A. Appendix: Counterfactual Learning to Rank

We also consider another important partial information setting: ranking evaluation and learning to rank based on implicit feedback (e.g. clicks, dwell time). Here the selection bias on the feedback signal is strongly influenced by position bias, since items lower in the ranking are less likely to be discovered by the user. However, it can be shown that this bias can be estimated (Joachims et al., 2017; Wang et al., 2018; Agarwal et al., 2019), and that the resulting estimates can serve as propensities in IPS-style estimators.

To connect the ranking setting with the contextual bandit setting more formally, now each context $x \sim P(\mathcal{X})$ represents a query and/or user profile. Given ranking function π, we use $\pi(x)$ to represent the ranking for query x. However, in the LTR setting, we no longer consider the actions atomic, but instead treat rankings as combinatorial actions where the reward decomposes as a weighted sum of component rewards. Formally speaking, the reward for rankings $\pi(x)$ is denoted as $\Delta(\pi(x)|x, r) := \sum_{d \in \mathbf{d}} \lambda(rank(d|\pi(x)))r(x, d)$, where $\lambda(\cdot)$ is a function that maps a rank to a score, $\mathbf{d}$ is the candidate set for query x and $rank(d|\pi(x))$ represents the rank of document d in the candidate set $\mathbf{d}$ under the ranking policy π given context x, and $r(x, d) \in \{0, 1\}$ is the relevance indicator. Then we can define the overall reward for a ranking policy π as

$$R(\pi) = \int \Delta(\pi(x)|x, r) dP(x) \tag{15}$$

Note that in this partial information setting we typically do not observe rewards for all (query, document) pairs. The only observable reward per component may be whether the user clicks the document or not, $c(x, \pi_0(x), d) \in \{0, 1\}$, and there is inherent ambiguity whether the lack of a click means lack of relevance or lack of discovery. Here we use a latent variable $o(x, \pi_0(x), d) \in \{0, 1\}$ to represent whether the user x observes document d under the logging policy π_0, which then leads to the following click model: a user clicks a document when the user observes it and the document is relevant, $c(x, \pi_0(x), d) = r(x, d) \cdot o(x, \pi_0(x), d)$. Note that the examination $o(x, \pi_0(x), d)$ is not observed by the system, but one can estimate a missingness model (Joachims et al., 2017), and use $p(x, \pi_0(x), d)$ be the (estimated) probability of $\mathbb{1}\{o(x, \pi_0(x), d) = 1\}$. We denote this probability value as the propensity of the observation. In practice one could estimate the propensities as outlined in (Agarwal et al., 2019; Fang et al., 2019). The logged data we get is in the format of $\mathcal{S} = \{\{(x_i, d_{ij}, p_{ij}, c_{ij})\}_{j=1}^{m_i}\}_{i=1}^{n}$ where m_i is the number of candidates for context x_i, p_{ij} is $p(x_i, \pi_0(x_i), d_{ij})$ and c_{ij} is $c(x_i, \pi_0(x_i), d_{ij})$. For additive ranking metrics, various estimators in the Interpolated Counterfactual Estimator Family can be written in the form

$$\hat{\mathrm{R}}^{\mathbf{w}}(\pi) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} \left[w_{ij}^{\alpha} \alpha_{ij} + o_{ij} w_{ij}^{\beta} \beta_{ij} + o_{ij} w_{ij}^{\gamma} \gamma_{ij} \right] \cdot \lambda(rank(d_{ij}|x_i, \pi(x_i)))$$

$$\text{with} \quad \alpha_{ij} := \hat{\delta}(x_i, d_{ij}), \beta_{ij} := \frac{r_{ij}}{p_{ij}}, \gamma_{ij} := \frac{\hat{\delta}(x_i, d_{ij})}{p_{ij}}.$$

where w_{ij}^{α}, w_{ij}^{β} and w_{ij}^{γ} are the weight functions of the three components of the Interpolated Counterfactual Estimator Family. Given a perfect reward model and logged propensities, we can get unbiased estimate of additive ranking metrics if the weight functions sum to 1. The weights of various estimators in this setting are shown in Table 3.

Table 3. The weight functions for different estimators of the family $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ in the ranking setting.

Estimator	w_{ij}^{α}	w_{ij}^{β}	w_{ij}^{γ}
DM	1	0	0
IPS	0	1	0
cIPS	0	$\min\{Mp_{ij}, 1\}$	0
SB	$1 - \tau$	τ	0
SWITCH	$\mathbb{1}\left\{\frac{1}{p_{ij}} > M\right\}$	$\mathbb{1}\left\{\frac{1}{p_{ij}} \leq M\right\}$	0
CAB	$1 - \min\{Mp_{ij}, 1\}$	$\min\{Mp_{ij}, 1\}$	0

Note that estimators with $w_{ij}^{\gamma} \neq 0$ (DR, CAB-DR) are not applicable in this setting since the third term $o_{ij} w_{ij}^{\gamma} \gamma_{ij}$ depends on o_{ij}, which is not observed nor fully captured by c_{ij}. However, the second term is computable since the unobserved o_{ij} and r_{ij} are captured by c_{ij} through $o_{ij}\beta_{ij} = \frac{o_{ij} r_{ij}}{p_{ij}} = \frac{c_{ij}}{p_{ij}}$. The SWITCH estimator is applicable for learning in this setting since the weights of the estimator do not depend on the ranking policy to be learned.

B. Appendix: Proofs

In this appendix, we provide proofs of the main theorems.

