Supplementary Material

Proof to Theorem 1

Proof. Let us define

$$\mathrm{E}_\delta^j[y] = \mathrm{E}\left[\delta(Y,y)K_j(X^1,X^2)\right]$$

Under the i.i.d. assumption, it is straightforward to show that

$$\mathrm{E}[a_p^j[y]] = \mathrm{E}[a_\delta^j[y]] = \mathrm{E}_\delta^j[y]$$

Following the McDiarmid's inequality, for any $\epsilon > 0$, we have

$$\Pr\left(\left|a_p^j[y] - \mathrm{E}_\delta^j[y]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{2\kappa_j^2}\right)$$

$$\Pr\left(\left|a_\delta^j[y] - \mathrm{E}_\delta^j[y]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{2\kappa_j^2}\right)$$

Using the following inequality and the union bound,

$$\left|a_\delta^j[y] - a_p^j[y]\right| \leq \left|a_\delta^j[y] - \mathrm{E}_\delta^j[y]\right| + \left|a_p^j[y] - \mathrm{E}_\delta^j[y]\right|$$

we can complete the theorem. $\square$

Proof to Theorem 3

Proof. Let us define

$$\widehat{\mathrm{E}}_\delta^j[y] = \mathrm{E}\left[\delta(\widehat{Y},y)K_j(X^1,X^2)\right]$$

Using the assumption (1.a) and (1.b), we have

$$\begin{aligned}
&\widehat{\mathrm{E}}_\delta^j[y] = \mathrm{E}_{X^1,X^2}\mathrm{E}_{\widehat{Y}|X^1,X^2}[\delta(\widehat{Y},y)K_j(X^1,X^2)] \\
&= \mathrm{E}_{X^1,X^2}[\Pr(\widehat{Y}=y|X^1,X^2)K_j(X^1,X^2)] \\
&= \mathrm{E}_{X^1,X^2}\left[c_y \Pr(Y=y|X^1,X^2)K_j(X^1,X^2)\right] \\
&+ \mathrm{E}_{X^1,X^2}\left[(1-c_{\bar{y}})\Pr(Y=\bar{y}|X^1,X^2)K_j(X^1,X^2)\right] \\
&= c_y\mathrm{E}_\delta^j[y] + (1-c_{\bar{y}})\mathrm{E}_\delta^j[\bar{y}] \\
&= (c_y + c_{\bar{y}} - 1)\mathrm{E}_\delta^j[y] + (1-c_{\bar{y}})\mathrm{E}[K_j(X^1,X^2)]
\end{aligned}$$

where we use the fact $\mathrm{E}_\delta^j[y] + \mathrm{E}_\delta^j[\bar{y}] = \mathrm{E}[K_j(X^1,X^2)]$. Let us define

$$\widehat{c}_\delta^j[y] = (c_y + c_{\bar{y}} - 1)a_\delta^j[y] + (1-c_{\bar{y}})\frac{1}{n}\sum_i K_j(\mathbf{x}_i^1,\mathbf{x}_i^2)$$

Under the i.i.d. assumption, it is straightforward to show that

$$\mathrm{E}[\widehat{a}_\delta^j[y]] = \mathrm{E}[\widehat{c}_\delta^j[y]] = \widehat{\mathrm{E}}_\delta^j[y]$$

Following the McDiarmid's inequality, for any $\epsilon > 0$, we have

$$\Pr\left(\left|\widehat{a}_\delta^j[y] - \widehat{\mathrm{E}}_\delta^j[y]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{2\kappa_j^2}\right)$$

$$\Pr\left(\left|\widehat{c}_\delta^j[y] - \widehat{\mathrm{E}}_\delta^j[y]\right| \geq \epsilon\right) \leq 2\exp\left(-\frac{\epsilon^2 n}{2\kappa_j^2}\right)$$

Then we have

$$\Pr\left(\left|\widehat{c}_\delta^j[y] - \widehat{a}_\delta^j[y]\right| \geq \epsilon\right) \leq 4\exp\left(-\frac{\epsilon^2 n}{8\kappa_j^2}\right)$$

Dividing both sides of $\left|\widehat{c}_\delta^j[y] - \widehat{a}_\delta^j[y]\right| \geq \epsilon$ by $c_y + c_{\bar{y}} - 1$, we have

$$\Pr\left(\left|a_\delta^j[y] - \widehat{b}_\delta^j[y]\right| \geq \frac{\epsilon}{c_y + c_{\bar{y}} - 1}\right) \leq 4\exp\left(-\frac{\epsilon^2 n}{8\kappa_j^2}\right)$$

Replacing ϵ with $(c_y + c_{\bar{y}} - 1)\epsilon$, we complete the proof. $\square$

