Supplementary Material

We present the simulation Scenario II in which the imputation model is correctly specified and exactly sparse in Appendix A. We also compared the SAS estimation and inference with unsupervised regression that uses only \( S \) to derive a proxy outcome in the simulation Scenario III very similar to Scenario I. The proofs of Theorems 1, 7, Corollary 5 and Propositions 9 and 10 are given in Appendix B. The technical details are put in Appendix C. Definitions and existing results are stated in Appendix D.

Appendix A. Additional Simulation

Scenario II: The imputation model is correctly specified and exactly sparse. In the second scenario, we first generate \( S_i \) from
\[
S_{i,1} = \left[ \varsigma \{ \nu Z_{i,1}^s + \alpha^T (Z_{i}^x \sqrt{1 - p^{-1}} + Z_{i}^u / \sqrt{p}) \} - \mu_S \right] / \sigma_S.
\]
and \( S_{i,j} = \{ \varsigma (Z_{i,j}^s) - \mu_X \} / \sigma_X \) for \( j = 2, \ldots, p \), and then generate \( Y_i \) from a sparse model
\[
\mathbb{P}(Y_i = 1 | X_i) = \text{expit}(\theta S_{i,1}).
\]
We chose \( \mu_s \approx 0.66 \) and \( \sigma_s \approx 1 \) such that \( S_{i,1} \) is roughly mean 0 and variance 1. Under this setting, the imputation model holds with \( s_\gamma = 1 \). The factor \( \nu \) and the coefficients \( \alpha \) control the predictiveness of \( X \) for \( S_1 \) and \( Y \) while \( \theta \) controls the predictiveness of \( W \) for \( Y \). We consider two \( \alpha \) of different sparsity patterns,

- Sparse \( (s_\alpha = 3) \): \( \alpha = (0.3, 0.212, 0.212, 0_97^T) \)
- Dense \( (s_\alpha = 500) \): \( \alpha = (0.211, 0.039_29^T, 0.004_470^T) \),

where \( a_{k \times 1} = (a, \ldots, a)^{T}_{k \times 1} \) for any \( a \). Similar to Scenario I, the sparsity of \( \alpha \) regulates the approximate sparsity of \( \beta \) measured by (33) (See Table 1). We consider two sets of \( (\nu, \theta) \) to allow \( W \) to be either moderately or strongly predictive of \( Y \):

- Moderate: \( \nu = 0.4, \theta = 2 \); and
- Strong: \( \nu = 0.6, \theta = 3.7 \).

The layouts of Tables A2 and A3 are different from those of 2 and Table 3 because of the different data generating mechanism. The distribution of \( Y_i | X_i \) is not affected by the distribution of \( S_i \) in Scenario I, while the property does not hold in Scenario II.

Table A1: AUC Table for simulations with 500 labels under Scenario II. The AUCs are evaluated on an independent testing set of size 100. We approximately measure the sparsity by \( S(\nu) = \|\nu\|_1^2 / \|\nu\|_2^2 \).

| Scenario | Prediction Accuracy (AUC) |
|----------|---------------------------|
| Surrogate | \( S(\beta) \) | \( S(\gamma) \) | Oracle | SLASSO | SAS |
| Strong   | 159 | 1.10 | 0.715 | 0.660 | 0.702 |
| Moderate | 128 | 1.06 | 0.715 | 0.665 | 0.704 |
| Strong   | 26.4 | 1.09 | 0.710 | 0.691 | 0.708 |
| Moderate | 18.4 | 1.03 | 0.709 | 0.693 | 0.707 |
Table A2: Comparison of SAS Estimation to the supervised LASSO (SLASSO) with Bias, Empirical standard error (ESE) and root mean-squared error (rMSE) of the linear predictions $x^T_{\text{new}} \hat{\beta}$ under Scenario II with 500 labels, moderate or large $S(\beta_0)$ and strong or moderate surrogates.

| Moderate Surrogates | Strong Surrogates |
|---------------------|-------------------|
| SLASSO             | SAS               | SLASSO           | SAS               | SLASSO          | SAS           |
| Type               | Bias | ESE  | rMSE | Bias | ESE  | rMSE | Bias | ESE  | rMSE | Bias | ESE  | rMSE | Bias | ESE  | rMSE |
| Risk prediction model approximately sparse |
| $x^L_{\text{new}}$ | 0.505 | 0.378 | 0.631 | 0.163 | 0.283 | 0.327 | 0.349 | 0.278 | 0.446 | 0.316 | 0.253 | 0.328 |
| $x^M_{\text{new}}$ | -0.140 | 0.331 | 0.359 | -0.047 | 0.272 | 0.276 | -0.113 | 0.282 | 0.304 | -0.058 | 0.217 | 0.225 |
| $x^H_{\text{new}}$ | -0.713 | 0.512 | 0.878 | -0.262 | 0.313 | 0.408 | -0.678 | 0.469 | 0.925 | -0.210 | 0.241 | 0.320 |
| $x^S_{\text{new}}$ | -0.111 | 0.143 | 0.181 | -0.058 | 0.081 | 0.190 | -0.190 | 0.142 | 0.237 | -0.037 | 0.063 | 0.072 |
| $x^I_{\text{new}}$ | -0.437 | 0.098 | 0.448 | -0.119 | 0.076 | 0.141 | -0.155 | 0.098 | 0.183 | -0.065 | 0.061 | 0.089 |
| $x^D_{\text{new}}$ | -0.349 | 0.093 | 0.361 | -0.138 | 0.078 | 0.158 | -0.150 | 0.093 | 0.176 | -0.112 | 0.063 | 0.129 |
| Large $S(\beta_0)$ |
| $x^L_{\text{new}}$ | 0.366 | 0.266 | 0.453 | 0.142 | 0.224 | 0.265 | 0.482 | 0.398 | 0.625 | 0.117 | 0.300 | 0.322 |
| $x^M_{\text{new}}$ | -0.060 | 0.275 | 0.282 | -0.035 | 0.213 | 0.216 | -0.199 | 0.337 | 0.391 | -0.082 | 0.299 | 0.310 |
| $x^H_{\text{new}}$ | -0.656 | 0.475 | 0.810 | -0.272 | 0.257 | 0.374 | -0.749 | 0.503 | 0.903 | -0.214 | 0.325 | 0.389 |
| $x^S_{\text{new}}$ | -0.236 | 0.139 | 0.274 | -0.087 | 0.079 | 0.117 | -0.054 | 0.138 | 0.148 | 0.005 | 0.063 | 0.063 |
| $x^I_{\text{new}}$ | -0.173 | 0.096 | 0.197 | -0.078 | 0.077 | 0.109 | -0.049 | 0.097 | 0.120 | -0.094 | 0.057 | 0.110 |
| $x^D_{\text{new}}$ | -0.144 | 0.092 | 0.171 | -0.101 | 0.080 | 0.129 | -0.059 | 0.093 | 0.371 | -0.154 | 0.060 | 0.166 |

