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
bwyyuan@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 \( V_{IPS}^\pi \)

From (11), (2), (3), and Assumption 1, we have
\[
V_{IPS}^\pi = E_{Pr(u)} E_{\beta(A|u,V)} E_{Pr(c|u,V)} \left( \frac{\pi(A|u,V,\theta)}{\beta(A|u,V)} r(c,A) \right)
\]
\[
= E_{Pr(u)} E_{Pr(c|u,V)} E_{Pr(A|u,V)} \left( \sum_{A \in G(A,K)} \frac{\pi(A|u,V,\theta)}{\beta(A|u,V)} r(c,A) \right)
\]
\[
= E_{Pr(u)} E_{Pr(c|u,V)} \pi(A|u,V,\theta) E_{Pr(A|u,V)} r(c,A) = V^\pi.
\]

2 DETAILED DERIVATIONS OF THEOREM 1

To prove Theorem 1, we need the following Lemma.

Lemma 1 (Hoeffding’s inequality \( \delta \)-version). Assume \( X_1, \ldots, X_m \) to be i.i.d. with 0 mean and \( |X_i| \leq M \) almost surely. Then with probability at least \( 1 - \delta \), we have
\[
\left| \frac{1}{m} \sum_{i=1}^{m} X_i \right| \leq M \sqrt{\frac{2}{m} \log \frac{2}{\delta}}.
\]

Proof of Lemma 1.

Hoeffding’s inequality states that for every positive \( t \), \( P\left( \left| \frac{1}{m} \sum_{i=1}^{m} X_i \right| \geq t \right) \leq 2 \exp\left( -\frac{mt^2}{2m^2} \right) \). Now let \( \delta = 2 \exp\left( -\frac{mt^2}{2m^2} \right) \). 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,
\[
\hat{V}_{IPS}^\pi(\theta) = \frac{1}{m} \sum_{i=1}^{m} w_i^\pi r(c_i, A_i).
\]

Let \( X_i = w_i^\pi r(c_i, A_i) - V^\pi \), which are i.i.d. From (2), (4), Assumptions 2 and 3, we know that \( |X_i| \leq K \max \). As we have \( E[X_i] = 0 \),

By taking \( \delta' \) instead, Lemma 1 tells us that for any single \( \pi \),
\[
P\left( \left| \hat{V}_{IPS}^\pi - V^\pi \right| \leq K \max \sqrt{\frac{2}{m} \log \frac{2}{\delta'}} \right) \geq 1 - \delta'. \tag{A.1}
\]

According to the union bound, for a countable set of events \( S_1, S_2, S_3, \ldots \), we have
\[
P\left( \bigcup_{i=1}^{\infty} S_i \right) \leq \sum_{i=1}^{\infty} P(S_i). \tag{A.2}
\]

With the De Morgan’s Law, we have
\[
P\left( \bigcap_{i=1}^{\infty} S_i^c \right) \leq \sum_{i=1}^{\infty} \left( 1 - P(S_i) \right).
\]

\[
1 - \sum_{i=1}^{\infty} \left( 1 - P(S_i) \right) \leq \sum_{i=1}^{\infty} \left( 1 - P(S_i) \right).
\] \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
\[
P\left( \bigcap_{\pi \in \mathcal{H}} \left\{ |\hat{V}_{IPS}^\pi - V^\pi| \leq K \max \sqrt{\frac{2}{m} \log \frac{2}{\delta'}} \right\} \right) \geq 1 - N \delta', \tag{A.4}
\]

\[
P\left( \bigcap_{\pi \in \mathcal{H}} \left\{ |\hat{V}_{IPS}^\pi - V^\pi| \leq K \max \sqrt{\frac{2}{m} \log \frac{2N}{\delta}} \right\} \right) \geq 1 - \delta,
\]

where \( \delta = N \delta' \). This completes our proof.

3 DETAILED DERIVATIONS OF THEOREM 2

To prove Theorem 2, we need the following Lemma.

Lemma 2. Given \( u, A_{1:k} \), and \( c_{A_k} \), we have
\[
E_{P_\theta(A_{k+1} | u, A_{1:k}, c_{A_k})} \left( \prod_{j=k+1}^{K} W_{A_j} \right) = 1. \tag{A.5}
\]

Proof of Lemma 2.
From our assumption \( P_\beta(c_{A_k} | u, A_{1:k}, A_{k+1:K}) = P_\beta(c_{A_k} | u, A_{1:k}) \), we have

\[
P_\beta(A_{k+1:K} | u, A_{1:k}, c_{A_k}) = P_\beta(u, A_{1:k}, A_{k+1:K}, c_{A_k}) \]

\[
= \frac{P_\beta(u, A_{1:k}, A_{k+1:K}) P_\beta(c_{A_k} | u, A_{1:k}, A_{k+1:K})}{P_\beta(u, A_{1:k}) P_\beta(c_{A_k} | u, A_{1:k})} \]  \hspace{1cm} (A.6)

\[
= \frac{P_\beta(A_{k+1:K} | u, A_{1:k}) P_\beta(c_{A_k} | u, A_{1:k}, A_{k+1:K})}{P_\beta(c_{A_k} | u, A_{1:k})} = P_\beta(A_{k+1:K} | u, A_{1:k}).
\]

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

\[
P_\beta(A_{k+1:K} | u, A_{1:k}) = \beta(A_{k+1:K} | u, A_{1:k}; V).
\]

