}
6.9	3.8	5.5	9.0	8.0	6.0	2.0	2.0	2.0	2.0	2.1	2.1	2.2	2.6	3.2	4.4	6.8	3.5	5.0

$$
\begin{aligned}
q_\pi(s_t, a_t) &= r_{t+1} + \frac{1}{2} r_{t+2} + \frac{1}{4} r_{t+3} + \sum_{k=3}^{\infty} \frac{1}{2^k} r_{t+k+1} \\
&\leq r_{t+1} + \frac{1}{2} r_{t+2} + \frac{1}{4} r_{t+3} + 1.25
\end{aligned}
$$

CAT Model



$$L_{proxy} = -\sum_i^C y_i \log(f(c)_i)$$

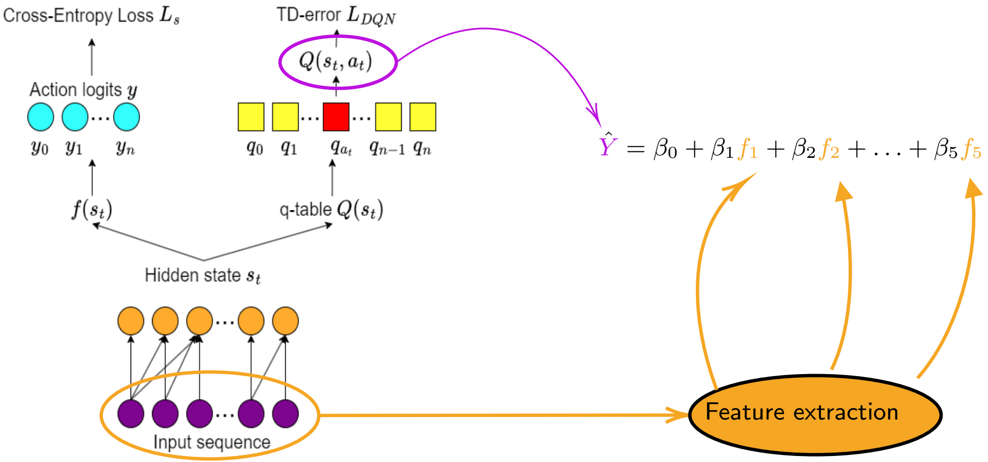
$$f(c)_i = \frac{e^{c_i}}{\sum_j^C e^{c_j}}$$

CAT Results

Model	HR@5	NDCG@5	HR@20	NDCG@20
GRU	0.1601	0.1248	0.2306	0.1456
GRU-SQN	***0.1921**	**0.1519**	***0.2698**	**0.1743**
GRU-CAT	0.1644	0.1282	*0.2384	0.1495

Bottom-up Approach

Feature Importance



Feature importance results

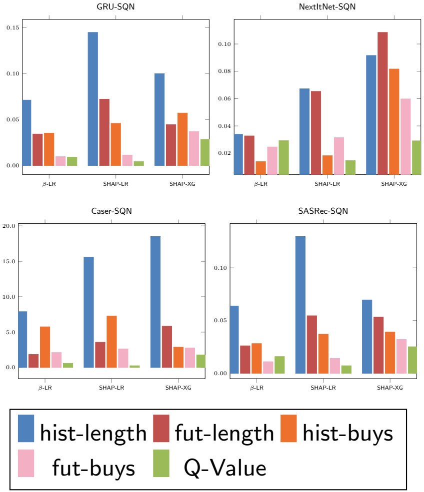
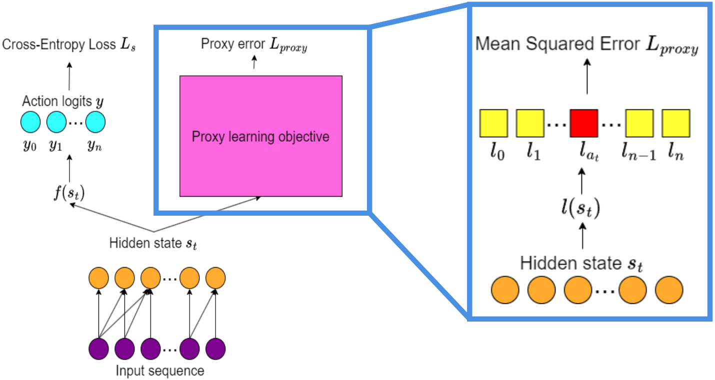


