Supplementary Materials for Practical Counterfactual Policy Learning for Top-K Recommendations

Yaxu Liu*†
National Taiwan University
d08944012@ntu.edu.tw

Jui-Nan Yen†
National Taiwan University
juinanyen@gmail.com

Bowen Yuan†‡
Amazon
bwyuan@amazon.com

Rundong Shi
Meituan
shirundong@meituan.com

Peng Yan
Meituan
yanpeng04@meituan.com

Chih-Jen Lin
National Taiwan University
cjlin@csie.ntu.edu.tw

1 DETAILED DERIVATIONS OF THE UNBIASEDNESS OF $V^{\pi}_{\mathbf{IPS}}$

From (11), (2), (3), and Assumption 1, we have

$$\begin{aligned}
V^{\pi}_{\text{IPS}} &= \mathbb{E}_{\Pr(\mathbf{u})}\mathbb{E}_{\beta(\mathbf{A}|\mathbf{u};V)}\mathbb{E}_{\Pr(\mathbf{c}|\mathbf{u};V)}\left[\frac{\pi(\mathbf{A}\mid\mathbf{u};V,\boldsymbol{\theta})}{\beta(\mathbf{A}\mid\mathbf{u};V)}r(\mathbf{c},\mathbf{A})\right]\\
&= \mathbb{E}_{\Pr(\mathbf{u})}\mathbb{E}_{\Pr(\mathbf{c}|\mathbf{u};V)}\mathbb{E}_{\beta(\mathbf{A}|\mathbf{u};V)}\left[\frac{\pi(\mathbf{A}\mid\mathbf{u};V,\boldsymbol{\theta})}{\beta(\mathbf{A}\mid\mathbf{u};V)}r(\mathbf{c},\mathbf{A})\right]\\
&= \mathbb{E}_{\Pr(\mathbf{u})}\mathbb{E}_{\Pr(\mathbf{c}|\mathbf{u};V)}\left[\sum_{\substack{\mathbf{A}\in G(\mathbb{A},K)\\ \beta(\mathbf{A}|\mathbf{u};V)\neq 0}}\pi(\mathbf{A}\mid\mathbf{u};V,\boldsymbol{\theta})r(\mathbf{c},\mathbf{A})\right]\\
&= \mathbb{E}_{\Pr(\mathbf{u})}\mathbb{E}_{\Pr(\mathbf{c}|\mathbf{u};V)}\left[\sum_{\mathbf{A}\in G(\mathbb{A},K)}\pi(\mathbf{A}\mid\mathbf{u};V,\boldsymbol{\theta})r(\mathbf{c},\mathbf{A})\right]\\
&= \mathbb{E}_{\Pr(\mathbf{u})}\mathbb{E}_{\pi(\mathbf{A}|\mathbf{u};V,\boldsymbol{\theta})}\mathbb{E}_{\Pr(\mathbf{c}|\mathbf{u};V)}\left[r(\mathbf{c},\mathbf{A})\right] = V^{\pi}.
\end{aligned}$$

2 DETAILED DERIVATIONS OF THEOREM 1

To prove Theorem 1, we need the following Lemma.

LEMMA 1 (HOEFFDING'S INEQUALITY δ-VERSION). *Assume $X_1,\ldots,X_m$ to be i.i.d.with 0 mean and $|X_i|\le M$ almost surely. Then with probability at least $1-\delta$, we have*

$$\left|\frac{1}{m}\sum_{i=1}^{m}X_i\right| \le M\sqrt{\frac{2}{m}\log\frac{2}{\delta}}.$$

Proof of Lemma 1.

Hoeffding's inequality states that for every positive t, $P(|\frac{1}{m}\sum_{i=1}^{m}X_i| \ge t) \le 2\exp(-\frac{mt^2}{2M^2})$. Now let $\delta = 2\exp(-\frac{mt^2}{2M^2})$. Solving for t, and we get $t = M\sqrt{\frac{2}{m}\log\frac{2}{\delta}}$, which completes our proof.

Proof of Theorem 1.