B.1. Proof of Theorem 1

Theorem 1 (Bias of the Interpolated Counterfactual Estimator Family)**.** *For contexts $x_1, x_2, \cdots, x_n$ drawn i.i.d from some distribution $P(\mathcal{X})$ and for actions $y_i \sim \pi_0(\mathcal{Y}|x_i)$, under Condition 1 the bias of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ with weighting functions $\mathbf{w} = (\mathrm{w}^{\alpha}, \mathrm{w}^{\gamma}, \mathrm{w}^{\gamma})$ is*

$$\begin{aligned}\mathbb{E}_x\,\mathbb{E}_{y\sim\pi}\Big[\mathrm{w}^{\alpha}\,\Delta - \mathrm{w}^{\beta}\,\zeta\delta + \mathrm{w}^{\gamma}(\Delta - \zeta(\delta + \Delta))\\ + (\mathrm{w}^{\alpha} + \mathrm{w}^{\beta} + \mathrm{w}^{\gamma})\delta - \delta\Big]\end{aligned} \tag{8}$$

Proof. For simplicity, we make the following notations throughout the proof. We let $\zeta := \zeta(x, y)$ denote the multiplicative deviation of the propensity estimate from the true propensity model, and $\Delta := \Delta(x, y)$ be the additive deviation of the reward model from the true reward. Recall

$$\zeta(x, y) = 1 - \frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)} \tag{16}$$

$$\Delta(x, y) = \hat{\delta}(x, y) - \delta(x, y). \tag{17}$$

Moreover, the $\sigma_r^2 := \sigma_r^2(x, y)$ is used to denote the randomness in reward $r(x, y)$ with $\sigma_r^2(x, y) = \mathbb{V}_r(r(x, y)|x, y)$. Moreover, we denote the true IPS weight $\frac{\pi(y|x)}{\pi_0(y|x)}$ as $c(x, y)$ with the estimated version being $\hat{c}(x, y)$. Also, let $\mathrm{w}^{\alpha} := \mathrm{w}^{\alpha}_{\mathrm{xy}}, \mathrm{w}^{\beta} := \mathrm{w}^{\beta}_{\mathrm{xy}}$ and $\mathrm{w}^{\gamma} := \mathrm{w}^{\gamma}_{\mathrm{xy}}$ be the abbreviation for the weighting functions.

We will start the proof by calculating the expectation of three different components of $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$. For the $\alpha_{i\bar{y}}$ component, this term is independent of the distribution of y_i, and we have:

$$\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\sum_{\bar{y}\in\mathcal{Y}}\pi(\bar{y}|x)\,\mathrm{w}^{\alpha}_{\mathrm{i}\bar{\mathrm{y}}}\,\alpha_{i\bar{y}}\Big] = \mathbb{E}_x\Big[\sum_{\bar{y}\in\mathcal{Y}}\pi(\bar{y}|x)\,\mathrm{w}^{\alpha}_{\mathrm{x}\bar{\mathrm{y}}}\,\hat{\delta}(x, \bar{y})\Big] = \mathbb{E}_x\,\mathbb{E}_{y\sim\pi}\Big[\mathrm{w}^{\alpha}_{\mathrm{xy}}(\delta + \Delta)\Big] \tag{18}$$

For the IPS term β_i, with $x_i \sim P(\mathcal{X})$ and $y_i \sim \pi_0(\mathcal{Y}\,|x)$.

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\pi(y_i|x_i)\,\mathrm{w}^{\beta}_{\mathrm{i}}\,\beta_i\Big] &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi_0}\,\mathbb{E}_r\Big[\mathrm{w}^{\beta}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\hat{\pi}_0(y|x)}r(x, y)\Big]\\ &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi_0}\Big[\mathrm{w}^{\beta}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\pi_0(y|x)}\frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)}\delta\Big]\\ &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi_0}\Big[c\,\mathrm{w}^{\beta}_{\mathrm{xy}}(1 - \zeta)\delta\Big]\\ &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi}\Big[\mathrm{w}^{\beta}_{\mathrm{xy}}(1 - \zeta)\delta\Big]\end{aligned} \tag{19}$$

where the second equation follows from the fact that conditioning on (x, y), $\mathbb{E}_r[r(x, y)|x, y] = \delta(x, y)$. For the third term γ_i, we have

$$\begin{aligned}\mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\pi(y_i|x_i)\,\mathrm{w}^{\gamma}_{\mathrm{i}}\,\gamma_i\Big] &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi_0}\Big[\mathrm{w}^{\gamma}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\hat{\pi}_0(y|x)}\hat{\delta}(x, y)\Big]\\ &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi_0}\Big[\mathrm{w}^{\gamma}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\pi_0(y|x)}\frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)}(\delta + \Delta)\Big]\\ &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi_0}\Big[c\,\mathrm{w}^{\gamma}_{\mathrm{xy}}(1 - \zeta)(\delta + \Delta)\Big]\\ &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi}\Big[\mathrm{w}^{\gamma}_{\mathrm{xy}}(1 - \zeta)(\delta + \Delta)\Big]\end{aligned} \tag{20}$$

Combining these three terms and using the formula that $Bias(\hat{\mathrm{R}}^{\mathbf{w}}(\pi)) = \mathbb{E}[\hat{\mathrm{R}}^{\mathbf{w}}(\pi)] - \mathbb{E}_x\,\mathbb{E}_{y\sim\pi}\,\mathbb{E}_r[r]$, we have

$$Bias(\hat{\mathrm{R}}^{\mathbf{w}}(\pi)) = \mathbb{E}_x\,\mathbb{E}_{y\sim\pi}\left[\mathrm{w}^{\alpha}\,\Delta - \mathrm{w}^{\beta}\,\zeta\delta + \mathrm{w}^{\gamma}(\Delta - \zeta(\delta+\Delta)) + (\mathrm{w}^{\alpha} + \mathrm{w}^{\beta} + \mathrm{w}^{\gamma})\delta - \delta\right] \tag{21}$$