**Scenario III: Similar to Scenario I.** In the third scenario, we repeat the data generation process of Scenario I except for difference values for $\xi$:

- $s_\alpha = 3, \theta = 0.6 : \quad \xi = (-0.593, 0.330, 0.330, 0.005_{497\times1})^T$,
- $s_\alpha = 3, \theta = 1 : \quad \xi = (-0.801, 0.163, 0.163, 0.002_{497\times1})^T$,
- $s_\alpha = 500, \theta = 0.6 : \quad \xi = (-0.650, 0.064_{29\times1}, 0.011_{470\times1})^T$,
- $s_\alpha = 500, \theta = 1 : \quad \xi = (-0.831, 0.032_{29\times1}, 0.005_{470\times1})^T$.

We focus on the comparison between SAS estimation and inference and an unsupervised learning (UL) approach. For the UL approach, a proxy outcome $\tilde{Y}$ is derived directly from the dichotomized informative surrogate $S_1$

$$\tilde{Y} = \mathbb{I}(S_1 \geq s_*)$$

where the threshold $s_*$ is chose in order to match the prevalence $\mathbb{E}(\tilde{Y}) \approx \mathbb{E}(Y)$. Then, the UL estimation of $\beta$ is obtained by regression $\tilde{Y}$ to $X$ under the logistic regression model over all $N$ observations. Classical inference is used for construction of UL confidence intervals.

The layouts of Tables A5 and A6 are different from those of 2 and Table 3 because of the different benchmark method. The supervised learning methods SLASSO and Debiased SLASSO are not affected by the distribution of $S_i$ in Scenario I, while UL considered in Scenario III uses $S_i$ to construct its proxy outcome. UL obviously failed with ineffective classification (AUC < 0.50 in Table A4), large bias in Table A5 and severe under-covering confidence intervals in Table A6. The performance of SAS is solid as in Scenario I.
Table A3: Bias, Empirical standard error (ESE) along with empirical coverage of the 95% confidence intervals (CP) for the debiased supervised LASSO (SLASSO) and debiased SAS estimator of linear predictions $x^T_{new} \beta_0$ under Scenario II with 500 labels, moderate or large $S(\beta_o)$ and strong or moderate surrogates.

| Type                        | Bias | ESE  | ASE  | CP   | Bias | ESE  | ASE  | CP   |
|-----------------------------|------|------|------|------|------|------|------|------|
| Risk prediction model       |      |      |      |      |      |      |      |      |
| approximately sparse, moderate surrogates |      |      |      |      |      |      |      |      |
| $x^0_{new}$                 | -0.236 | 1.930 | 1.915 | 0.944 | -0.014 | 1.786 | 1.791 | 0.950 |
| $x^1_{new}$                 | -0.044 | 2.031 | 1.997 | 0.944 | -0.028 | 1.873 | 1.853 | 0.947 |
| $x^2_{new}$                 | 0.364  | 2.110 | 2.084 | 0.944 | -0.045 | 1.943 | 1.924 | 0.947 |
| $x^3_{new}$                 | 0.133  | 0.156 | 0.127 | 0.848 | -0.028 | 0.133 | 0.130 | 0.944 |
| $x^4_{new}$                 | 0.149  | 0.121 | 0.139 | 0.848 | -0.014 | 0.141 | 0.105 | 0.948 |
| Risk prediction model       |      |      |      |      |      |      |      |      |
| approximately sparse, strong surrogates |      |      |      |      |      |      |      |      |
| $x^0_{new}$                 | -0.070 | 1.953 | 1.935 | 0.947 | -0.021 | 1.371 | 1.366 | 0.949 |
| $x^1_{new}$                 | -0.031 | 2.019 | 1.986 | 0.946 | -0.026 | 1.408 | 1.401 | 0.949 |
| $x^2_{new}$                 | 0.148  | 2.073 | 2.055 | 0.948 | -0.010 | 1.458 | 1.444 | 0.949 |
| $x^3_{new}$                 | 0.029  | 0.153 | 0.127 | 0.894 | -0.016 | 0.103 | 0.096 | 0.928 |
| $x^4_{new}$                 | 0.018  | 0.134 | 0.126 | 0.938 | -0.004 | 0.081 | 0.079 | 0.944 |
| $x^5_{new}$                 | 0.134  | 0.128 | 0.141 | 0.842 | -0.007 | 0.081 | 0.083 | 0.956 |

