Proofs

Miaoyan Wang, May 20, 2020

Proof. Note that $f \in \mathcal{F}$ implies that, there exist $\boldsymbol{B}^{\text{true}} \in \mathbb{R}^{d_1 \times d_2}$, such that $\text{rank}(\boldsymbol{B}^{\text{true}}) \leq r$ and f has the following representation,

$$\mathbb{P}(Y=1|\boldsymbol{X}) = f(\boldsymbol{X}) = \langle \boldsymbol{X},\ \boldsymbol{B}^{\text{true}} \rangle, \quad \text{for all } \boldsymbol{X} \in \mathbb{R}^{d_1 \times d_2}.$$

The low-rank SMM minimizes the sample version of the following population function

$$L_\pi(\boldsymbol{B}) \stackrel{\text{def}}{=} \mathbb{E}[\ell_\pi(\boldsymbol{B})] = \mathbb{E}\left\{\sum_{y_i=1}[1-\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle]_+ + \sum_{y_i=-1}[1+\langle \boldsymbol{X}_i, \boldsymbol{B}\rangle]_+\right\},$$

where the expectation is over $(\boldsymbol{X}_i, y_i) \sim_{\text{i.i.d}} \mathcal{X} \times \mathcal{Y}$. Straightforward calculation shows that

$$\frac{1}{n}L_\pi(\boldsymbol{B}) = \mathbb{E}\left\{[1-Y\langle \boldsymbol{X}, \boldsymbol{B}\rangle]_+ \mathbb{1}\left(Y=1\right) + [1-Y\langle \boldsymbol{X}, \boldsymbol{B}\rangle]_+ \mathbb{1}\left(Y=-1\right)\right\},$$

where $(\boldsymbol{X}, y) \sim \mathcal{X} \times \mathcal{Y}$. Let $\hat{\boldsymbol{B}} = \arg\min_{\{\boldsymbol{B}:\, \text{rank}(\boldsymbol{B}) \leq r]} L(\boldsymbol{B})$. We will prove that $\hat{\boldsymbol{B}} = \boldsymbol{B}^{\text{true}}$. Note that

$$\begin{aligned}
\frac{1}{n}L_\pi(\hat{\boldsymbol{B}}) &= \mathbb{E}\left\{\left[1-Y\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle\right]_+ \mathbb{1}\left(Y=1\right) + \left[1-Y\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle\right]_+ \mathbb{1}\left(Y=-1\right)\right\} \\
&= \mathbb{E}\{\left(1-Y\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle\right)\mathbb{1}(Y=1)\} \\
&= \mathbb{E}(Y=1) - \mathbb{E}\left\{Y\mathbb{1}(Y=1)\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle\right\}.
\end{aligned}$$

We note that

$$\begin{aligned}
\mathbb{E}\left\{Y\mathbb{1}(Y=1)\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle\right\} &= \mathbb{E}_{\boldsymbol{X}}\left\{\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle \mathbb{E}_{(\boldsymbol{X},Y)}[Y\mathbb{1}(Y=1)|\boldsymbol{X}]\right\} \\
&= \mathbb{E}_{\boldsymbol{X}}\left\{\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle \left\{\mathbb{P}(\boldsymbol{Y}=1|\boldsymbol{X}) - \frac{1}{2}\right)\right\} \\
&= \mathbb{E}_{\boldsymbol{X}}\left\{\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle \left(\langle \boldsymbol{X}, \boldsymbol{B}^{\text{true}}\rangle - \frac{1}{2}\right)\right\}.
\end{aligned}$$

Therefore, the optimal $\hat{\boldsymbol{B}}$ must satisfy $\text{sign}\langle \boldsymbol{X}, \hat{\boldsymbol{B}}\rangle = \text{sign}\left(\langle \boldsymbol{X}, \boldsymbol{B}^{\text{true}}\rangle - \frac{1}{2}\right)$. $\square$