□

B.2. Proof of Theorem 2

Theorem 2 (Variance of the Interpolated Counterfactual Estimator Family)**.** *Under the same conditions as in Theorem 1, the variance of* $\hat{\mathrm{R}}^{\mathbf{w}}(\pi)$ *with weighting functions* $\mathbf{w} = (\mathrm{w}^{\alpha}, \mathrm{w}^{\beta}, \mathrm{w}^{\gamma})$ *is*

$$\begin{aligned}
&\frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_{\pi}[\mathrm{w}^{\alpha}\,\Delta - \mathrm{w}^{\beta}\,\zeta\delta + \mathrm{w}^{\gamma}(\Delta - \zeta(\delta+\Delta))\\
&+ (\mathrm{w}^{\alpha} + \mathrm{w}^{\beta} + \mathrm{w}^{\gamma})\delta]\Big) + \mathbb{E}_x\mathbb{E}_{\pi}\Big[(\mathrm{w}^{\beta})^2 c(1-\zeta)^2\sigma_r^2\Big]\\
&+ \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(\mathrm{w}^{\beta}\,c(1-\zeta)\delta + \mathrm{w}^{\gamma}\,c(1-\zeta)(\delta+\Delta))\Big]\Big\}
\end{aligned} \tag{9}$$

Proof. We follow the same notation as in Appendix B.1. Let $\hat{\mathrm{R}}^{\mathbf{w}}_{\mathrm{i}}(\pi) := \sum_{\bar{y}\in\mathcal{Y}} \pi(\bar{y}|x_i)\; w^{\alpha}_{i\bar{y}}\;\alpha_{i\bar{y}} + \pi(y_i|x_i)\; w^{\beta}_i\;\beta_i + \pi(y_i|x_i)\; w^{\gamma}_i\;\gamma_i$ with the abbreviated version defined as $\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)$, and it is easy to see that $\mathbb{V}(\hat{\mathrm{R}}^{\mathbf{w}}(\pi)) = \frac{1}{n}\,\mathbb{V}(\hat{\mathrm{R}}^{\mathbf{w}}_{\mathrm{i}}(\pi))$.

$$\begin{aligned}
\mathbb{V}(\hat{\mathrm{R}}^{\mathbf{w}}_{\mathrm{i}}(\pi)) &= \mathbb{V}_x\left(\mathbb{E}_{y\sim\pi_0,r}[\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x]\right) + \mathbb{E}_x\left[\mathbb{V}_{y\sim\pi_0,r}(\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x)\right]\\
&= \mathbb{V}_x\left(\mathbb{E}_{y\sim\pi_0,r}[\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x]\right) + \mathbb{E}_x\left[\mathbb{E}_{y\sim\pi_0}[\mathbb{V}_r(\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y)|x]\right] + \mathbb{E}_x\left[\mathbb{V}_{y\sim\pi_0}(\mathbb{E}_r[\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y]|x)\right]
\end{aligned} \tag{22}$$

For the first term, using the bias formula in Appendix B.1, it is easy to see that

$$\mathbb{V}_x\left(\mathbb{E}_{y\sim\pi_0,r}[\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x]\right) = \mathbb{V}_x\left(\mathbb{E}_{y\sim\pi}[\mathrm{w}^{\alpha}_{\mathrm{xy}}\,\Delta - \mathrm{w}^{\beta}_{\mathrm{xy}}\,\zeta\delta + \mathrm{w}^{\gamma}_{\mathrm{xy}}(\Delta - \zeta(\delta+\Delta)) + (\mathrm{w}^{\alpha}_{\mathrm{x}\bar{\mathrm{y}}} + \mathrm{w}^{\beta}_{\mathrm{xy}} + \mathrm{w}^{\gamma}_{\mathrm{xy}})\delta|x]\right) \tag{23}$$

For the second term, we will calculate $\mathbb{V}_r(\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y)$ first.

$$\begin{aligned}
\mathbb{V}_r(\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y) &= \mathbb{V}_r\left(\mathrm{w}^{\beta}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\hat{\pi}_0(y|x)}r(x,y)|x,y\right)\\
&= \mathbb{V}_r\left(\mathrm{w}^{\beta}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\pi_0(y|x)}\,\frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)}r|x,y\right)\\
&= c^2(\mathrm{w}^{\beta}_{\mathrm{xy}})^2(1-\zeta)^2\,\mathbb{V}_r(r|x,y)\\
&= c^2(\mathrm{w}^{\beta}_{\mathrm{xy}})^2(1-\zeta)^2\sigma_r^2
\end{aligned} \tag{24}$$

where the first equality follows from the fact that conditioning on (x,y), $\sum_{\bar{y}\in\mathcal{Y}} \pi(\bar{y}|x)\;\mathrm{w}^{\alpha}_{\mathrm{x}\bar{\mathrm{y}}}\;\alpha_{x\bar{y}} + \pi(y|x)\;\mathrm{w}^{\gamma}_{\mathrm{xy}}\;\gamma_{xy}$ is just a constant, and we use the formula $\mathbb{V}(a+X) = \mathbb{V}(X)$ for any constant a, random variable X.

Then for the term $\mathbb{E}_x\left[\mathbb{E}_{y\sim\pi_0}[\mathbb{V}_r(\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y)|x]\right]$, we have

$$\begin{aligned}
\mathbb{E}_x\left[\mathbb{E}_{y\sim\pi_0}[\mathbb{V}_r(\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y)|x]\right] &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi_0}\left[c^2(\mathrm{w}^{\beta}_{\mathrm{xy}})^2(1-\zeta)^2\sigma_r^2\right]\\
&= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi}\left[c(\mathrm{w}^{\beta}_{\mathrm{xy}})^2(1-\zeta)^2\sigma_r^2\right]
\end{aligned} \tag{25}$$

Similarly, for the third term, we will calculate $\mathbb{E}_r[\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y]$ first.