Large $S(\beta_o)$, moderate surrogates

| Type                        | Bias | ESE  | ASE  | CP   | Bias | ESE  | ASE  | CP   |
|-----------------------------|------|------|------|------|------|------|------|------|
| Risk prediction model       |      |      |      |      |      |      |      |      |
| approximately sparse, strong surrogates |      |      |      |      |      |      |      |      |
| $x^0_{new}$                 | -0.092 | 1.942 | 1.925 | 0.950 | 0.004  | 1.796 | 1.792 | 0.951 |
| $x^1_{new}$                 | -0.034 | 1.995 | 1.969 | 0.947 | -0.018 | 1.852 | 1.835 | 0.951 |
| $x^2_{new}$                 | 0.082  | 2.061 | 2.036 | 0.946 | -0.027 | 1.912 | 1.890 | 0.948 |
| $x^3_{new}$                 | -0.009 | 0.155 | 0.125 | 0.876 | -0.027 | 0.131 | 0.125 | 0.922 |
| $x^4_{new}$                 | 0.090  | 0.126 | 0.125 | 0.952 | -0.009 | 0.104 | 0.103 | 0.950 |
| $x^5_{new}$                 | 0.119  | 0.126 | 0.139 | 0.894 | -0.012 | 0.108 | 0.108 | 0.940 |

Large $S(\beta_o)$, strong surrogates

| Type                        | Bias | ESE  | ASE  | CP   | Bias | ESE  | ASE  | CP   |
|-----------------------------|------|------|------|------|------|------|------|------|
| Risk prediction model       |      |      |      |      |      |      |      |      |
| approximately sparse, strong surrogates |      |      |      |      |      |      |      |      |
| $x^0_{new}$                 | -0.221 | 1.929 | 1.926 | 0.949 | 0.022  | 1.353 | 1.349 | 0.951 |
| $x^1_{new}$                 | 0.032  | 2.047 | 2.017 | 0.947 | -0.003 | 1.427 | 1.414 | 0.950 |
| $x^2_{new}$                 | 0.442  | 2.137 | 2.104 | 0.940 | -0.039 | 1.479 | 1.469 | 0.951 |
| $x^3_{new}$                 | 0.176  | 0.150 | 0.128 | 0.936 | -0.018 | 0.094 | 0.099 | 0.946 |
| $x^4_{new}$                 | 0.030  | 0.128 | 0.129 | 0.936 | -0.002 | 0.079 | 0.077 | 0.952 |
| $x^5_{new}$                 | 0.167  | 0.125 | 0.142 | 0.804 | -0.008 | 0.082 | 0.080 | 0.954 |

Table A4: AUC Table for simulations with 500 labels under Scenario III. The AUCs are evaluated on an independent testing set of size 100. We approximately measure the sparsity by $S(v) = \|v\|_1/\|v\|_2$.

| Scenario | Surrogate | $S(\beta_o)$ | $S(\gamma_o)$ | Prediction Accuracy (AUC) |
|----------|-----------|---------------|---------------|---------------------------|
|          |           |               |               | Oracle | UL  | SAS |
| Strong   | 174.08    | 2.31          |               | 0.72   | 0.43 | 0.69 |
| Moderate | 174.08    | 2.25          |               | 0.72   | 0.46 | 0.69 |
| Strong   | 28.27     | 2.32          |               | 0.72   | 0.45 | 0.69 |
| Moderate | 28.27     | 2.24          |               | 0.72   | 0.47 | 0.69 |
Table A5: Bias, Empirical standard error (ESE) and mean-squared error (MSE) for the unsupervised learning (UL) and SAS estimator of linear predictions $x_{\text{new}}^T \hat{\beta}_0$ with under Scenario III 500 labels, approximately sparse or dense $\beta_0$ and strong or moderate surrogates.

| Type | Bias | ESE | MSE | Bias | ESE | MSE |
|------|------|-----|-----|------|-----|-----|
| Moderate $S(\beta_0)$, moderate surrogates | | | | | | |
| $x_{\text{new}}^L$ | 0.42 | 0.92 | 1.01 | 0.02 | 0.42 | 0.42 |
| $x_{\text{new}}^M$ | 0.20 | 2.31 | 2.32 | -0.03 | 0.51 | 0.51 |
| $x_{\text{new}}^H$ | -3.39 | 3.90 | 5.16 | 0.09 | 0.59 | 0.60 |
| $x_{\text{new}}^S$ | -3.73 | 0.07 | 3.73 | -0.09 | 0.14 | 0.17 |
| $x_{\text{new}}^I$ | -0.42 | 0.14 | 0.44 | -0.19 | 0.10 | 0.22 |
| $x_{\text{new}}^D$ | -0.03 | 0.15 | 0.16 | -0.18 | 0.09 | 0.20 |
| Large $S(\beta_0)$, moderate surrogates | | | | | | |
| $x_{\text{new}}^L$ | 0.61 | 0.70 | 0.93 | 0.04 | 0.37 | 0.37 |
| $x_{\text{new}}^M$ | -0.04 | 2.54 | 2.54 | -0.03 | 0.49 | 0.49 |
| $x_{\text{new}}^H$ | -2.53 | 4.66 | 5.30 | 0.07 | 0.59 | 0.59 |
| $x_{\text{new}}^S$ | -3.56 | 0.11 | 3.56 | -0.14 | 0.14 | 0.19 |
| $x_{\text{new}}^I$ | -0.70 | 0.14 | 0.72 | -0.11 | 0.09 | 0.15 |
| $x_{\text{new}}^D$ | -0.30 | 0.13 | 0.33 | -0.12 | 0.09 | 0.15 |
| Moderate $S(\beta_0)$, strong surrogates | | | | | | |
| $x_{\text{new}}^L$ | 0.89 | 0.67 | 1.11 | -0.05 | 0.38 | 0.38 |
| $x_{\text{new}}^M$ | -0.15 | 2.51 | 2.51 | -0.04 | 0.54 | 0.54 |
| $x_{\text{new}}^H$ | -3.28 | 4.57 | 5.62 | 0.22 | 0.60 | 0.64 |
| $x_{\text{new}}^S$ | -3.83 | 0.07 | 3.83 | -0.09 | 0.13 | 0.15 |
| $x_{\text{new}}^I$ | -0.99 | 0.12 | 1.00 | -0.08 | 0.08 | 0.11 |
| $x_{\text{new}}^D$ | -0.57 | 0.12 | 0.58 | -0.10 | 0.08 | 0.13 |
| Large $S(\beta_0)$, strong surrogates | | | | | | |
| $x_{\text{new}}^L$ | 0.76 | 0.82 | 1.12 | -0.09 | 0.45 | 0.46 |
| $x_{\text{new}}^M$ | 0.09 | 2.35 | 2.35 | -0.03 | 0.57 | 0.58 |
| $x_{\text{new}}^H$ | -4.10 | 3.97 | 5.71 | 0.23 | 0.63 | 0.67 |
| $x_{\text{new}}^S$ | -3.99 | 0.09 | 3.99 | -0.05 | 0.13 | 0.14 |
| $x_{\text{new}}^I$ | -0.82 | 0.12 | 0.83 | -0.12 | 0.08 | 0.15 |
| $x_{\text{new}}^D$ | -0.37 | 0.14 | 0.40 | -0.13 | 0.08 | 0.15 |
Table A6: Bias, Empirical standard error (ESE), average of the estimated standard error (ASE) along with empirical coverage of the 95% confidence intervals (CP) for the unsupervised learning (UL) and debiased SAS estimator of linear predictions $x_{\text{new}}^T \beta_0$ under Scenario III with 500 labels, approximately sparse or dense $\beta_0$ and strong or moderate surrogates.

