Appendix A. Decompose the loss

We decompose the new loss:

$$\mathbb{E}_{\widetilde{D}}[\ell_B(\boldsymbol{\omega}) + \boldsymbol{\beta}\boldsymbol{\ell}_A(\boldsymbol{\omega})] \tag{14}$$

For simplicity, we omit $\boldsymbol{\omega}$ from $\ell f(x, \boldsymbol{\omega})$ in the following derivations. So we first decompose the first term: $\mathbb{E}_{\widetilde{D}}[\ell_B(\boldsymbol{\omega})]$

$$\begin{aligned}
&\mathbb{E}_{\widetilde{D}}[\ell_B(\boldsymbol{\omega})] \\
&= \sum_k \sum_l \sum_a \int_x P(Y = k \mid Z = l, A = a, x) P(x \mid Z = l, A = a) \delta_a \ell(f(x), k) dx P(Z = l) P(A = a) \\
&= \sum_k \sum_l \sum_a P(Z = l) P(A = a) \mathbb{E}_{D_{x|Z=l,A=a}}[P(Y = k \mid Z = l, A = a) \delta_a \ell(f(x), k)] \\
&= \sum_k \sum_l \sum_a P(Z = l) P(A = a) [\underbrace{\mathbb{E}_{D_{x|Z=l,A=a}} P(Y = k \mid Z = l, A = a) \cdot \mathbb{E}_{D_{x|Z=l,A=a}} \delta_a \ell(f(x), k)}_{A} \\
&+ \underbrace{\mathrm{Cov}_{D_{x|Z=l,A=a}}[P(Y = k \mid Z = l, A = a) \cdot \delta_a \ell(f(x), k)]}_{B}]
\end{aligned} \tag{15}$$

Expand Part A in Eq. (15), we can obtain:

$$\begin{aligned}
&\sum_a P(A = a) \delta_a [P(Z = 1) \cdot \mathbb{E}_{D_{x|Z=1,A=a}}(1 - \theta_a^-) \cdot \mathbb{E}_{D_{x|Z=1,A=a}} \ell(f(x), 1) \\
&+ P(Z = -1) \cdot \mathbb{E}_{D_{x|Z=-1,A=a}}(1 - \theta_a^+) \cdot \mathbb{E}_{D_{x|Z=-1,A=a}} \ell(f(x), -1)] \\
&+ \sum_a P(A = a) \delta_a \sum_k \sum_{l, l \neq k} P(Z = l) \cdot \mathbb{E}_{D_{x|Z=l,A=a}} P(Y = k \mid Z = l, A = a) \cdot \mathbb{E}_{D_{x|Z=l,A=a}} \ell(f(x), k)] \\
&= \sum_a P(A = a) \delta_a [\underbrace{\mathbb{E}_{D_x}(1 - \theta_a^- - \theta_a^+) \cdot \mathbb{E}_{D|a} \ell(f(x), Z))}_{C} \\
&+ \underbrace{P(Z = 1) \cdot \mathbb{E}_{D_{x|Z=1,A=a}} \theta_a^+ \cdot \mathbb{E}_{D_{x|Z=1,A=a}} \ell(f(x), 1) + P(Z = -1) \cdot \mathbb{E}_{D_{x|Z=-1,A=a}} \theta_a^- \cdot \mathbb{E}_{D_{x|Z=-1,A=a}} \ell(f(x), -1)]}_{D} \\
&+ \underbrace{\sum_a P(A = a) \delta_a \sum_k \sum_{l, l \neq k} P(Z = l) \cdot \mathbb{E}_{D_{x|Z=l,A=a}} P(Y = k \mid Z = l, A = a) \cdot \mathbb{E}_{D_{x|Z=l,A=a}} \ell(f(x), k)}_{E}
\end{aligned} \tag{16}$$

Expand part B in Eq. (15), we can get:

$$\begin{aligned}
&\sum_a P(A = a) \delta_a [\sum_k P(Z = k) \mathbb{E}_{D_{x|Z=k,A=a}}((P(Y = k \mid Z = k, A = a) - \mathbb{E}_{D_{x|Z=k,A=a}}(P(Y = k \mid Z = k, A = a)) \\
&\times (\ell(f(x), k) - \mathbb{E}_{D_{x|Z=k,A=a}}[\ell(f(x), k]) \\
&+ \sum_k \sum_{l, l \neq k} P(Z = l) \mathbb{E}_{D_{x|Z=l,A=a}}((P(Y = k \mid Z = l, A = a) - \mathbb{E}_{D_{x|Z=l,A=a}}(P(Y = k \mid Z = l, A = a)) \\
&\times (\ell(f(x), k) - \mathbb{E}_{D_{x|Z=l,A=a}}[\ell(f(x), k])]
\end{aligned} \tag{17}$$