Figure 1: Feature Importance for Different Models

HIST Model



$$L_{proxy} = \frac{1}{N}\sum_{i=1}^{N}(Y_i - l_{a_t})^2$$

Results

GRU				
Model	HR@5	NDCG@5	HR@20	NDCG@20
GRU	0.1601	0.1248	0.2306	0.1450
GRU-SQN	*0.1921	0.1519	*0.2698	0.1743
GRU-EVAL	*0.1962	*0.1545	*0.2718	*0.1762
GRU-cat	0.1644	0.1282	*0.2384	0.1495
GRU-hist	***0.1980**	***0.1549**	***0.2747**	***0.1770**

NIN				
Model	HR@5	NDCG@5	HR@20	NDCG@20
NIN	0.2282	0.1785	0.3215	0.2053
NIN-SQN	*0.3307	*0.2577	*0.4418	*0.2901
NIN-EVAL	***0.3308**	***0.2582**	***0.4421**	***0.2905**
NIN-cat	*0.2505	0.1959	*0.3483	0.2242
NIN-hist	*0.3001	*0.2348	*0.4110	*0.2667

Caser				
Model	HR@5	NDCG@5	HR@20	NDCG@20
Caser	0.1682	0.1361	0.2217	0.1514
Caser-SQN	*0.2020	*0.1601	*0.2715	*0.1801
Caser-EVAL	0.2000	0.1610	*0.2638	0.1794
Caser-cat	0.1765	0.1434	0.2307	0.1591
Caser-hist	***0.2399**	***0.1893**	***0.3296**	***0.2152**

SAS				
Model	HR@5	NDCG@5	HR@20	NDCG@20
SAS	0.2458	0.1872	0.3509	0.2176
SAS-SQN	***0.3012**	***0.2280**	***0.4227**	***0.2634**
SAS-EVAL	0.2963	*0.2242	*0.4195	*0.2600
SAS-cat	*0.2636	*0.1999	*0.3731	*0.2315
SAS-hist	*0.2960	*0.2231	0.3847	*0.2573

Conclusions and Future Work

- The HIST model achieved competitive results with the SQN model.

- The HIST model achieved competitive results with the SQN model.
- With the GRU and Caser self-supervised base, the HIST model outperformed SQN.

- The HIST model achieved competitive results with the SQN model.
- With the GRU and Caser self-supervised base, the HIST model outperformed SQN.
- Different reward schemes and models may learn other signals.

- The HIST model achieved competitive results with the SQN model.
- With the GRU and Caser self-supervised base, the HIST model outperformed SQN.
- Different reward schemes and models may learn other signals.
- The research serves as a necessary step to improve the understanding, explainability and performance of RL methods in recommendation.

Future Work

- Explore different signals.

Future Work

- Explore different signals.
- Deepen the understanding of the HIST model.

Future Work

- Explore different signals.
- Deepen the understanding of the HIST model.
- Extend research to different RL models.

Future Work

- Explore different signals.
- Deepen the understanding of the HIST model.
- Extend research to different RL models.
- Develop methods to evaluate the RL effect in the long term.

Thank you! Any questions?