| Type                      | Bias     | ESE     | ASE     | CP        | Bias     | ESE     | ASE     | CP        |
|---------------------------|----------|---------|---------|-----------|----------|---------|---------|-----------|
| Moderate $S(\beta_0)$, moderate surrogates |          |         |         |           |          |         |         |           |
| $x_{\text{SS}}^T$        | 0.42     | 0.92    | 0.40    | 0.53      | -0.27    | 1.95    | 1.92    | 0.94      |
| $x_{\text{SM}}^T$        | 0.20     | 2.31    | 0.42    | 0.26      | -0.04    | 2.09    | 2.03    | 0.94      |
| $x_{\text{SH}}^T$        | -3.39    | 3.90    | 0.45    | 0.12      | 0.43     | 2.18    | 2.14    | 0.94      |
| $x_{\text{SS}}^T$        | -3.73    | 0.07    | 0.06    | 0.00      | -0.04    | 0.16    | 0.15    | 0.93      |
| $x_{\text{SM}}^T$        | -0.42    | 0.14    | 0.03    | 0.00      | -0.03    | 0.13    | 0.12    | 0.92      |
| $x_{\text{SH}}^T$        | -0.03    | 0.15    | 0.03    | 0.28      | -0.02    | 0.12    | 0.12    | 0.95      |
| Moderate $S(\beta_0)$, strong surrogates |          |         |         |           |          |         |         |           |
| $x_{\text{SS}}^T$        | 0.61     | 0.70    | 0.41    | 0.56      | -0.22    | 1.98    | 1.96    | 0.95      |
| $x_{\text{SM}}^T$        | -0.04    | 2.54    | 0.43    | 0.29      | -0.05    | 2.06    | 2.02    | 0.95      |
| $x_{\text{SH}}^T$        | -2.53    | 4.66    | 0.45    | 0.11      | 0.47     | 2.16    | 2.10    | 0.94      |
| $x_{\text{SS}}^T$        | -3.56    | 0.11    | 0.06    | 0.00      | -0.03    | 0.16    | 0.15    | 0.92      |
| $x_{\text{SM}}^T$        | -0.70    | 0.14    | 0.03    | 0.00      | -0.02    | 0.13    | 0.12    | 0.91      |
| $x_{\text{SH}}^T$        | -0.30    | 0.13    | 0.03    | 0.02      | -0.02    | 0.13    | 0.12    | 0.93      |
| Moderate $S(\beta_0)$, strong surrogates |          |         |         |           |          |         |         |           |
| $x_{\text{SS}}^T$        | 0.89     | 0.67    | 0.41    | 0.37      | -0.22    | 1.69    | 1.67    | 0.95      |
| $x_{\text{SM}}^T$        | -0.15    | 2.51    | 0.42    | 0.27      | -0.04    | 1.78    | 1.71    | 0.94      |
| $x_{\text{SH}}^T$        | -3.28    | 4.57    | 0.44    | 0.10      | 0.48     | 1.86    | 1.79    | 0.93      |
| $x_{\text{SS}}^T$        | -3.83    | 0.07    | 0.07    | 0.00      | -0.02    | 0.14    | 0.13    | 0.93      |
| $x_{\text{SM}}^T$        | -0.99    | 0.12    | 0.03    | 0.00      | -0.00    | 0.11    | 0.10    | 0.92      |
| $x_{\text{SH}}^T$        | -0.57    | 0.12    | 0.03    | 0.00      | -0.01    | 0.11    | 0.10    | 0.93      |
| Large $S(\beta_0)$, strong surrogates |          |         |         |           |          |         |         |           |
| $x_{\text{SS}}^T$        | 0.76     | 0.82    | 0.40    | 0.41      | -0.27    | 1.67    | 1.65    | 0.95      |
| $x_{\text{SM}}^T$        | 0.09     | 2.35    | 0.42    | 0.26      | -0.03    | 1.81    | 1.73    | 0.94      |
| $x_{\text{SH}}^T$        | -4.10    | 3.97    | 0.45    | 0.10      | 0.45     | 1.88    | 1.83    | 0.94      |
| $x_{\text{SS}}^T$        | -3.99    | 0.09    | 0.07    | 0.00      | -0.03    | 0.14    | 0.14    | 0.93      |
| $x_{\text{SM}}^T$        | -0.82    | 0.12    | 0.03    | 0.00      | -0.01    | 0.11    | 0.10    | 0.93      |
| $x_{\text{SH}}^T$        | -0.37    | 0.14    | 0.03    | 0.01      | -0.01    | 0.11    | 0.11    | 0.97      |
Appendix B. Proofs of Main Results

We first summarize below notations used Section 3 for the conditional expectations given different part of the data.