Proof to Theorem 5

Proof. Let

$$\begin{aligned}
L(\lambda) &= \frac{1}{n}\sum_i \ln(\exp(\lambda_1^\top \mathbf{k}_i) + \exp(\lambda_0^\top \mathbf{k}_i)) \\
&\quad - \lambda_1^\top \widehat{\mathbf{b}}_1^* - \lambda_0^\top \widehat{\mathbf{b}}_0^* + \frac{\gamma}{2}\|\lambda_1\|_2^2 + \frac{\gamma}{2}\|\lambda_0\|_2^2 \\
&= g(\lambda) - tr(\lambda^\top \widehat{\mathbf{b}}^*) + \frac{\gamma}{2}\|\lambda\|_F^2
\end{aligned}$$

where $\lambda = (\lambda_1, \lambda_0)$, $\widehat{\mathbf{b}}^* = \left(\widehat{\mathbf{b}}_1^*, \widehat{\mathbf{b}}_0^*\right)$, $g(\lambda)$ is the sum of log-exponential function of λ, which is convex in λ. Assume λ^* is the optimal solution to minimizing $L(\lambda)$, λ^o is the optimal solution to minimizing $L(\lambda)$ with $\widehat{\mathbf{b}}^* = \left(\widehat{\mathbf{b}}_1^*, \widehat{\mathbf{b}}_0^*\right)$ replaced by $\widehat{\mathbf{b}}^o = \left(\widehat{\mathbf{b}}_1^o, \widehat{\mathbf{b}}_0^o\right)$, then we have

$$\begin{aligned}
L(\lambda^o) &\geq L(\lambda^*) + tr\left(\nabla L(\lambda^*)^\top(\lambda^o - \lambda^*)\right) + \frac{\gamma}{2}\|\lambda^o - \lambda^*\|_F^2 \\
&\geq L(\lambda^*) + \frac{\gamma}{2}\|\lambda^o - \lambda^*\|_F^2
\end{aligned}$$

where we use the fact that $L(\cdot)$ is a c_r-strongly convex function, and the optimality criterion that $tr\left(\nabla L(\lambda^*)^\top(\lambda^o - \lambda^*)\right) \geq 0$. Then

$$
\begin{aligned}
& L(\lambda^o) = g(\lambda^o) - tr({\lambda^o}^\top \widehat{\mathbf{b}}^*) + \frac{\gamma}{2}\|\lambda^o\|_F^2 \\
& = g(\lambda^o) - tr({\lambda^o}^\top \widehat{\mathbf{b}}^o) + \frac{\gamma}{2}\|\lambda^o\|_F^2 + tr({\lambda^o}^\top(\widehat{\mathbf{b}}^o - \widehat{\mathbf{b}}^*)) \\
& \leq g(\lambda^*) - tr({\lambda^*}^\top \widehat{\mathbf{b}}^o) + \frac{\gamma}{2}\|\lambda^*\|_F^2 + tr({\lambda^o}^\top(\widehat{\mathbf{b}}^o - \widehat{\mathbf{b}}^*)) \\
& \leq g(\lambda^*) - tr({\lambda^*}^\top \widehat{\mathbf{b}}^*) + \frac{\gamma}{2}\|\lambda^*\|_F^2 \\
& \quad + tr((\lambda^o - \lambda^*)^\top(\widehat{\mathbf{b}}^o - \widehat{\mathbf{b}}^*)) \\
& \leq L(\lambda^*) + \|\lambda^* - \lambda^o\|_F \|\widehat{\mathbf{b}}^* - \widehat{\mathbf{b}}^o\|_F
\end{aligned}
$$

Coming the above two bounds for $L(\lambda^o)$ together, we have

$$
\frac{\gamma}{2}\|\lambda^* - \lambda^o\|_F^2 \leq \|\lambda^* - \lambda^o\|_F \|\widehat{\mathbf{b}}^* - \widehat{\mathbf{b}}^o\|_F
$$

i.e.,

$$
\|\lambda^* - \lambda^o\|_F \leq \frac{2}{\gamma}\|\widehat{\mathbf{b}}^* - \widehat{\mathbf{b}}^o\|_F
$$

□

Proof of Theorem 6

Proof. Let $\lambda = (\lambda_1, \lambda_0)$ be the solution to (4) using noisy side information with $\widehat{\mathbf{b}} = (\widehat{\mathbf{b}}_1, \widehat{\mathbf{b}}_0)$, and $\lambda^* = (\lambda_1^*, \lambda_0^*)$ be the solution to (2) using the perfect side information, i.e. solution to (4) with $\widehat{\mathbf{b}}$ replaced by $\widehat{\mathbf{b}}^* = (\mathbf{a}_1, \mathbf{a}_0)$, where $\mathbf{a}_1, \mathbf{a}_0$ are defined as

$$
\begin{aligned}
\mathbf{a}_1 &= (a_\delta^1[y=1], \ldots, a_\delta^m[y=1])^\top \\
\mathbf{a}_0 &= (a_\delta^1[y=-1], \ldots, a_\delta^m[y=-1])^\top
\end{aligned}
$$