Full results - GRU click

Model	HR@5	NG@5	HR@10	NG@10	HR@15	NG@15	HR@20	NG@20
GRU	0.1205	0.0938	0.1472	0.1024	0.1633	0.1067	0.1751	0.1094
GRU-SQN	*0.1416	*0.1105	*0.1705	*0.1199	*0.1874	*0.1244	*0.1999	*0.1273
GRU-SAC	***0.1475**	***0.1148**	***0.1774**	***0.1245**	**0.1947**	***0.1291**	***0.2069**	***0.1320**
GRU-QVAL	*0.1400	*0.1089	*0.1680	*0.1180	*0.1849	*0.1224	*0.1970	*0.1253
GRU-EVAL	*0.1426	*0.1109	*0.1714	*0.1203	*0.1888	*0.1249	<u>*0.2009</u>	*0.1277
GRU-cat	*0.1228	*0.0955	*0.1497	*0.1042	*0.1665	*0.1086	*0.1784	*0.1114
GRU-cat3	*0.1249	*0.0969	*0.1514	*0.1054	*0.1677	*0.1097	*0.1797	*0.1126
GRU-hist	<u>*0.1434</u>	<u>*0.1118</u>	<u>*0.1717</u>	<u>*0.1210</u>	<u>*0.1880</u>	<u>*0.1253</u>	*0.1997	<u>*0.1281</u>
GRU-fut	0.0390	0.0312	0.0474	0.0339	0.0528	0.0353	0.0569	0.0375

Full results - NIN click

Model	HR@5	NG@5	HR@10	NG@10	HR@15	NG@15	HR@20	NG@20
NIN	0.1345	0.1059	0.1612	0.1145	0.1774	0.1188	0.1892	0.1216
NIN-SQN	***0.1673**	***0.1310**	<u>*0.1996</u>	***0.1414**	<u>*0.2178</u>	***0.1463**	<u>*0.2305</u>	***0.1493**
NIN-SAC	<u>*0.1671</u>	*0.1301	***0.1999**	*0.1407	***0.2186**	*0.1457	***0.2317**	*0.1488
NIN-QVAL	*0.1653	*0.1295	*0.1963	*0.1396	0.2141	*0.1443	*0.2273	*0.1474
NIN-EVAL	*0.1668	<u>*0.1306</u>	*0.1993	<u>*0.1411</u>	*0.2176	<u>*0.1459</u>	*0.2302	<u>*0.1489</u>
NIN-cat	*0.1431	*0.1121	*0.1721	*0.1215	*0.1890	*0.1259	*0.2014	*0.1289
NIN-cat3	*0.1436	*0.1129	*0.1732	*0.1225	0.1903	*0.1271	*0.2027	*0.1300
NIN-hist	*0.1638	*0.1284	*0.1951	*0.1386	*0.2130	*0.1433	*0.2256	0.1463
NIN-fut	0.0939	0.0750	0.1100	0.0802	0.1194	0.0827	0.1263	0.0830

Full results - Caser click

Model	HR@5	NG@5	HR@10	NG@10	HR@15	NG@15	HR@20	NG@20
Caser	0.1400	0.1107	0.1640	0.1185	0.1781	0.1222	0.1877	0.1245
Caser-SQN	*0.1560	0.1218	*0.1849	*0.1312	*0.2018	*0.1357	*0.2136	*0.1385
Caser-SAC	*0.1539	0.1190	*0.1836	*0.1286	*0.2012	*0.1333	*0.2132	*0.1361
Caser-QVAL	<u>*0.1608</u>	<u>*0.1269</u>	<u>*0.1883</u>	<u>*0.1358</u>	<u>*0.2040</u>	<u>*0.1400</u>	<u>*0.2151</u>	<u>*0.1426</u>
Caser-EVAL	0.1577	0.1246	0.1849	0.1334	0.2005	0.1375	0.2116	0.1401
Caser-cat	0.1415	0.1135	0.1652	0.1212	0.1790	0.1249	0.1888	0.1344
Caser-cat3	0.1565	*0.1241	*0.1830	*0.1327	*0.1982	*0.1367	0.2090	*0.1393
Caser-hist	**0.1669**	***0.1300**	***0.1979**	**0.1401**	***0.2160**	***0.1449**	***0.2283**	***0.1478**
Caser-fut	0.0248	0.0189	0.0316	0.0210	0.0358	0.0222	0.0393	0.0230