Then, from (11), we have

\[
\mathbb{E}_{P_\beta}(A_{k+1:K} | u, A_{1:k}, c_{A_k}) = \mathbb{E} P_\beta(A_{k+1:K} | u, A_{1:k}, c_{A_k}) \]

\[
= \mathbb{E} P_\beta(A_{k+1:K} | u, A_{1:k}, c_{A_k}) \pi(A_{k+1:K} | u, A_{1:k}; V, \theta) \]

\[
= \sum_{A_{k+1:K} \in G(A|A_{1:k}, K-k)} \mathbb{E} P_\beta(A_{k+1:K} | u, A_{1:k}, c_{A_k}) \pi(A_{k+1:K} | u, A_{1:k}; V, \theta) \]

\[
= \sum_{A_{k+1:K} \in G(A|A_{1:k}, K-k)} \beta(A_{k+1:K} | u, A_{1:k}; V) \pi(A_{k+1:K} | u, A_{1:k}; V, \theta) \]

\[
= \sum_{A_{k+1:K} \in G(A|A_{1:k}, K-k)} \pi(A_{k+1:K} | u, A_{1:k}; V, \theta) \]

\[
= \sum_{A_{k+1:K} \in G(A|A_{1:k}, K-k)} \pi(A_{k+1:K} | u, A_{1:k}; V, \theta), \]  \hspace{1cm} (A.10)

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

**Proof of Theorem 2.**

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

\[
\mathcal{V}^\pi_{\text{PS}} = \mathbb{E}_{P_\beta} \left[ \sum_{k=1}^{K} w_A c_{A_k} \right] \]

\[
= \sum_{k=1}^{K} \mathbb{E}_{P_\beta} \left[ w_A c_{A_k} \right] \]

\[
= \sum_{k=1}^{K} \mathbb{E}_{P_\beta}(u_A, A_{1:k}, c_{A_k}) \mathbb{E}_{P_\beta}(A_{k+1:K} | u_A, A_{1:k}, c_{A_k}) \left[ \left[ \prod_{j=1}^{k} w_{A_j} \right] \right] \]

\[
\times \left[ \left[ \prod_{j=k+1}^{K} w_{A_j} \right] c_{A_k} \right] \]  \hspace{1cm} (A.15)

\[
= \sum_{k=1}^{K} \mathbb{E}_{P_\beta}(u_A, A_{1:k}, c_{A_k}) \mathbb{E}_{P_\beta}(A_{k+1:K} | u_A, A_{1:k}, c_{A_k}) \left[ \left[ \prod_{j=1}^{k} w_{A_j} \right] \right] \]

\[
\times \mathbb{E}_{P_\beta}(A_{k+1:K} | u_A, A_{1:k}, c_{A_k}) \left[ \left[ \prod_{j=k+1}^{K} w_{A_j} \right] c_{A_k} \right] \]  \hspace{1cm} (A.16)

\[
= \sum_{k=1}^{K} \mathbb{E}_{P_\beta}(u_A, A_{1:k}, c_{A_k}) \left[ \left[ \prod_{j=1}^{k} w_{A_j} \right] c_{A_k} \right] \]  \hspace{1cm} (A.17)

\[
= \sum_{k=1}^{K} \mathbb{E}_{P_\beta}(u_A, A_{1:k}, c_{A_k}) \mathbb{E}_{P_\beta}(A_{k+1:K} | u_A, A_{1:k}, c_{A_k}) \left[ \left[ \prod_{j=1}^{k} w_{A_j} \right] c_{A_k} \right] \]  \hspace{1cm} (A.18)

\[
= \sum_{k=1}^{K} \mathbb{E}_{P_\beta} \left[ \left[ \prod_{j=1}^{k} w_{A_j} \right] c_{A_k} \right], \]  \hspace{1cm} (A.19)

\[
= \mathbb{E}_{P_\beta} \left[ \sum_{k=1}^{K} \left[ \prod_{j=1}^{k} w_{A_j} \right] c_{A_k} \right]. \]  \hspace{1cm} (A.20)

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

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

Let \( \hat{\gamma}_{ij}, \tilde{\gamma}_{ij} \) be the second-order approximation of \( \hat{\gamma}_{ij}(\hat{Y}_{ij}) \) at \( \tilde{y}_j \). By denoting \( \hat{\gamma}^*_{ij}(\tilde{y}_j) \) as \( E_{j_0} \) and \( \nabla \hat{\gamma}^*_{ij}(\tilde{y}_j) \) as \( E_{j_2} \), we have

\[
T(\hat{Y}_{ij}, \tilde{y}_j) = E_{j_0} + E_{j_1}(\hat{Y}_{ij} - \tilde{y}_j) + \frac{E_{j_2}}{2} (\hat{Y}_{ij} - \tilde{y}_j)^2
\]

\[
= \frac{1}{2} E_{j_2} (\hat{Y}_{ij} - \frac{E_{j_1} - E_{j_2} \tilde{y}_j}{E_{j_2}})^2
\]

\[
+ (E_{j_0} - E_{j_1} \tilde{y}_j + \frac{1}{2} E_{j_2} \tilde{y}_j^2 - \frac{1}{2} (E_{j_1} - E_{j_2} \tilde{y}_j)^2)
\]  \hspace{1cm} (A.21)

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