If we combine Eq. (17) with Part E in Eq. (16), we can obtain:

$$\sum_a P(A=a)\delta_a \sum_k [\sum_{l,l\neq k} P(Z=l)\mathbb{E}_{D_{x|Z=l,A=a}}(P(Y=k\mid Z=l,A=a)\ell(f(x),k)$$
$$+P(Z=k)\mathbb{E}_{D_{x|Z=k,A=a}}((P(Y=k\mid Z=k,A=a)-\mathbb{E}_{D_{x|Z=k,A=a}}(P(Y=k\mid Z=k,A=a))[\ell(f(x),k)]]$$
$$=\sum_a P(A=a)\delta_a[P(Z=1)\mathbb{E}_{D_{x|Z=1,A=a}}(1-\theta_a^- - \mathbb{E}_{D_{x|Z=1,A=a}}(1-\theta_a^-))\ell(f(x),1)$$
$$+P(Z=-1)\mathbb{E}_{D_{x|Z=-1,A=a}}(1-\theta_a^+ - \mathbb{E}_{D_{x|Z=-1,A=a}}(1-\theta_a^+))\ell(f(x),-1)$$
$$+P(Z=-1)\mathbb{E}_{D_{x|Z=-1,A=a}}(\theta_a^+\ell(f(x),1)] + P(Z=1)\mathbb{E}_{D_{x|Z=1,A=a}}(\theta_a^-\ell(f(x),-1)] \tag{18}$$

Finally, we combine Eq. (18) with part C as well as part D in Eq. (16) and we can finally get the decomposed terms:

$$\mathbb{E}_{\widetilde{D}}[\ell_B(\boldsymbol{\omega})]$$
$$=\sum_a P(A=a)\delta_a[(1-\theta_a^+-\theta_a^-)\mathbb{E}_{D|A=a}\ell(f(x),Z)+\sum_k\sum_l P(Z=l)\mathbb{E}_{D_{x|l,a}}\delta_a\theta_a^{\mathrm{sgn}(k)}\ell(f(x),k)]$$
$$=\sum_a P(A=a)[\mathbb{E}_{D|A=a}\ell(f(x),Z)+\sum_{k\in[C]}\sum_{l\in[C]}P(Z=l)\mathbb{E}_{D_{x|l,a}}\delta_a\theta_a^{\mathrm{sgn}(k)}\ell(f(x),k)]$$
$$=\mathbb{E}_D[\ell(f(X),Z)]+\sum_a P(A=a)\sum_k\sum_l P(Z=l)\mathbb{E}_{D_{x|l,a}}\delta_a\theta_a^{\mathrm{sgn}(k)}\ell(f(x),k) \tag{19}$$

Now we then decompose the second and third term in Eq. (5).

$$\mathbb{E}_{\widetilde{D}}[\boldsymbol{\beta\ell}_A(\boldsymbol{\omega})]$$
$$=\mathbb{E}_{\widetilde{D}}[-\beta_0\cdot\mathbb{E}_{Y|\widetilde{D},A=0}(1-a_i)\ell(f(x),Y)-\beta_1\cdot\mathbb{E}_{Y|\widetilde{D},A=1}a_i\ell(f(x),Y)]$$
$$=\mathbb{E}_{\widetilde{D}}[\lambda\cdot(\mathbb{E}_{Y|\widetilde{D},A=0}(1-a_i)\ell(f(x),Y)-\mathbb{E}_{Y|\widetilde{D},A=1}a_i\ell(f(x),Y))$$
$$+(-\beta_0-\lambda)\cdot\mathbb{E}_{Y|\widetilde{D},A=0}(1-a_i)\ell(f(x),Y)+(-\beta_1+\lambda)\cdot\mathbb{E}_{Y|\widetilde{D},A=1}a_i\ell(f(x),Y)]$$
$$=\lambda\cdot[\mathbb{E}_{\widetilde{D}|A=0}\ell(f(x),Y)-\mathbb{E}_{\widetilde{D}|A=1}\ell_{A=1}(f(x),Y)]+(-\beta_0-\lambda)\int_x\sum_k P(X=x,Y=k,A=0)(1-0)\ell(f(x),k)dx$$
$$+(-\beta_1+\lambda)\int_x\sum_k P(X=x,Y=k,A=1)(1)\ell(f(x),k)dx$$
$$=\lambda\cdot\mathbb{E}_{\widetilde{D}}[\ell_{A=0}(f(x),Y)-\ell_{A=1}(f(x),Y))-P(A=0)(\beta_0+\lambda)\sum_k\sum_l P(Z=l)\mathbb{E}_{D_{x|l,0}}P(Y=k)\ell(f(x),k)$$
$$-P(A=1)(\beta_1-\lambda)\sum_k\sum_l P(Z=l)\mathbb{E}_{D_{x|l,1}}P(Y=k)\ell(f(x),k)$$
$$=\lambda\cdot[\mathbb{E}_{\widetilde{D}|A=0}\ell(f(x),Y)-\mathbb{E}_{\widetilde{D}|A=1}\ell(f(x),Y)]-\sum_a P(A=a)\sum_k\sum_l P(Z=l)\mathbb{E}_{D_{x|l,a}}\gamma_a\cdot P(Y=k)\ell(f(x),k) \tag{20}$$