Full results - SAS click

Model	HR@5	NG@5	HR@10	NG@10	HR@15	NG@15	HR@20	NG@20
SAS	0.1635	0.1249	0.1982	0.1361	0.2176	0.1413	0.2313	0.1445
SAS-SQN	0.1835	*0.1397	*0.2228	*0.1524	*0.2445	*0.1582	*0.2597	*0.1618
SAS-SAC	***0.1852**	***0.1398**	***0.2266**	***0.1532**	***0.2496**	***0.1593**	***0.2649**	***0.1629**
SAS-QVAL	*0.1804	*0.1373	*0.2189	*0.1498	*0.2409	*0.1556	*0.2557	*0.1591
SAS-EVAL	0.1815	*0.1382	*0.2212	*0.1511	*0.2433	*0.1569	*0.2584	0.1605
SAS-cat	*0.1689	*0.1285	*0.2057	*0.1404	*0.2259	*0.1458	*0.2399	*0.1491
SAS-cat3	0.1669	0.1271	*0.2034	0.1390	*0.2239	0.1444	*0.2380	0.1477
SAS-hist	*0.1812	*0.1376	*0.2204	*0.1503	*0.2424	*0.1561	*0.2576	*0.1597
SAS-fut	0.0099	0.0071	0.0138	0.0083	0.0168	0.0091	0.0193	0.0097

Full results - GRU buy

Model	HR@5	NG@5	HR@10	NG@10	HR@15	NG@15	HR@20	NG@20
GRU	0.1601	0.1248	0.1932	0.1355	0.2151	0.1413	0.2306	0.1456
GRU-SQN	*0.1921	0.1519	*0.2304	0.1643	*0.2531	0.1703	*0.2698	0.1743
GRU-SAC	<u>0.1973</u>	<u>*0.1546</u>	<u>*0.2338</u>	<u>*0.1664</u>	<u>*0.2575</u>	<u>*0.1727</u>	<u>*0.2734</u>	<u>*0.1764</u>
GRU-QVAL	*0.1884	*0.1487	*0.2233	*0.1599	*0.2448	*0.1656	*0.2601	*0.1693
GRU-EVAL	*0.1962	*0.1545	*0.2333	*0.1664	*0.2555	*0.1723	*0.2718	*0.1762
GRU-cat	0.1644	0.1282	*0.2004	*0.1400	*0.2225	0.1457	*0.2384	0.1495
GRU-cat3	*0.1696	*0.1325	*0.2044	*0.1438	*0.2272	*0.1498	*0.2435	*0.1536
GRU-hist	***0.1980**	***0.1549**	***0.2352**	***0.1670**	***0.2588**	***0.1732**	***0.2747**	***0.1770**
GRU-fut	0.0505	0.0400	0.0623	0.0438	0.0697	0.0457	0.0756	0.0471

Full results - NIN buy

Model	HR@5	NG@5	HR@10	NG@10	HR@15	NG@15	HR@20	NG@20
NIN	0.2282	0.1785	0.2746	0.1935	0.3018	0.2007	0.3215	0.2053
NIN-SQN	*0.3307	*0.2577	***0.3890**	*0.2767	*0.4200	*0.2849	*0.4418	*0.2901
NIN-SAC	*0.3303	***0.2594**	*0.3880	***0.2781**	***0.4201**	***0.2867**	***0.4422**	***0.2919**
NIN-QVAL	*0.3216	*0.2511	0.3791	*0.2675	0.4094	*0.2778	0.4310	*0.2829
NIN-EVAL	***0.3308**	*0.2582	*0.3885	*0.2770	*0.4199	*0.2853	*0.4421	*0.2905
NIN-cat	*0.2505	0.1959	*0.2998	0.2119	*0.3280	0.2194	*0.3483	0.2242
NIN-cat3	*0.2543	*0.1973	*0.3046	*0.2136	*0.3346	*0.2215	*0.3553	*0.2264
NIN-hist	*0.3001	*0.2348	*0.3566	*0.2529	*0.3896	*0.2616	*0.4110	*0.2667
NIN-fut	0.1980	0.1581	0.2301	0.1685	0.2490	0.1735	0.2623	0.1767