**Definition 11** The conditional expectation for samples with index in set $S$ conditionally on subset of the data $\mathcal{D}$ is denoted as

$$E_{i \in S}\{f(Y_i, X_i, S_i) \mid \mathcal{D}\}, \quad S \subseteq \{1, \ldots, n + N\}, \mathcal{D} \subset \mathcal{L} \cup \mathcal{U}.$$  

We denote the conditional expectation of unlabeled data given labelled data by $E_{i>n}\{f(W_i) \mid \mathcal{L}\}$ and the conditional probability of new copy of data given current data by $P^\text{new}_i\{f(W_i) \mid \mathcal{D}\}$. With $\mathcal{L}$ and $\mathcal{U}$ partitioned into $K$ folds indexed respectively by $\{I_k, k = 1, \ldots, K\}$ and $\{J_k, k = 1, \ldots, K\}$, we denote the conditional expectation of fold-$k$ labelled data and unlabeled data given the out-of-fold data respectively by

$$E_{i \in I_k}\{f(Y_i, X_i, S_i) \mid \mathcal{D}^c_k\} \quad \text{and} \quad E_{i \in J_k}\{f(W_i) \mid \mathcal{D}^c_k\},$$

where $\mathcal{D}^c_k = \{S_i, X_i, i \in J_k\} \cup \{Y_i, S_i, X_i, i \in I_k\}$.

**B1 Proof of Theorem 1**

Our proof shares the general steps with the restricted strong convexity framework laid down in Negahban et al. (2010) while we have a delicate analysis of the symmetrized Bregman divergence to establish the improved rate of estimation under semi-supervised learning setting. To bound $\hat{\beta}$ through the symmetrized Bregman divergence $(\hat{\beta} - \beta_0)^\dagger \ell^\dagger (\beta_0; \tilde{\gamma})$, instead of directly applying the Hölder’s bound, we first split it into two parts,

$$\begin{align*}
(\hat{\beta} - \beta_0)^\dagger \ell^\dagger (\beta_0; \tilde{\gamma}) &= (\hat{\beta} - \beta_0)^\dagger \left[ \ell^\dagger (\beta_0; \tilde{\gamma}) - E\{\ell^\dagger (\beta_0; \tilde{\gamma}) \mid \mathcal{L}\} + E\{\ell^\dagger (\beta_0; \gamma_0) \mid \mathcal{L}\} \right] \\
&+ (\hat{\beta} - \beta_0)^\dagger E\left\{ \ell^\dagger (\beta_0; \tilde{\gamma}) - \ell^\dagger (\beta_0; \gamma_0) \mid \mathcal{L}\right\} 
\end{align*}$$

(A.1)

and discuss which part dominates the estimation error. When the first variance term in (A.1) is dominant, the bias from $\tilde{\gamma}$ becomes eligible. Then, we should recover the usual error bound for LASSO as if $\gamma_0$ is used. When the second bias term in (A.1) is dominant, the error bound of $\hat{\beta}$ can be controlled by the error bound of $\tilde{\gamma}$. Combining the error bounds in the two cases, we obtain the oracle inequalities.

**Lemma 12** On event

$$\Omega = \left\{ \ell_{pl}(\beta_0 + \Delta) - \ell_{pl}(\beta_0) - \Delta^\dagger \ell_{pl}^{}(\beta_0) \right\},$$

$$\geq \kappa_{nc,1}\|\Delta\|_2\{\|\Delta\|_2 - \kappa_{nc,2}\sqrt{\log(p)/N}\|\Delta\|_1\}, \forall \|\Delta\|_2 \leq 1,$$

setting $\lambda_\beta \geq \sqrt{\log(p)/N}$ such that

$$\lambda_\beta \geq 3 \|\ell^\dagger (\beta_0; \tilde{\gamma}) - E\{\ell^\dagger (\beta_0; \tilde{\gamma}) \mid \mathcal{L}\} + E\{\ell^\dagger (\beta_0; \gamma_0) \mid \mathcal{L}\} \|_\infty + \kappa_{nc,1}\kappa_{nc,2}\sqrt{\frac{\log(p)}{N}},$$

28
we have the oracle inequalities for estimation error of $\hat{\beta}$,

$$
\|\hat{\beta} - \beta\|_2 \leq \max \left\{ 14\sqrt{\beta_0} \lambda_\beta / \kappa_{nc,1}, (1 - \rho) 7M \sigma_{\max}^2 \| \gamma_0 - \hat{\gamma} \|_2 / \kappa_{nc,1} \right\},
$$

$$
\|\hat{\beta} - \beta\|_1 \leq \max \left\{ 84\sqrt{\beta_0} \lambda_\beta / \kappa_{nc,1}, (1 - \rho)^2 21M^2 \sigma_{\max}^4 \| \gamma_0 - \hat{\gamma} \|_2^2 / (\kappa_{nc,1} \lambda_\beta) \right\}.
$$

The constants $\kappa_{nc,1}, \kappa_{nc,2}$ are the restrictive strong convexity parameters specified in Lemma 19.