Recalling the definition,

$$\hat{V}^{\pi}_{\text{IPS}}(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m}\mathrm{w}^{i}_{\mathbf{A}}r(\mathbf{c}_i,\mathbf{A}_i).$$

Let $X_i = \mathrm{w}^{i}_{\mathbf{A}}r(\mathbf{c}_i,\mathbf{A}_i) - V^{\pi}$, which are i.i.d. From (2), (4), Assumptions 2 and 3, we know that $|X_i| \le Kw_{\max}$. As we have $\mathbb{E}[X_i]=0$, By taking δ' instead, Lemma 1 tells us that for any single π,

$$P\left(|\hat{V}^{\pi}_{\text{IPS}} - V^{\pi}| \le Kw_{\max}\sqrt{\frac{2}{m}\log\frac{2}{\delta'}}\right) \ge 1-\delta'. \tag{A.1}$$

According to the union bound, for a countable set of events $S_1, S_2, S_3, \cdots$, we have

$$P\left(\bigcup_{i=1}^{\infty}S_i\right) \le \sum_{i=1}^{\infty}P(S_i). \tag{A.2}$$

With the De Morgan's Law, we have

$$\begin{aligned}
P\left(\bigcup_{i=1}^{\infty}S_i^c\right) &\le \sum_{i=1}^{\infty}P(S_i^c),\\
P\left(\left(\bigcap_{i=1}^{\infty}S_i\right)^c\right) &\le \sum_{i=1}^{\infty}\Big(1-P(S_i)\Big),\\
1-P\left(\bigcap_{i=1}^{\infty}S_i\right) &\le \sum_{i=1}^{\infty}\Big(1-P(S_i)\Big),\\
P\left(\bigcap_{i=1}^{\infty}S_i\right) &\ge 1-\sum_{i=1}^{\infty}\Big(1-P(S_i)\Big),
\end{aligned} \tag{A.3}$$

where S_i^c means the complement of S_i. Then, by applying (A.1) to (A.3) over the finite policy class $\mathcal{H}$, where $|\mathcal{H}| = N$, we get

$$\begin{aligned}
P\left(\bigcap_{\pi\in\mathcal{H}}\left\{|\hat{V}^{\pi}_{\text{IPS}} - V^{\pi}| \le Kw_{\max}\sqrt{\frac{2}{m}\log\frac{2}{\delta'}}\right\}\right) &\ge 1-N\delta',\\
P\left(\sup_{\pi\in\mathcal{H}}\left\{|\hat{V}^{\pi}_{\text{IPS}} - V^{\pi}| \le Kw_{\max}\sqrt{\frac{2}{m}\log\frac{2}{\delta'}}\right\}\right) &\ge 1-N\delta',\\
P\left(\sup_{\pi\in\mathcal{H}}\left\{|\hat{V}^{\pi}_{\text{IPS}} - V^{\pi}| \le Kw_{\max}\sqrt{\frac{2}{m}\log\frac{2N}{\delta}}\right\}\right) &\ge 1-\delta,
\end{aligned} \tag{A.4}$$

where $\delta \equiv N\delta'$. This completes our proof.

3 DETAILED DERIVATIONS OF THEOREM 2

To prove Theorem 2, we need the following Lemma.

LEMMA 2. *Given $\mathbf{u}$, $\mathbf{A}_{1:k}$, and $\mathrm{c}_{\mathrm{A}_k}$, we have*

$$\mathbb{E}_{P_{\beta}(\mathbf{A}_{k+1:K}|\mathbf{u},\mathbf{A}_{1:k},\mathrm{c}_{\mathrm{A}_k})}\left[\Big(\prod_{j=k+1}^{K}\mathrm{w}_{\mathrm{A}_j}\Big)\right] = 1. \tag{A.5}$$

Proof of Lemma 2.

---

*Work done at Meituan as an intern.

†Contribute equally for this work.

‡Work done at National Taiwan University.