Full results - Caser buy

Model	HR@5	NG@5	HR@10	NG@10	HR@15	NG@15	HR@20	NG@20
Caser	0.1682	0.1361	0.1947	0.1446	0.2105	0.1488	0.2217	0.1514
Caser-SQN	*0.2020	*0.1601	*0.2367	*0.1713	*0.2568	*0.1766	*0.2715	*0.1801
Caser-SAC	*0.1979	*0.1559	*0.2237	*0.1684	*0.2581	*0.1741	<u>*0.2737</u>	*0.1778
Caser-QVAL	<u>*0.2050</u>	<u>*0.1644</u>	<u>*0.2391</u>	<u>*0.1755</u>	<u>*0.2594</u>	<u>*0.1808</u>	*0.2731	<u>*0.1841</u>
Caser-EVAL	0.2000	0.1610	*0.2321	0.1714	*0.2505	0.1762	*0.2638	0.1794
Caser-cat	0.1765	0.1434	0.2032	0.1521	0.2193	0.1564	0.2307	0.1591
Caser-cat3	0.1936	*0.1565	*0.2254	*0.1668	*0.2420	*0.1712	*0.2541	*0.1740
Caser-hist	***0.2399**	***0.1893**	***0.2857**	***0.2041**	***0.3112**	***0.2108**	***0.3296**	***0.2152**
Caser-fut	0.0299	0.0222	0.0378	0.0247	0.0435	0.0262	0.0481	0.0273

Full results - SAS buy

Model	HR@5	NG@5	HR@10	NG@10	HR@15	NG@15	HR@20	NG@20
SAS	0.2458	0.1872	0.2995	0.2046	0.3283	0.2123	0.3509	0.2176
SAS-SQN	*0.3012	*0.2280	*0.3657	*0.2490	*0.3989	*0.2577	*0.4227	*0.2634
SAS-SAC	***0.3143**	***0.2379**	***0.3810**	***0.2596**	***0.4153**	***0.2687**	***0.4382**	***0.2741**
SAS-QVAL	*0.2860	*0.2163	*0.3471	*0.2361	*0.3799	*0.2448	*0.4010	*0.2498
SAS-EVAL	0.2963	*0.2242	*0.3619	*0.2454	*0.3962	*0.2545	*0.4195	*0.2600
SAS-cat	*0.2636	*0.1999	*0.3190	*0.2178	*0.3504	*0.2261	*0.3731	*0.2315
SAS-cat3	0.2572	0.1948	*0.3130	0.2129	*0.3437	0.2211	*0.3651	0.2261
SAS-hist	*0.2960	*0.2231	0.3315	*0.2428	0.3630	*0.2519	0.3847	*0.2573
SAS-fut	0.0207	0.0144	0.0285	0.0170	0.0349	0.0187	0.0400	0.0199

Deriving Theorem 1

Let's consider a sequential recommendation problem with two items $\mathcal{I} = \{x_1, x_2\}$ and only two possible interaction sequences that any user can take: $\{x_1, x_2\}$ and $\{x_2\}$. For simplicity, let's consider that the training and testing sets are identical, meaning that overfitting for the training test will lead to optimal performances in the testing set.
This set $\mathcal{D}$ contains one trace $\{x_1, x_2\}$ and $n > 1$ copies of the trace $\{x_2\}$.