$$\begin{aligned}
\mathbb{E}_r[\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y] &= \mathbb{E}_r\Big[\sum_{\bar{y}\in\mathcal{Y}} \pi(\bar{y}|x)\,\mathrm{w}^{\alpha}_{\mathrm{x\bar{y}}}\,\alpha_{i\bar{y}} + \pi(y|x)\,\mathrm{w}^{\beta}_{\mathrm{xy}}\,\beta_{xy} + \pi(y|x)\,\mathrm{w}^{\gamma}_{\mathrm{xy}}\,\gamma_{xy}|x,y\Big] \\
&= \sum_{\bar{y}\in\mathcal{Y}} \pi(\bar{y}|x)\,\mathrm{w}^{\alpha}_{\mathrm{x\bar{y}}}\,\hat{\delta}(x,\bar{y}) + \mathbb{E}_r\Big[\mathrm{w}^{\beta}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\pi_0(y|x)}\frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)}r|x,y\Big] + \mathrm{w}^{\gamma}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\pi_0(y|x)}\frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)}(\delta+\Delta) \\
&= \sum_{\bar{y}\in\mathcal{Y}} \pi(\bar{y}|x)\,\mathrm{w}^{\alpha}_{\mathrm{x\bar{y}}}\,\hat{\delta}(x,\bar{y}) + \mathrm{w}^{\beta}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\pi_0(y|x)}\frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)}\delta + \mathrm{w}^{\gamma}_{\mathrm{xy}}\,\frac{\pi(y|x)}{\pi_0(y|x)}\frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)}(\delta+\Delta) \\
&= \sum_{\bar{y}\in\mathcal{Y}} \pi(\bar{y}|x)\,\mathrm{w}^{\alpha}_{\mathrm{x\bar{y}}}\,\hat{\delta}(x,\bar{y}) + \mathrm{w}^{\beta}_{\mathrm{xy}}\,c(1-\zeta)\delta + \mathrm{w}^{\gamma}_{\mathrm{xy}}\,c(1-\zeta)(\delta+\Delta)
\end{aligned} \tag{26}$$

For the term $\mathbb{V}_{y\sim\pi_0}\Big(\mathbb{E}_r[\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y]|x\Big)$, since the first term $\sum_{\bar{y}\in\mathcal{Y}}\pi(\bar{y}|x)\,\mathrm{w}^{\alpha}_{\mathrm{x\bar{y}}}\,\hat{\delta}(x,\bar{y})$ is independent of y, then we have

$$\begin{aligned}
\mathbb{V}_{y\sim\pi_0}\Big(\mathbb{E}_r[\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y]|x\Big) &= \mathbb{V}_{y\sim\pi_0}\Big(\sum_{\bar{y}\in\mathcal{Y}} \pi(\bar{y}|x_i)\,\mathrm{w}^{\alpha}_{\mathrm{x\bar{y}}}\,\hat{\delta}(x_i,\bar{y}) + \mathrm{w}^{\beta}_{\mathrm{xy}}\,c(1-\zeta)\delta + \mathrm{w}^{\gamma}_{\mathrm{xy}}\,c(1-\zeta)(\delta+\Delta)|x\Big) \\
&= \mathbb{V}_{y\sim\pi_0}\Big(\mathrm{w}^{\beta}_{\mathrm{xy}}\,c(1-\zeta)\delta + \mathrm{w}^{\gamma}_{\mathrm{xy}}\,c(1-\zeta)(\delta+\Delta)|x\Big)
\end{aligned} \tag{27}$$

Then taking the outer expectation over x, we have:

$$\mathbb{E}_x\Big[\mathbb{V}_{y\sim\pi_0}(\mathbb{E}_r[\mathrm{R}^{\mathbf{w}}_{\mathrm{xy}}(\pi)\,|x,y]|x)\Big] = \mathbb{E}_x\Big[\mathbb{V}_{y\sim\pi_0}(\mathrm{w}^{\beta}_{\mathrm{xy}}\,c(1-\zeta)\delta + \mathrm{w}^{\gamma}_{\mathrm{xy}}\,c(1-\zeta)(\delta+\Delta)|x)\Big] \tag{28}$$

Summing all the three terms together, and using the formula $\mathbb{V}(\hat{\mathrm{R}}^{\mathbf{w}}(\pi)) = \frac{1}{n}\,\mathbb{V}(\hat{\mathrm{R}}^{\mathbf{w}}_{\mathrm{i}}(\pi))$ for $i.i.d$ $\hat{\mathrm{R}}^{\mathbf{w}}_{\mathrm{i}}$, we have:

$$\begin{aligned}
\mathbb{V}(\hat{\mathrm{R}}^{\mathbf{w}}(\pi)) =& \frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_\pi[\mathrm{w}^{\alpha}\,\Delta - \mathrm{w}^{\beta}\,\zeta\delta + \mathrm{w}^{\gamma}(\Delta-\zeta(\delta+\Delta)) + (\mathrm{w}^{\alpha}+\mathrm{w}^{\beta}+\mathrm{w}^{\gamma})\delta]\Big) \\
&+ \mathbb{E}_x\mathbb{E}_\pi\Big[(\mathrm{w}^{\beta})^2 c(1-\zeta)^2\sigma_r^2\Big] + \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(\mathrm{w}^{\beta}\,c(1-\zeta)\delta + \mathrm{w}^{\gamma}\,c(1-\zeta)(\delta+\Delta))\Big]\Big\}
\end{aligned} \tag{29}$$

□

B.3. Proof of Theorem 3

Theorem 3 (Bias of CAB). *For contexts $x_1, x_2, \cdots, x_n$ drawn i.i.d from some distribution $P(\mathcal{X})$ and for actions $y_i \sim \pi_0(\mathcal{Y}\,|x_i)$, under Condition 1 the bias of $\hat{R}_{CAB}(\pi)$ is*

$$\mathbb{E}_x\mathbb{E}_\pi[-\delta\zeta\mathbb{1}\{\hat{c}\le M\} + \{\Delta(1-\frac{M}{c(1-\zeta)}) - \frac{M}{c(1-\zeta)}\delta\zeta\}\mathbb{1}\{\hat{c} > M\}] \tag{30}$$

Proof. Note CAB falls into the class of counterfactual estimator with the weighting functions $\mathrm{w}^{\alpha}_{\mathrm{i\bar{y}}} = 1 - \min\Big\{M\frac{\hat{\pi}_0(\bar{y}|x_i)}{\pi(\bar{y}|x_i)}, 1\Big\}$, $\mathrm{w}^{\beta}_{\mathrm{i}} = \min\Big\{M\frac{\hat{\pi}_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\Big\}$, $\mathrm{w}^{\gamma}_{\mathrm{i}} = 0$.