From our assumption $P_\beta(c_{A_k}|\mathbf{u}, \mathbf{A}_{1:k}, \mathbf{A}_{k+1:K}) = P_\beta(c_{A_k}|\mathbf{u}, \mathbf{A}_{1:k})$, we have

$$
\begin{aligned}
&P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u}, \mathbf{A}_{1:k}, c_{A_k}) \\
=&\frac{P_\beta(\mathbf{u}, \mathbf{A}_{1:k}, \mathbf{A}_{k+1:K}, c_{A_k})}{P_\beta(\mathbf{u}, \mathbf{A}_{1:k}, c_{A_k})} \\
=&\frac{P_\beta(\mathbf{u}, \mathbf{A}_{1:k})P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u}, \mathbf{A}_{1:k})P_\beta(c_{A_k}|\mathbf{u}, \mathbf{A}_{1:k}, \mathbf{A}_{k+1:K})}{P_\beta(\mathbf{u}, \mathbf{A}_{1:k})P_\beta(c_{A_k}|\mathbf{u}, \mathbf{A}_{1:k})} \\
=&\frac{P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u}, \mathbf{A}_{1:k})P_\beta(c_{A_k}|\mathbf{u}, \mathbf{A}_{1:k}, \mathbf{A}_{k+1:K})}{P_\beta(c_{A_k}|\mathbf{u}, \mathbf{A}_{1:k})} \\
=&P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u}, \mathbf{A}_{1:k}).
\end{aligned}
\tag{A.6}
$$

According to our definition in Section 2.1, $P_\beta(\cdot)$ is the probability distribution decided by the policy β. Thus, from the structure of β in (19), we have

$$P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u}, \mathbf{A}_{1:k}) = \beta(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V}).$$

Then, from (11), we have

$$
\begin{aligned}
&\mathbb{E}_{P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u},\mathbf{A}_{1:k},c_{A_k})}\left[(\prod_{j=k+1}^{K} w_{A_j})\right] && \text{(A.7)}\\
=&\mathbb{E}_{P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u},\mathbf{A}_{1:k},c_{A_k})}\left[\frac{\pi(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V}, \boldsymbol{\theta})}{\beta(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V})}\right] && \text{(A.8)}\\
=&\sum_{\substack{\mathbf{A}_{k+1:K}\in G(\mathbb{A}\setminus \mathbf{A}_{1:k}, K-k)\\ P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u},\mathbf{A}_{1:k},c_{A_k})\neq 0}} P_\beta(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}, c_{A_k}) \\
&\times \frac{\pi(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V}, \boldsymbol{\theta})}{\beta(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V})} && \text{(A.9)}\\
=&\sum_{\substack{\mathbf{A}_{k+1:K}\in G(\mathbb{A}\setminus \mathbf{A}_{1:k}, K-k)\\ \beta(\mathbf{A}_{k+1:K}|\mathbf{u},\mathbf{A}_{1:k};\boldsymbol{V})\neq 0}} \beta(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V})\frac{\pi(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V}, \boldsymbol{\theta})}{\beta(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V})} && \text{(A.10)}\\
=&\sum_{\substack{\mathbf{A}_{k+1:K}\in G(\mathbb{A}\setminus \mathbf{A}_{1:k}, K-k)\\ \beta(\mathbf{A}_{k+1:K}|\mathbf{u},\mathbf{A}_{1:k};\boldsymbol{V})\neq 0}} \pi(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V}, \boldsymbol{\theta}) && \text{(A.11)}\\
=&\sum_{\mathbf{A}_{k+1:K}\in G(\mathbb{A}\setminus \mathbf{A}_{1:k}, K-k)} \pi(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V}, \boldsymbol{\theta}), && \text{(A.12)}
\end{aligned}
$$

where (A.10) comes from (A.6), and the equation (A.12) comes from Assumption 1. Finally, we have $\sum_{\mathbf{A}_{k+1:K}\in G(\mathbb{A}\setminus \mathbf{A}_{1:k}, K-k)} \pi(\mathbf{A}_{k+1:K} \mid \mathbf{u}, \mathbf{A}_{1:k}; \boldsymbol{V}, \boldsymbol{\theta}) = 1$, which follows from the fact that π is a probability distribution. This completes our proof.

Proof of Theorem 2.