An optimal solution f_* to the cross-entropy loss – assuming that f_* has enough capacity to encode such a solution – is the following: $f_*(x_1|\emptyset) = \frac{1}{n+1}$, $f_*(x_2|\emptyset) = \frac{n}{n+1}$, and $f_*(x_2|\{x_1\}) = 1$. As a result, the NIP performance of f_* over $\mathcal{D}$ is the following:

$$\mathcal{N}(\mathcal{D}, f_*) = \frac{1}{n+2}\left(n \cdot s(1) + s(2) + s(1)\right) = \frac{n+1}{n+2},$$

Deriving Theorem 1

On the other hand, an optimal policy π_* for $\mathcal{D}$ first recommends x_1 because the expected discounted return of recommending x_1 is $q_*(\emptyset, x_1) = r(1+\gamma)$ whereas the Q-function of recommending x_2 is $q_*(\emptyset, x_2) = r$. Thus, as long as we define a positive reward r for interacting with an item and we use a discount of $\gamma > 0$, then $\pi_*(x_1|\emptyset) = 1$. And once the user interacts with x_1, the next recommendation will be x_2: $\pi_*(x_2|x_1) = 1$. Therefore, according to the NIP performance, π_* will fail at recommending x_1 instead of x_2 in n sequences in $\mathcal{D}$:

$$\mathcal{N}(\mathcal{D}, \pi_*) = \frac{1}{n+2}\left(n \cdot s(2) + s(1) + s(1)\right) = \frac{2}{n+2},$$

Hence, the ratio between the scores of f_* and π_* is the following:

$$\frac{\mathcal{N}(\mathcal{D}, f_*)}{\mathcal{N}(\mathcal{D}, \pi_*)} = n + \frac{1}{2} > n.$$

Then, as we increase the value of n the optimal policy π_* can perform arbitrarily worse than f_* according to any consistent NIP metric.

Deriving Theorem 2

Let $\mathcal{R}_\pi$ be a sequential recommender that ranks items according to a policy π and $\mathcal{R}_f$ be a sequential recommender that ranks items according to f_*. Let $C_\mathcal{D} : \mathcal{I}^* \to \mathbb{N}$ be a function that returns the number of times a subsequence appears in $\mathcal{D}$. In particular, $C_\mathcal{D}(x_{1:t})$ is equal to the number of sequences in $\mathcal{D}$ that begins with $x_{1:t}$. Then,

$$f_*(x_{1:t}, y) = \frac{C(x_{1:t} \circ y \mid \mathcal{D})}{C(x_{1:t} \mid \mathcal{D})} \quad \text{for all } y \in \mathcal{I}, \tag{1}$$

where $x \circ y$ represents the concatenation of x and y.

$$\mathcal{N}(\mathcal{D}, \mathcal{R}) = \frac{1}{N} \sum_{x_s \subset \mathcal{D}} \sum_{y \in \mathcal{I}} C(x_s \circ y \mid \mathcal{D}) \cdot s(\mathcal{R}(x_s, y))$$

Let $\mathcal{N}(\mathcal{D}, \mathcal{R}, x_s)$ be the following:

$$\mathcal{N}(\mathcal{D}, \mathcal{R}, x_s) = \sum_{y \in \mathcal{I}} C(x_s \circ y \mid \mathcal{D}) \cdot s(\mathcal{R}(x_s, y))$$

Deriving Theorem 2

Then, we will prove that, for all subsequence $x_s \subset \mathcal{D}$:

$$\begin{aligned}\mathcal{N}(\mathcal{D}, \mathcal{R}_f, x_s) &= \sum_{y \in \mathcal{I}} C(x_s \circ y \mid \mathcal{D}) \cdot s(\mathcal{R}_f(x_s, y)) \\ &\geq \sum_{y \in \mathcal{I}} C(x_s \circ y \mid \mathcal{D}) \cdot s(\mathcal{R}_\pi(x_s, y)) = \mathcal{N}(\mathcal{D}, \mathcal{R}_\pi, x_s) \end{aligned} \tag{2}$$

$\mathcal{R}_f(x_s, y)$ ranks the items according to $C(x_s \circ y \mid \mathcal{D})$.