Using Theorem 1, the bias for $\hat{R}_{CAB}(\pi)$ is:

$$\begin{aligned}
Bias(\hat{R}_{CAB}(\pi)) &= \mathbb{E}_x\,\mathbb{E}_{y\sim\pi}\Big[\mathrm{w}^{\alpha}_{\mathrm{xy}}\,\Delta - \mathrm{w}^{\beta}_{\mathrm{xy}}\,\zeta\delta + \mathrm{w}^{\gamma}_{\mathrm{xy}}(\Delta-\zeta(\delta+\Delta)) + (\mathrm{w}^{\alpha}_{\mathrm{x\bar{y}}}+\mathrm{w}^{\beta}_{\mathrm{xy}}+\mathrm{w}^{\gamma}_{\mathrm{xy}})\delta - \delta\Big] \\
&= \mathbb{E}_x\mathbb{E}_{y\sim\pi}\Big[(1-\min\{\frac{M}{\hat{c}(x,y)},1\})\Delta - \min\{\frac{M}{\hat{c}(x,y)},1\}\zeta\delta\Big] \\
&= \mathbb{E}_x\mathbb{E}_{y\sim\pi}\Big[-\zeta\delta\,\mathbb{1}\{\hat{\mathrm{c}}\le \mathrm{M}\} + \{(1-\frac{M}{\hat{c}(x,y)})\Delta - \frac{M}{\hat{c}(x,y)}\zeta\delta\}\,\mathbb{1}\{\hat{\mathrm{c}} > \mathrm{M}\}\Big] \\
&= \mathbb{E}_x\mathbb{E}_{y\sim\pi}\Big[-\zeta\delta\,\mathbb{1}\{\hat{\mathrm{c}}\le \mathrm{M}\} + \{(1-\frac{M}{c(1-\zeta)})\Delta - \frac{M}{c(1-\zeta)}\zeta\delta\}\,\mathbb{1}\{\hat{\mathrm{c}} > \mathrm{M}\}\Big]
\end{aligned} \tag{31}$$

while the last equality follows from the fact that $\hat{c}(x,y) := \frac{\pi(y|x)}{\hat{\pi}_0(y|x)} = \frac{\pi(y|x)}{\pi_0(y|x)}\frac{\pi_0(y|x)}{\hat{\pi}_0(y|x)} = c(x,y)(1-\zeta(x,y))$ □

B.4. Proof of Theorem 4

Theorem 4 (Variance of CAB). *Under the same conditions as in Theorem 3, the variance of* $\hat{R}_{CAB}(\pi)$

$$\mathbb{V}(\hat{R}_{CAB}(\pi)) = \frac{1}{n}\Big\{\mathbb{V}_x(\mathbb{E}_\pi[\delta - \delta\zeta\mathbb{1}\{\hat{c} \le M\} + (\Delta(1-\frac{M}{c(1-\zeta)}) - \frac{M}{c(1-\zeta)}\delta\zeta)\mathbb{1}\{\hat{c} > M\}]) \\ + \mathbb{E}_x\mathbb{E}_\pi[c(1-\zeta)^2\sigma_r^2\mathbb{1}\{\hat{c}\le M\} + \frac{M^2}{c}\sigma_r^2\mathbb{1}\{\hat{c}>M\}] + \mathbb{E}_x[\mathbb{V}_{\pi_0}(c(1-\zeta)\delta\mathbb{1}\{\hat{c}\le M\} + M\delta\mathbb{1}\{\hat{c}>M\})]\Big\} \tag{32}$$

Proof. The result follows by plugging in the weighting function for CAB with $\mathrm{w}^{\alpha}_{\mathrm{i\bar{y}}} = 1 - \min\Big\{M\frac{\hat{\pi}_0(\bar{y}|x_i)}{\pi(\bar{y}|x_i)}, 1\Big\}$, $\mathrm{w}^{\beta}_{\mathrm{i}} = \min\Big\{M\frac{\hat{\pi}_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\Big\}$, $\mathrm{w}^{\gamma}_{\mathrm{i}} = 0$ in Theorem 2.

For the term $\mathbb{V}_x\Big(\mathbb{E}_\pi[\mathrm{w}^\alpha\,\Delta - \mathrm{w}^\beta\,\zeta\delta + \mathrm{w}^\gamma(\Delta - \zeta(\delta+\Delta)) + (\mathrm{w}^\alpha + \mathrm{w}^\beta + \mathrm{w}^\gamma)\delta]\Big)$, using the result from Theorem 3, we have:

$$\begin{aligned}&\mathbb{V}_x\Big(\mathbb{E}_\pi[\mathrm{w}^\alpha\,\Delta - \mathrm{w}^\beta\,\zeta\delta + \mathrm{w}^\gamma(\Delta - \zeta(\delta+\Delta)) + (\mathrm{w}^\alpha + \mathrm{w}^\beta + \mathrm{w}^\gamma)\delta]\Big) \\ &= \mathbb{V}_x\Big(\mathbb{E}_\pi[\delta - \delta\zeta\mathbb{1}\{\hat{c}\le M\} + (\Delta(1-\frac{M}{c(1-\zeta)}) - \frac{M}{c(1-\zeta)}\delta\zeta)\mathbb{1}\{\hat{c}>M\}]\Big)\end{aligned} \tag{33}$$