First, we have

$$
\begin{aligned}
& |\widehat{p}(y|\mathbf{x}^1, \mathbf{x}^2) - p(y|\mathbf{x}^1, \mathbf{x}^2)| \\
= \quad & \frac{\left|(\lambda_1 - \lambda_1^* + \lambda_2 - \lambda_2^*)^\top \mathbf{k}(\mathbf{x}^1, \mathbf{x}^2)\right|}{1 + \exp\left((\widetilde{\lambda}_1 - \widetilde{\lambda}_2)^\top \mathbf{k}(\mathbf{x}^1, \mathbf{x}^2)\right)} \\
\leq \quad & \|\lambda - \lambda^*\|_F \kappa \leq \frac{2\kappa}{\gamma}\|\widehat{\mathbf{b}} - \widehat{\mathbf{b}}^*\|_F \\
\leq \quad & \frac{2\kappa}{\gamma} \sum_{y \in \{-1,+1\}} \sum_{j=1}^m \left|a_\delta^j[y] - \widehat{b}_\delta^j[y]\right|
\end{aligned}
$$

where we use the mean value theorem and $\widetilde{\lambda}$ is a point on the line segment $[\lambda, \lambda^*]$. Using Theorem 3, we have the following inequalities hold with probability at least $1 - \delta$,

$$
\left|a_\delta^j[y] - \widehat{b}_\delta^j[y]\right| \leq \frac{\kappa_j}{c}\sqrt{\frac{8}{n}\ln\frac{8m}{\delta}}, j = 1, \ldots, m, y = \pm 1
$$

We complete the proof by combining the above results.

□

Proof of Theorem 7

Proof. We define

$$
\widetilde{b}_\delta^j[y] = \frac{\widehat{a}_\delta^j[y]}{\widehat{c}} - \frac{1 - \widehat{c}_{\bar{y}}}{n\widehat{c}} \sum_{i=1}^n K_j(\mathbf{x}_i^1, \mathbf{x}_i^2)
$$

where $\widehat{c} = \widehat{c}_+ + \widehat{c}_- - 1$. Define

$$
\begin{aligned}
\widetilde{\mathbf{b}}_1 &= (\widetilde{b}_\delta^1[y=1], \ldots, \widetilde{b}_\delta^m[y=1])^\top \\
\widetilde{\mathbf{b}}_0 &= (\widetilde{b}_\delta^1[y=-1], \ldots, \widetilde{b}_\delta^m[y=-1])^\top
\end{aligned}
$$

Let $\widetilde{p}(y|\mathbf{x}^1, \mathbf{x}^2)$ be the classification model learned from the noisy side information using corrupted $\widehat{c}_+$ and $\widehat{c}_-$, and $\widehat{p}(y|\mathbf{x}^1, \mathbf{x}^2)$ be the classification model learned from the noisy side information using perfect c_+ and c_-. Using the analysis in the proof of Theorem 6, we have

$$
\begin{aligned}
\left|\widetilde{p}(y|\mathbf{x}^1, \mathbf{x}^2) - \widehat{p}(y|\mathbf{x}^1, \mathbf{x}^2)\right| \quad & \leq \quad \frac{2}{\gamma}\left\|\widetilde{\mathbf{b}} - \widehat{\mathbf{b}}\right\|_F \\
& \leq \quad \frac{2}{\gamma} \sum_{y=\pm 1} \sum_{j=1}^m |\widetilde{b}_\delta^j[y] - \widehat{b}_\delta^j[y]|
\end{aligned}
$$

Since

$$
\left|\widetilde{b}_\delta^j[y] - \widehat{b}_\delta^j[y]\right| \leq 2\kappa\left|\frac{1}{\widehat{c}} - \frac{1}{c}\right| \leq \frac{8\kappa\Delta}{\rho^2}
$$

we have

$$
\left|\widetilde{p}(y|\mathbf{x}^1, \mathbf{x}^2) - \widehat{p}(y|\mathbf{x}^1, \mathbf{x}^2)\right| \leq \frac{32\kappa m\Delta}{\gamma\rho^2}
$$

Using the fact

$$
\begin{aligned}
& \left|\widetilde{p}(y|\mathbf{x}^1, \mathbf{x}^2) - p(y|\mathbf{x}^1, \mathbf{x}^2)\right| \leq \\
& \quad \left|\widetilde{p}(y|\mathbf{x}^1, \mathbf{x}^2) - \widehat{p}(y|\mathbf{x}^1, \mathbf{x}^2)\right| + \left|\widehat{p}(y|\mathbf{x}^1, \mathbf{x}^2) - p(y|\mathbf{x}^1, \mathbf{x}^2)\right|
\end{aligned}
$$

and the result in Theorem 6, we have the theorem. □