We will prove that Equation 2 holds by showing that no other ranking could achieve a higher $\mathcal{N}(\mathcal{D}, \mathcal{R}_f, x_s)$ value than $\mathcal{R}_f$, for any x_s. Let's assume that there exists a policy π such that its ranking $\mathcal{R}_\pi(x_s, \cdot)$ for the next item given the interaction sequence x_s has higher $\mathcal{N}(\mathcal{D}, \mathcal{R}_\pi, x_s)$ value. For now, let's consider that the only difference between the rankings $\mathcal{R}_f(x_s, \cdot)$ and $\mathcal{R}_\pi(x_s, \cdot)$ is in the location of two items y_1 and y_2 that are swapped. That is, $\mathcal{R}_f(x_s, y_1) = \mathcal{R}_\pi(x_s, y_2)$ and $\mathcal{R}_f(x_s, y_2) = \mathcal{R}_\pi(x_s, y_1)$. Let's say that f_* ranks y_1 higher than y_2. Let p_1^f be the position of y_1 according to f_* and p_2^f be the position of y_2. Conversely, let p_1^π be the position of y_1 according to π and p_2^π be the position of y_2.

Deriving Theorem 2

Since f_* ranks y_1 higher than y_2, then $C_1 = C(x_s \circ y_1 \mid \mathcal{D}) \geq C(x_s \circ y_2 \mid \mathcal{D}) = C_2$. Therefore,

$$C_1 = C_2 + \epsilon \ ,$$

for some $\epsilon \geq 0$. In addition, we know that the scoring function s is non-increasing. That means that:

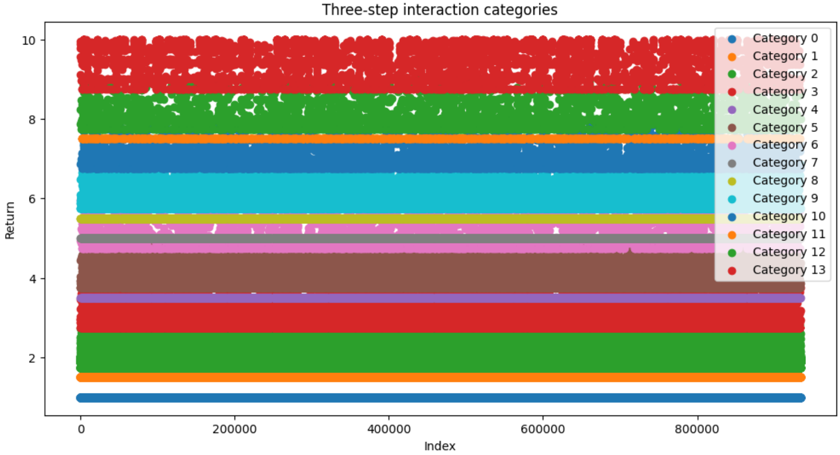
$$s(p_1^f) = s(p_2^f) + \beta \ ,$$

where $\beta \geq 0$. We now prove that the swap cannot increase the value of $\mathcal{N}(\mathcal{D}, \mathcal{R}_f, x_s)$.