For the term $\mathbb{E}_x\mathbb{E}_\pi\Big[(\mathrm{w}^\beta)^2c(1-\zeta)^2\sigma_r^2\Big]$, we have

$$\begin{aligned}\mathbb{E}_x\mathbb{E}_\pi\Big[(\mathrm{w}^\beta)^2c(1-\zeta)^2\sigma_r^2\Big] &= \mathbb{E}_x\mathbb{E}_\pi\Big[\min\{(\frac{M}{\hat{c}(x,y)})^2, 1\}c(1-\zeta)^2\sigma_r^2\Big] \\ &= \mathbb{E}_x\,\mathbb{E}_\pi\,\Big[c(1-\zeta)^2\sigma_r^2\,\mathbb{1}\{\hat{\mathrm{c}}\le\mathrm{M}\} + \frac{M^2}{\hat{c}^2(x,y)}c(1-\zeta)^2\sigma_r^2\,\mathbb{1}\{\hat{\mathrm{c}}>\mathrm{M}\}\,\Big] \\ &= \mathbb{E}_x\,\mathbb{E}_\pi\,\Big[c(1-\zeta)^2\sigma_r^2\,\mathbb{1}\{\hat{\mathrm{c}}\le\mathrm{M}\} + \frac{M^2}{c^2(1-\zeta)^2}c(1-\zeta)^2\sigma_r^2\,\mathbb{1}\{\hat{\mathrm{c}}>\mathrm{M}\}\,\Big] \\ &= \mathbb{E}_x\,\mathbb{E}_\pi\,\Big[c(1-\zeta)^2\sigma_r^2\,\mathbb{1}\{\hat{\mathrm{c}}\le\mathrm{M}\} + \frac{M^2}{c}\sigma_r^2\,\mathbb{1}\{\hat{\mathrm{c}}>\mathrm{M}\}\,\Big]\end{aligned} \tag{34}$$

For the last term $\mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(\mathrm{w}^\beta\,c(1-\zeta)\delta + \mathrm{w}^\gamma\,c(1-\zeta)(\delta+\Delta))\Big]$, then

$$\begin{aligned}\mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(\mathrm{w}^\beta\,c(1-\zeta)\delta)\Big] &= \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(\min\{\frac{M}{\hat{c}(x,y)}, 1\}c(1-\zeta)\delta)\Big] \\ &= \mathbb{E}_x\,\Big[\mathbb{V}_{\pi_0}(c(1-\zeta)\delta\,\mathbb{1}\{\hat{\mathrm{c}}\le\mathrm{M}\} + \frac{M}{\hat{c}(x,y)}c(1-\zeta)\delta\,\mathbb{1}\{\hat{\mathrm{c}}>\mathrm{M}\})\Big] \\ &= \mathbb{E}_x\,\Big[\mathbb{V}_{\pi_0}(c(1-\zeta)\delta\,\mathbb{1}\{\hat{\mathrm{c}}\le\mathrm{M}\} + M\delta\,\mathbb{1}\{\hat{\mathrm{c}}>\mathrm{M}\})\Big]\end{aligned} \tag{35}$$

Combining all, we have

$$\mathbb{V}(\hat{R}_{CAB}(\pi)) = \frac{1}{n}\Big\{\mathbb{V}_x\Big(\mathbb{E}_\pi[\delta - \delta\zeta\mathbb{1}\{\hat{c}\le M\} + (\Delta(1-\frac{M}{c(1-\zeta)}) - \frac{M}{c(1-\zeta)}\delta\zeta)\mathbb{1}\{\hat{c}>M\}]\Big) \\ + \mathbb{E}_x\mathbb{E}_\pi\Big[c(1-\zeta)^2\sigma_r^2\mathbb{1}\{\hat{c}\le M\} + \frac{M^2}{c}\sigma_r^2\mathbb{1}\{\hat{c}>M\}\Big] + \mathbb{E}_x\Big[\mathbb{V}_{\pi_0}(c(1-\zeta)\delta\mathbb{1}\{\hat{c}\le M\} + M\delta\mathbb{1}\{\hat{c}>M\})\Big]\Big\} \tag{36}$$

□

B.5. Proof of Bias and Variance of CAB-DR

Theorem 5 (Bias of CAB-DR). *For contexts $x_1, x_2, \cdots, x_n$ drawn i.i.d from some distribution $P(\mathcal{X})$ and for actions $y_i \sim \pi_0(\mathcal{Y}|x_i)$, under Condition 1 the bias of $\hat{R}_{CABDR}(\pi)$ is*

$$\mathbb{E}_x\,\mathbb{E}_\pi\left[\zeta\Delta\mathbb{1}\{\hat{c} \le M\} + \Delta(1 - \frac{M}{c})\mathbb{1}\{\hat{c} > M\}\right] \tag{37}$$

Proof. CAB-DR is also an instance in the Interpolated Counterfactual Estimator Family with the weighting function: $\mathrm{w}^{\alpha}_{\mathrm{i\bar{y}}} = 1, \mathrm{w}^{\beta}_{\mathrm{i}} = \min\left\{M\frac{\hat{\pi}_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\right\}, \mathrm{w}^{\gamma}_{\mathrm{i}} = -\min\left\{M\frac{\hat{\pi}_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\right\}$. Using Theorem 1, the bias for CAB-DR is:

$$\begin{aligned}
Bias(\hat{R}_{CABDR}(\pi)) &= \mathbb{E}_x\,\mathbb{E}_\pi\left[\mathrm{w}^{\alpha}\,\Delta - \mathrm{w}^{\beta}\,\zeta\delta + \mathrm{w}^{\gamma}(\Delta - \zeta(\delta+\Delta)) + (\mathrm{w}^{\alpha} + \mathrm{w}^{\beta} + \mathrm{w}^{\gamma})\delta - \delta\right] \\
&= \mathbb{E}_x\,\mathbb{E}_\pi\left[\Delta - \min\{\frac{M}{\hat{c}(x,y)}, 1\}\zeta\delta - \min\{\frac{M}{\hat{c}(x,y)}, 1\}(\Delta - \zeta(\delta+\Delta))\right] \\
&= \mathbb{E}_x\,\mathbb{E}_\pi\left[\Delta - \min\{\frac{M}{\hat{c}(x,y)}, 1\}(\Delta - \zeta\Delta)\right] \\
&= \mathbb{E}_x\,\mathbb{E}_\pi\left[\zeta\Delta\,\mathbb{1}\{\mathrm{\hat{c}} \le \mathrm{M}\} + \{\Delta[1 - \frac{M}{\hat{c}(x,y)}(1-\zeta)]\}\,\mathbb{1}\{\mathrm{\hat{c}} > \mathrm{M}\}\right] \\
&= \mathbb{E}_x\,\mathbb{E}_\pi\left[\zeta\Delta\mathbb{1}\{\hat{c} \le M\} + \Delta(1 - \frac{M}{c})\mathbb{1}\{\hat{c} > M\}\right]
\end{aligned} \tag{38}$$