From (11), (20), and Lemma 2, we have

$$
\begin{aligned}
V^{\pi}_{\text{IPS}} =&\mathbb{E}_{P_\beta}\left[\sum_{k=1}^{K} w_{\mathbf{A}} c_{A_k}\right] && \text{(A.13)}\\
=&\sum_{k=1}^{K}\mathbb{E}_{P_\beta}\left[w_{\mathbf{A}} c_{A_k}\right] && \text{(A.14)}\\
=&\sum_{k=1}^{K}\mathbb{E}_{P_\beta(\mathbf{u},\mathbf{A}_{1:k},c_{A_k})}\mathbb{E}_{P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u},\mathbf{A}_{1:k},c_{A_k})}\Big[(\prod_{j=1}^{k} w_{A_j}) \\
&\times (\prod_{j=k+1}^{K} w_{A_j}) c_{A_k}\Big] && \text{(A.15)}\\
=&\sum_{k=1}^{K}\mathbb{E}_{P_\beta(\mathbf{u},\mathbf{A}_{1:k},c_{A_k})}\Big[(\prod_{j=1}^{k} w_{A_j}) c_{A_k} \\
&\times \mathbb{E}_{P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u},\mathbf{A}_{1:k},c_{A_k})}\left[(\prod_{j=k+1}^{K} w_{A_j})\right]\Big] && \text{(A.16)}\\
=&\sum_{k=1}^{K}\mathbb{E}_{P_\beta(\mathbf{u},\mathbf{A}_{1:k},c_{A_k})}\left[(\prod_{j=1}^{k} w_{A_j}) c_{A_k}\right] && \text{(A.17)}\\
=&\sum_{k=1}^{K}\mathbb{E}_{P_\beta(\mathbf{u},\mathbf{A}_{1:k},c_{A_k})}\mathbb{E}_{P_\beta(\mathbf{A}_{k+1:K}|\mathbf{u},\mathbf{A}_{1:k},c_{A_k})}\left[(\prod_{j=1}^{k} w_{A_j}) c_{A_k}\right] && \text{(A.18)}\\
=&\sum_{k=1}^{K}\mathbb{E}_{P_\beta}\left[(\prod_{j=1}^{k} w_{A_j}) c_{A_k}\right], && \text{(A.19)}\\
=&\mathbb{E}_{P_\beta}\left[\sum_{k=1}^{K}(\prod_{j=1}^{k} w_{A_j}) c_{A_k}\right], && \text{(A.20)}
\end{aligned}
$$

where (A.14), (A.16), and (A.20) rely on the linearity of expectation $\mathbb{E}[\mathbf{X}+\mathbf{Y}] = \mathbb{E}[\mathbf{X}] + \mathbb{E}[\mathbf{Y}]$ and $\mathbb{E}[a\mathbf{X}] = a\mathbb{E}[\mathbf{X}]$.

4 DETAILED DERIVATIONS OF THE SECOND-ORDER APPROXIMATION AS SQUARED LOSS

Let $T(\hat{Y}_{ij}, \tilde{y}_j)$ be the second-order approximation of $\ell^{-}_{\log}(\hat{Y}_{ij})$ at $\tilde{y}_j$. By denoting $\ell^{-}_{\log}(\tilde{y}_j)$ as E_{j0}, $\nabla\ell^{-}_{\log}(\tilde{y}_j)$ as E_{j1} and $\nabla^2\ell^{-}_{\log}(\tilde{y}_j))$ as E_{j2}, we have

$$
\begin{aligned}
T(\hat{Y}_{ij}, \tilde{y}_j) =&E_{j0} + E_{j1}(\hat{Y}_{ij} - \tilde{y}_j) + \frac{E_{j2}}{2}(\hat{Y}_{ij} - \tilde{y}_j)^2 \\
=&\frac{1}{2}E_{j2}(\hat{Y}_{ij} + \frac{E_{j1} - E_{j2}\tilde{y}_j}{E_{j2}})^2 \\
&+ (E_{j0} - E_{j1}\tilde{y}_j + \frac{1}{2}E_{j2}\tilde{y}_j^2 - \frac{1}{2}\frac{(E_{j1} - E_{j2}\tilde{y}_j)^2}{E_{j2}}),
\end{aligned}
\tag{A.21}
$$

where the fist part is a squared loss and the second part can be omitted as a constant for $\hat{Y}_{ij}$.