We next prove the oracle inequalities. First, we note that by the definition of $\hat{\beta}$,

$$
\ell^t(\hat{\beta}; \hat{\gamma}) + \lambda_\beta \|\hat{\beta}\|_1 \leq \ell^t(\beta_0; \hat{\gamma}) + \lambda_\beta \|\beta_0\|_1. \quad (A.2)
$$

Denote the standardized estimation error as $\delta = (\hat{\beta} - \beta_0) / \|\beta - \beta_0\|_2$. Due to convexity of the loss function, we have for $t = \|\beta - \beta_0\|_2 \wedge 1$

$$
\ell^t(\beta_0 + t\delta; \hat{\gamma}) + \lambda_\beta \|\beta_0 + t\delta\|_1 \leq \ell^t(\beta_0; \hat{\gamma}) + \lambda_\beta \|\beta_0\|_1. \quad (A.3)
$$

By the triangle inequality $\|\beta_0\|_1 - \|\beta_0 + t\delta\|_1 \leq t\|\delta\|_1$, we have from (A.3)

$$
\ell^t(\beta_0 + t\delta; \hat{\gamma}) - \ell^t(\beta_0; \hat{\gamma}) \leq t\lambda_\beta \|\delta\|_1 \quad (A.4)
$$

To apply the restricted strong convexity of the complete data loss (11) established in Lemma 19, we show that the second order approximation error of the imputed loss is equivalent to that of the complete data loss,

$$
\ell^t(\beta_0 + t\delta; \hat{\gamma}) - \ell^t(\beta_0; \hat{\gamma}) - t\delta^T \hat{\ell}^t(\beta_0; \hat{\gamma}) = \ell_{pl}(\beta_0 + t\delta) - \ell_{pl}(\beta_0) - t\delta^T \hat{\ell}_{pl}(\beta_0).
$$

Then by applying the restricted strong convexity event $\Omega$, we obtain

$$
\ell^t(\beta_0 + t\delta; \hat{\gamma}) - \ell^t(\beta_0; \hat{\gamma}) - t\delta^T \hat{\ell}^t(\beta_0; \hat{\gamma}) \geq t^2 \kappa_{nc,1} - t\kappa_{nc,1} \kappa_{nc,2} \sqrt{\log(p)/N} \|\delta\|_1. \quad (A.5)
$$

Applying (A.5) to (A.4), we have with large probability

$$
t\delta^T \hat{\ell}^t(\beta_0; \hat{\gamma}) + t^2 \kappa_{nc,1} - t\kappa_{nc,1} \kappa_{nc,2} \sqrt{\log(p)/N} \|\delta\|_1 \leq t\lambda_\beta \|\delta\|_1
$$

where $\|\delta\|_2 = 1$ from definition. Thus, we have reach

$$
t\kappa_{nc,1} \leq \lambda_\beta \|\delta\|_1 - \delta^T \hat{\ell}^t(\beta_0; \hat{\gamma}) + \kappa_{nc,1} \kappa_{nc,2} \sqrt{\log(p)/N} \|\delta\|_1. \quad (A.6)
$$

Next, we analyze $\delta^T \hat{\ell}^t(\beta_0; \hat{\gamma})$ by the decomposition

$$
\left|\delta^T \hat{\ell}^t(\beta_0; \hat{\gamma})\right| = \delta^T \left[ \hat{\ell}^t(\beta_0; \hat{\gamma}) - \mathbb{E} \left\{ \hat{\ell}^t(\beta_0; \hat{\gamma}) \mid \mathcal{L} \right\} + \mathbb{E} \left\{ \hat{\ell}^t(\beta_0; \gamma_0) \mid \mathcal{L} \right\} \right] + \delta^T \mathbb{E} \left\{ \hat{\ell}^t(\beta_0; \hat{\gamma}) - \hat{\ell}^t(\beta_0; \gamma_0) \mid \mathcal{L} \right\}
$$

$$
\leq \|\delta\|_1 \left\| \hat{\ell}^t(\beta_0; \hat{\gamma}) - \mathbb{E} \left\{ \hat{\ell}^t(\beta_0; \hat{\gamma}) \mid \mathcal{L} \right\} + \mathbb{E} \left\{ \hat{\ell}^t(\beta_0; \gamma_0) \mid \mathcal{L} \right\} \right\|_\infty
$$

$$
+ (1 - \rho) \left\| \mathbb{E}_{i > n} \left\{ g(\beta_0^T X_i) - g(\hat{\gamma}^T W_i) \right\} \mid \mathcal{L} \right\|_2. \quad (A.7)
$$
To establish the rate for $L_2$-norm of $\mathbb{E}\{\hat{\ell}^\dagger(\beta_0; \hat{\gamma}) \mid \mathcal{L}\}$, we note that
\[
\mathbb{E}_{i > n}[X_i \{g(\beta_0^i X_i) - g(\hat{\gamma}_i^T W_i)\} \mid \mathcal{L}] = \mathbb{E}_{i > n}[X_i \{Y_i - g(\hat{\gamma}_i^T W_i)\} \mid \mathcal{L}].
\tag{A.8}
\]
By the characterization of $\gamma_0$ as in (6), we may rewrite (A.8) as
\[
\mathbb{E}_{i > n}[X_i \{Y_i - g(\hat{\gamma}_i^T W_i)\} \mid \mathcal{L}] = \mathbb{E}_{i > n}[g'(\gamma_0^i W_i)X_i W_i^\dagger \mid \mathcal{L}] (\gamma_0 - \hat{\gamma}),
\tag{A.9}
\]
where $\gamma_u = u\hat{\gamma} + (1 - u)\gamma_0$ for some $u \in [0, 1]$. Under Assumptions 1b and 2a, as well as the fact that $X_i$ is a sub-vector of $W_i$, we have
\[
\left\| \mathbb{E}\{\hat{\ell}^\dagger(\beta_0; \hat{\gamma}) \mid \mathcal{L}\} \right\|_2 \leq \left\| \mathbb{E}_{i \in \mathcal{U}}[g'(\gamma_0^i W_i)W_i^{\otimes 2} \mid \mathcal{L}] \right\|_2 \|\gamma_0 - \hat{\gamma}\|_2 
\leq M \left\| \mathbb{E}(W_i^{\otimes 2}) \right\|_2 \|\gamma_0 - \hat{\gamma}\|_2 \leq M \sigma^2_{\max} \|\gamma_0 - \hat{\gamma}\|_2,
\tag{A.10}
\]
where for any vector $x$, $x^{\otimes 2} = xx^\dagger$. By the bound for (A.10) and the definition of $\lambda_\beta$, we have the bound from (A.6)
\[
t_{\kappa_{sc,1}} \leq 2\lambda_\beta \|\delta\|_1 + (1 - \rho) M \sigma^2_{\max} \|\gamma_0 - \hat{\gamma}\|_2.
\tag{A.11}
\]
Hence, we can reach an immediate bound for estimation error from (A.11) without considering the sparsity of $\beta_0$. We shall proceed to derive a sharper bound that involves the sparsity of $\beta_0$. We separately analyze two cases.