□

Theorem 6 (Variance of CAB-DR). *Under the same conditions as in Theorem 3, the variance of $\hat{R}_{CABDR}(\pi)$*

$$\begin{aligned}
\mathbb{V}(\hat{R}_{CABDR}(\pi)) = \frac{1}{n}\Big\{&\mathbb{V}_x\left(\mathbb{E}_\pi[\delta + \zeta\Delta\mathbb{1}\{\hat{c} \le M\} + \Delta(1 - \frac{M}{c})\mathbb{1}\{\hat{c} > M\}]\right) \\
&+ \mathbb{E}_x\mathbb{E}_\pi\left[c(1-\zeta)^2\sigma_r^2\mathbb{1}\{\hat{c} \le M\} + \frac{M^2}{c}\sigma_r^2\mathbb{1}\{\hat{c} > M\}\right] \\
&+ \mathbb{E}_x\left[\mathbb{V}_{\pi_0}(c(1-\zeta)(-\Delta)\mathbb{1}\{\hat{c} \le M\} - M\Delta\mathbb{1}\{\hat{c} > M\})\right]\Big\}
\end{aligned} \tag{39}$$

Proof. The proof follows by using Theorem 2 with the weights $\mathrm{w}^{\alpha}_{\mathrm{i\bar{y}}} = 1, \mathrm{w}^{\beta}_{\mathrm{i}} = \min\left\{M\frac{\hat{\pi}_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\right\}, \mathrm{w}^{\gamma}_{\mathrm{i}} = -\min\left\{M\frac{\hat{\pi}_0(y_i|x_i)}{\pi(y_i|x_i)}, 1\right\}$.

For the first term, following directly from Appendix B.5, it is easy to see

$$\mathbb{V}_x\Big(\mathbb{E}_\pi[\mathrm{w}^{\alpha}\,\Delta - \mathrm{w}^{\beta}\,\zeta\delta + \mathrm{w}^{\gamma}(\Delta - \zeta(\delta+\Delta)) + (\mathrm{w}^{\alpha} + \mathrm{w}^{\beta} + \mathrm{w}^{\gamma})\delta]\Big) = \mathbb{V}_x\left(\mathbb{E}_\pi[\delta + \zeta\Delta\mathbb{1}\{\hat{c} \le M\} + \Delta(1 - \frac{M}{c})\mathbb{1}\{\hat{c} > M\}]\right) \tag{40}$$

For the second term $\mathbb{E}_x\mathbb{E}_\pi[(\mathrm{w}^{\beta})^2c(1-\zeta)^2\sigma_r^2]$, since CAB and CAB-DR has the same weighting function w^{β}, this term is exactly the same for CAB and CAB-DR.

For the third term $\mathbb{E}_x\left[\mathbb{V}_{\pi_0}(\mathrm{w}^{\beta}\,c(1-\zeta)\delta + \mathrm{w}^{\gamma}\,c(1-\zeta)(\delta+\Delta))\right]$, we have

$$\begin{aligned}
\mathbb{E}_x\left[\mathbb{V}_{\pi_0}(\mathrm{w}^{\beta}\,c(1-\zeta)\delta + \mathrm{w}^{\gamma}\,c(1-\zeta)(\delta+\Delta))\right] &= \mathbb{E}_x\left[\mathbb{V}_{\pi_0}(\min\{\frac{M}{\hat{c}(x,y)}, 1\}c(1-\zeta)\delta - \min\{\frac{M}{\hat{c}(x,y)}, 1\}c(1-\zeta)(\delta+\Delta))\right] \\
&= \mathbb{E}_x\left[\mathbb{V}_{\pi_0}(-\min\{\frac{M}{\hat{c}(x,y)}, 1\}c(1-\zeta)\Delta)\right] \\
&= \mathbb{E}_x\left[\mathbb{V}_{\pi_0}(c(1-\zeta)(-\Delta)\,\mathbb{1}\{\mathrm{\hat{c}} \le \mathrm{M}\} - M\Delta\,\mathbb{1}\{\mathrm{\hat{c}} > \mathrm{M}\})\right]
\end{aligned} \tag{41}$$

Combining all the three terms will give us the variance for $\hat{R}_{CABDR}(\pi)$. □

C. Experiment Details

In this section, we provide experiment details for both the BLBF and LTR settings.

C.1. BLBF

In the BLBF experiment, specifically, given a supervised dataset $\{(x_i, y_i^*)\}_{i=1}^n$, where x is $i.i.d$ drawn from a certain fixed distribution $P(\mathcal{X})$ and $y^* \in \{1, 2, \cdots, k\}$ denotes the true class label. For a particular logging policy π_0, the logged bandit data is simulated by sampling $y_i \sim \pi_0(\mathcal{Y}|x_i)$ and a deterministic loss $r(x_i, y_i)$ is revealed. In our experiments, the loss is defined as $r(x_i, y_i) = \mathbb{1}\{y_i \neq y_i^*\} - 1$. The resulting logged contextual bandit data $\mathcal{S} = \{x_i, y_i, r(x_i, y_i), \pi_0(y_i|x_i)\}$ is then used to evaluate the performance of different estimators.

For evaluation, we split each dataset equally into train and test sets. For the train set, we use 10% of the full-information data to train the logger π_0 and loss predictor $\pi_r(x)$, with loss estimates defined by $\hat{\delta}(x_i, y) = \mathbb{1}\{\pi_r(x_i) \neq y\} - 1$. The policy π we want to evaluate is a multiclass logistic regression trained on the whole train set. Finally, we use the full-information test set to generate the contextual bandit datasets $\mathcal{S}$ for off-policy evaluation of sizes $n = 200, 500, 2000$. We evaluate the policy π with different estimators on the logged bandit feedback of different sizes and treat the performance on the full-information test set as ground truth $R(\pi)$. The performance is measured by MSE. We repeat each experiment 500 times and calculate the bias, variance and MSE.