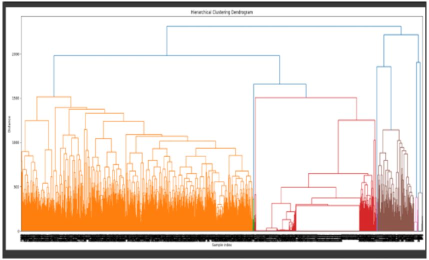
$$\begin{aligned}
C_1 \cdot s(p_1^f) + C_2 \cdot s(p_2^f) &= (C_2 + \epsilon) \cdot (s(p_2^f) + \beta) + C_2 \cdot s(p_2^f) \\
&\geq C_2 \cdot s(p_2^f) + C_2 \cdot \beta + \epsilon \cdot s(p_2^f) + C_2 \cdot s(p_2^f) \\
&= C_2 \cdot s(p_2^\pi) + s(p_1^\pi) \cdot C_1
\end{aligned}$$

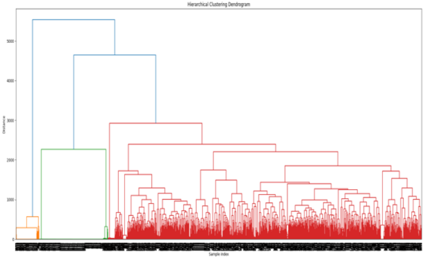
Thus, swapping the order of y_1 and y_2 cannot increase the value of $\mathcal{N}(\mathcal{D}, \mathcal{R}_f, x_s)$. And, for the same reason, making more than one swaps cannot increase the value of $\mathcal{N}(\mathcal{D}, \mathcal{R}_f, x_s)$. This means that, regardless the policy π, $\mathcal{N}(\mathcal{D}, \mathcal{R}_f, x_s) \geq \mathcal{N}(\mathcal{D}, \mathcal{R}_\pi, x_s)$. And since this relation holds for all $x_s \subset \mathcal{D}$, $\mathcal{N}(\mathcal{D}, f_*) \geq \mathcal{N}(\mathcal{D}, \pi_*)$ – proving the theorem.

Three-Step categorization (CAT2 model)



Clustering approach-Hierarchical Clustering





Clustering approach-Features

Table 1: Initial list of features before filtering

Feature	Description
Interaction	Categorical feature denoting the interaction type (click, buy) at the target timestamp
Interaction_2	Categorical feature denoting the interaction type (click, buy, done) at timestamp $t + 1$
Interaction_3	Categorical feature denoting the interaction type (click, buy, done) at timestamp $t + 2$
Is_done	Binary feature denoting whether the sequence finishes at timestamp t
hist-length	Number of past user interactions in the sequence.
fut-length	Number of future user interactions in the sequence.
total-length	Number of interactions in the complete sequence.
Q-Value (eval)	The expected return following the sequence in the history log.
hist-buys	Number of items the user bought in past interactions.
fut-buys	Number of items the user will buy in future interactions.
Steps2Buy	Number of steps until the next buy interaction in the sequence.

Clustering approach-Results

Table 2: Feature importance values for clustering models

Feature	Kmeans-4	Kmeans-8	Hierarchical
hist-length	**0.418**	**0.362**	**0.448**
total-length	0.255	0.255	0.237
fut-length	0.113	0.117	0.115
Q-Value (EVAL)	0.073	0.085	0.060
fut-buys	0.092	0.057	0.037
hist-buys	0.031	0.035	0.030
Steps2Buy	0.021	0.028	0.029

Parameter Setting

Table 3: Model parameters used in training. **Batch**: Batch size used. **lr**: learning rate. **h_factor**: Hidden factor or item embedding size. **filter#** Number of horizontal filters used in Caser. **f_sizes**: The size of the horizontal filters in Caser. **Head#** : Number of heads in self-attention in SASRec. **dropout**: Dropout Rate. **CR**: Click Reward. **BR**: Buy Reward

Model	Optimizer	Epochs	Batch	lr	γ	h_factor	filter#	f_sizes	Head#	dropout	CR	BR
GRU	Adam	50	256	0.005	0.5	64	-	-	-	0	1	5
NIN	Adam	50	256	0.005	0.5	64	-	-	-	0	1	5
Caser	Adam	50	256	0.005	0.5	64	16	[2,3,4]	-	0.1	1	5
SAS	Adam	50	256	0.005	0.5	64	-		1	0.1	1	5