**Case 1:** $$(1 - \rho) M \sigma^2_{\max} \|\gamma_0 - \hat{\gamma}\|_2 \geq \|\delta\|_1 \lambda_\beta / 3.$$ 

In this case, the estimation error is dominated by $\hat{\gamma} - \gamma_0$. We simply have from (A.11)
\[
t_{\kappa_{sc,1}} \leq 7(1 - \rho) M \sigma^2_{\max} \|\gamma_0 - \hat{\gamma}\|_2,
\]
\[
t_{\kappa_{sc,1}} \|\delta\|_1 \lambda_\beta / 3 \leq 7(1 - \rho) M^2 \sigma^4_{\max} \|\gamma_0 - \hat{\gamma}\|_2^2.
\]
Thus, we have
\[
\|\hat{\beta} - \beta_0\|_2 \leq (1 - \rho) 7M \sigma^2_{\max} \|\gamma_0 - \hat{\gamma}\|_2 / \kappa_{sc,1},
\]
\[
\|\hat{\beta} - \beta_0\|_1 \leq (1 - \rho)^2 21M^2 \sigma^4_{\max} \|\gamma_0 - \hat{\gamma}\|_2^2 / (\kappa_{sc,1} \lambda_\beta). 
\tag{A.12}
\]
If case 1 does not hold, then instead

**Case 2:** $$(1 - \rho) M \sigma^2_{\max} \|\gamma_0 - \hat{\gamma}\|_2 \leq \|\delta\|_1 \lambda_\beta / 3.$$ 

In this case, the estimation error is comparable to that when we have the true $\gamma_0$ for the imputation. Thus, the sparsity of $\beta_0$ may affect the estimation error.

Following the typical approach to establish the cone condition for $\delta$, we analyze the symmetrized Bregman’s divergence,
\[
(\hat{\beta} - \beta_0)^\dagger \left\{ \hat{\ell}^\dagger(\beta; \gamma) - \ell^\dagger(\beta_0; \gamma) \right\} = \|\hat{\beta} - \beta_0\|_2 \delta^\dagger \left\{ \hat{\ell}^\dagger(\beta; \gamma) - \ell^\dagger(\beta_0; \gamma) \right\}.
\tag{A.14}
\]
Due to the convexity of the loss \( \ell^\dagger(\cdot; \gamma) \) under Assumption 1b, the symmetrized Bregman’s divergence (A.14) is nonnegative through a mean-value theorem,

\[
(\hat{\beta} - \beta_o)^T \left\{ \ell^\dagger(\hat{\beta}; \gamma) - \ell^\dagger(\beta_o; \gamma) \right\} \geq \inf_{\beta \in \mathbb{R}^{p+1}} \frac{1}{N} \sum_{i > n} g'((\hat{\beta} - \beta_o)^T X_i)((\hat{\beta} - \beta_o)^T X_i)^2 \geq 0.
\]

Denote the indices set of nonzero coefficient in \( \beta_o \) as \( O_\beta = \{ j : \beta_{0,j} \neq 0 \} \). We denote \( \delta_{O_\beta} \) and \( \delta_{\bar{O}_\beta} \) as the sub-vectors for \( \delta \) at positions in \( O_\beta \) and at positions not in \( O_\beta \), respectively. The solution \( \hat{\beta} \) satisfies the KKT condition

\[
\| \ell^\dagger(\hat{\beta}; \gamma) \|_\infty \leq \lambda_\beta, \quad \ell^\dagger(\hat{\beta}; \gamma)_j = -\lambda_\beta \text{sign}(\hat{\beta}_j), \ j : \hat{\beta}_j \neq 0.
\]

From the KKT condition and the definitions of \( \delta \) and \( O_\alpha \), we have

\[
\delta_j \ell^\dagger(\hat{\beta}; \gamma)_j \leq |\delta_j| \lambda_\beta, \ j \in O_\beta; \quad \delta_j \ell^\dagger(\hat{\beta}; \gamma)_j = \frac{-\hat{\beta}_j \lambda_\beta \text{sign}(\hat{\beta}_j)}{\| \beta - \beta_o \|_2} = -\lambda_\beta |\delta_j|, \ j \in O_\beta. \quad (A.15)
\]