For learning, we first split the original dataset into training (48%), validation (32%) and test sets (20%). Following (Swaminathan & Joachims, 2015a), the policy we want to learn lies on the space $\mathcal{F} := \{\pi_w : w \in \mathbb{R}^p\}$ with π_w as the stochastic linear rules defined by:

$$\pi_w(y|x) = \frac{\exp(w^T \phi(x, y))}{\mathbb{Z}(x)} \tag{42}$$

Here, $\phi(x, y)$ denotes the joint feature map between context x and action y, and $\mathbb{Z}(x)$ is a normalization factor. The training objective is defined by $\pi^{est} = \mathrm{argmin}_{\pi_w \in \mathcal{F}} \hat{R}^{est}(\pi_w) + \lambda ||w||_2$, where λ is selected through the lowest $\hat{R}^{IPS}(\pi)$ on the validation set. To avoid local minimum, the objective is optimized via L-BFGS using scikit-learn with 10 random starts. The performance of the learned policy π^{est} is measured via expected error on the test set, defined as: $\frac{1}{n_{test}} \sum_{i=1}^{n_{test}} \mathbb{E}_{y_i \sim \pi^{est}(\mathcal{Y}|x_i)}[\mathbb{1}\{y_i \neq y_i^*\}]$. Similar to evaluation, we use 20% of the training data to train the multiclass logistics regression as logging policy with default hyperparameter. While for the estimated loss $\hat{\delta}(x, y)$, we train the logistic regression using 10% training data with tuned hyperparameter selected from the validation set. All the result is averaged over 10 runs with $n = 5000$.

C.2. LTR

In the LTR experiment, we use 10% of the training set for learning a DM, which reflects that we typically have a small amount of manual relevance judgements. The DM is a binary Gradient Boosted Decision Tree Classifier calibrated by Sigmoid Calibration (Platt et al., 1999). We use $\lambda(rank) = rank$ as the performance metric which can be interpreted as the average rank of the relevant results. The examination probability (propensity) that we use is $p(x, \pi_0(x), d) = \frac{1}{rank(d|\pi_0(x))}$. For the evaluation experiments, to get a ranking policy for evaluation, we train a ranking SVM (Joachims, 2002) on the remaining 90% training data. As input to the estimators, different amounts of click data are generated from the test set. For each experiment, we generate the log data 100 times and report the bias, variance, and MSE with respect to the estimated ground truth from the full-information test set.

For the learning experiments, we derived a concrete learning algorithm based on propensity SVM-Rank that conducts learning from biased user feedback using different estimators. The SVM-style algorithm (Joachims, 2002; 2006; Joachims et al., 2017) optimizes an upper bound on different estimators and details are in Appendix C.3. We compare the performance of different estimators using different amounts of simulated user feedback with the proposed learning algorithm. As input to the propensity SVM-Rank, different amount of click data is simulated from the 90% training data. Specifically, we present the performance using 1 sweep of the data in Table 2. We grid search C for propensity SVM-Rank and M for different estimators and conduct hyperparameter selection with 90 percentile cIPS on user feedback data simulated from the validation set for 5 sweeps. All the experiments are run for 5 times and the average is presented.

C.3. Generalized Propensity SVM-Rank

We now derive a concrete learning algorithm that conducts learning from biased user feedback using different estimators from the Interpolated Counterfactual Estimator Family. It is based on SVM-Rank (Joachims, 2002; 2006; Joachims et al., 2017) but we expect other learning to rank methods can also be adapted to the estimators.

The generalized propensity SVM-Rank learns a linear scoring function $f(x, d) = w \cdot \phi(x, d)$ with $\phi(x, d)$ describing how context x and document d interact. It optimizes the following objective

$$\begin{aligned} \hat{w} = argmin_{w,\xi} \frac{1}{2} w \cdot w + \frac{C}{n} \sum_i \sum_j \left[w_{ij}^{\alpha} \alpha_{ij} + o_{ij} w_{ij}^{\beta} \beta_{ij} \right] \sum_{k \neq j} \xi_{ijk} \\ s.t. \quad \forall i, j, k \neq j \quad w \cdot [\phi(x_i, d_{ij}) - \phi(x_i, d_{ik})] > 1 - \xi_{ijk}, \\ \forall i, j, k \neq j \quad \xi_{ijk} \geq 0 \end{aligned} \tag{43}$$

where w is the parameter of the generalized propensity SVM-Rank and C is a regularization parameter. The training objective optimizes an upper bound on the estimator with average rank of positive examples metric($\lambda(rank) = rank$) since

$$\begin{aligned} & \sum_i \sum_j \left[w_{ij}^{\alpha} \alpha_{ij} + o_{ij} w_{ij}^{\beta} \beta_{ij} \right] (rank(d_{ij} | x_i, \pi(x_i)) - 1) \\ = & \sum_i \sum_j \left[w_{ij}^{\alpha} \alpha_{ij} + o_{ij} w_{ij}^{\beta} \beta_{ij} \right] \cdot \sum_{k \neq j} \mathbb{1}\{ w \cdot [\phi(x_i, d_{ik}) - \phi(x_i, d_{ij})] > 0 \} \\ \leq & \sum_i \sum_j \left[w_{ij}^{\alpha} \alpha_{ij} + o_{ij} w_{ij}^{\beta} \beta_{ij} \right] \cdot \sum_{k \neq j} \max(1 - w \cdot [\phi(x_i, d_{ij}) - \phi(x_i, d_{ik})], 0) \\ \leq & \sum_i \sum_j \left[w_{ij}^{\alpha} \alpha_{ij} + o_{ij} w_{ij}^{\beta} \beta_{ij} \right] \cdot \sum_{k \neq j} \xi_{ijk} \end{aligned}$$