where $\gamma_a=\begin{cases}\beta_0+\lambda \text{ if } a=0\\ \beta_1-\lambda \text{ if } a=1\end{cases}$. Without loss of generality, we assume $\mathbb{E}_{Y|\widetilde{D},A=0}(1-a_i)\ell(f(x),Y)>\mathbb{E}_{Y|\widetilde{D},A=1}a_i\ell(f(x),Y)$ for Eq. (20).

Appendix B. Derive the relationship between selection bias and label bias

Let $\widetilde{N}_{\text{sgn}(y),a}$, $\hat{N}_{\text{sgn}(y),a}$ and $N_{\text{sgn}(y),a}$ denote the number of instances in group with membership of $(\text{sgn}(y), a)$. Here $\widetilde{N}_{.}$ is for the observed data with both biases. $\hat{N}_{.}$ is for the data with selection bias only.

$$\widetilde{N}_{+1,1} = (1 - \theta_1^-) \cdot \hat{N}_{+1,1} + \theta_1^+ \cdot \hat{N}_{-1,1} \tag{21}$$

Let ε_0^- denotes the bias rate combining the selection bias and label bias.

$$\widetilde{N}_{+1,1} = (1 - \varepsilon_1^-) \cdot N_{+1,1} + \varepsilon_1^+ \cdot N_{-1,1} \tag{22}$$

We assume the selection bias is proportion to the ratio of positive labeled instances in unprotected group, i.e.,

$$\begin{aligned} \frac{\hat{N}_{+1,1}}{\hat{N}_{+1,1} + N_{-1,1}} = \frac{r}{\sigma} = \frac{N_{+1,1}}{\sigma(N_{+1,1} + N_{-1,1})} \\ \hat{N}_{+1,1} = \frac{1-r}{\sigma - r} N_{+1,1} \end{aligned} \tag{23}$$

Then we can derive the relationship between ε_1^+ and θ_1^+ by

$$\begin{aligned} (1 - \varepsilon_1^-) \cdot N_{+1,1} + \varepsilon_1^+ \cdot N_{-1,1} = (1 - \theta_1^-) \cdot \hat{N}_{+1,1} + \theta_1^+ \cdot \hat{N}_{-1,1} \\ (1 - \theta_1^+) \frac{1-r}{\sigma - r} N_{+1,1} = (1 - \varepsilon_1^-) N_{+1,1} \\ \theta_1^- = \frac{\sigma - r}{1 - r} \varepsilon_1^- + \frac{1 - \sigma}{1 - r} \end{aligned} \tag{24}$$

Appendix C. Synthetic data generating process

- Generate $W \sim N(0, \sigma)$ (we use $\sigma = I^{15 \times 15}$, and dimension of W is 15).

- Generate $a_i \sim \text{Bernoulli}(\alpha)$, (we set $\alpha = 0.1$ and n = 2000).
- Generate $x_i^j \sim \text{Bernoulli}(\frac{1}{j+1}^r)$ for $j = 0, ..., k-2$, where k is the dimension of W, which is 15. r controls the discrepancy between the rarity of features. We sample each dimension i according to a Bernoulli proportional to $\frac{1}{i}$ making some dimensions common and others rare (we set $r = 0.5$).

- Generate unbiasd label $z_i = \max(0, \text{sign}(w_{\text{gen}}^T x_i))$

- Generate biased label $y_i \sim g(y \mid z_i, a_i, x_i, \beta)$ where $g(y_i \mid z_i, a_i, x_i, \beta) = \begin{cases} \beta \text{ if } y_i \neq z_i \wedge z = a_i \\ 1 - \beta \end{cases}$ and β controls the amount of label bias (We set $\beta = 0.5$).