Applying (A.15) to (A.14), we have the upper bound,

\[
\delta^T \left\{ \ell^\dagger(\beta_o; \gamma) - \ell^\dagger(\beta_o; \gamma) \right\} = \sum_{j \in O_\beta} \delta_j \ell^\dagger(\hat{\beta}; \gamma)_j + \sum_{j \in \bar{O}_\beta} \delta_j \ell^\dagger(\hat{\beta}; \gamma)_j + \delta^T \ell^\dagger(\beta_o; \gamma)
\]

\[
\leq \lambda_\beta \sum_{j \in O_\beta} |\delta_j| - \lambda_\beta \sum_{j \in \bar{O}_\beta} |\delta_j| + \delta^T \ell^\dagger(\beta_o; \gamma)
\]

\[
\leq \lambda_\beta \| \delta_{O_\beta} \|_1 - \lambda_\beta \| \delta_{\bar{O}_\beta} \|_1 + \| \delta^T \ell^\dagger(\beta_o; \gamma) \|.
\]

Then, we apply (A.10), the definition of \( \lambda_\beta \) and (A.13),

\[
0 \leq \lambda_\beta \| \delta_{O_\beta} \|_1 - \lambda_\beta \| \delta_{\bar{O}_\beta} \|_1 + \frac{2}{3} \lambda_\beta \| \delta \|_1 \quad \text{and} \quad \lambda_\beta \| \delta_{O_\beta} \|_1 \leq 5 \lambda_\beta \| \delta_{O_\beta} \|_1.
\]

Therefore, we can bound the \( L_1 \) norm of \( \delta \) by the cone property,

\[
\| \delta \|_1 \leq 6 \lambda_\beta \| \delta_{O_\beta} \|_1 \leq 6 \sqrt{s_\beta} \| \delta \|_2 = 6 \sqrt{s_\beta}. \quad (A.16)
\]

We then apply the cone condition (A.16) and the case condition (A.13) to the bound (A.11),

\[
t \kappa_{\text{res},1} \leq 14 \sqrt{s_\beta} \lambda_\beta, \quad \text{and} \quad t \kappa_{\text{res},1} \| \delta \|_1 \leq 84 s_\beta \lambda_\beta
\]

Thus, we obtain the rate for estimation error

\[
\| \hat{\beta} - \beta_o \|_2 \leq 14 \sqrt{s_\beta} \lambda_\beta / \kappa_{\text{res},1}, \quad \text{and} \quad \| \hat{\beta} - \beta_o \|_1 \leq 84 s_\beta \lambda_\beta / \kappa_{\text{res},1}. \quad (A.17)
\]

Since Case 1 and Case 2 are the complement of each other, one of them must occur. Thus, the bound of estimation error is controlled by the larger bound in the two cases,

\[
\| \hat{\beta} - \beta_o \|_2 \leq \max \left\{ 14 \sqrt{s_\beta} \lambda_\beta / \kappa_{\text{res},1}, (1 - \rho) 7 M^2 \sigma^2_{\max} \| \gamma_0 - \hat{\gamma} \|_2 / \kappa_{\text{res},1} \right\},
\]

\[
\| \hat{\beta} - \beta_o \|_1 \leq \max \left\{ 84 s_\beta \lambda_\beta / \kappa_{\text{res},1}, (1 - \rho)^2 21 M^2 \sigma^4_{\max} \| \gamma_0 - \hat{\gamma} \|_2 / (\kappa_{\text{res},1} \lambda_\beta) \right\},
\]

which is our oracle inequality in Lemma 12.
Consistency  We next show that the oracle inequality leads to the consistency under dimension condition (26). To show
\[
\left\| \hat{\ell}^t(\beta_0; \hat{\gamma}) - \mathbb{E}\{\hat{\ell}^t(\beta_0; \hat{\gamma}) \mid \mathcal{L}\} + \mathbb{E}\{\ell^t(\beta_0; \gamma_0) \mid \mathcal{L}\} \right\| \leq \left\| \hat{\ell}^t(\beta_0; \hat{\gamma}) - \mathbb{E}\{\hat{\ell}^t(\beta_0; \hat{\gamma}) \mid \mathcal{L}\} \right\| + \left\| \mathbb{E}\{\ell^t(\beta_0; \gamma_0) \mid \mathcal{L}\} \right\| = O_p\left(\sqrt{\log(p)/N}\right),
\]
we express the term of interest as the sum of the following empirical processes
\[
\ell^t(\beta_0; \hat{\gamma}) - \mathbb{E}\{\ell^t(\beta_0; \hat{\gamma}) \mid \mathcal{L}\} = \frac{1}{N} \sum_{i>n} \left( X_i \{ g(\beta_0^T X_i) - Y_i + Y_i - g(\hat{\gamma}^TW_i) \} - \mathbb{E}_{i>n} \left[ X_i \{ g(\beta_0^T X_i) - Y_i + Y_i - g(\hat{\gamma}^TW_i) \} \mid \mathcal{L} \right] \right),
\]
\[
\mathbb{E}\{\ell^t(\beta_0; \gamma_0) \mid \mathcal{L}\} = \frac{1}{N} \sum_{i=1}^n X_i \{ g(\beta_0^T X_i) - Y_i \}.
\]
Under Assumption 1a and 1a, \(X_i\) and \(g(\beta_0^T X_i) - Y_i\) are sub-Gaussian. According to Lemma 21, the event \(\| \hat{\gamma} - \gamma_0 \|_2 \leq 1\) occurs with large probability, on which we have a bound for the sub-Gaussian norm of \(Y_i - g(\hat{\gamma}^TW_i)\) by Lemma 14.
\[
\| Y_i - g(\hat{\gamma}^TW_i) \|_{\psi_2} \leq \max\{2\nu_1, M\sqrt{2}\sigma_{\max}\}, \quad i > n \quad (A.18)
\]
Thus, we obtain from (A.18) that \(Y_i - g(\hat{\gamma}^TW_i)\) is sub-Gaussian with large probability. Thus by the properties of sub-Gaussian random variables in Lemma 17-d and 17-f, we have established that the elements in the summands of \(\ell^t(\beta_0; \hat{\gamma})\) are all sub-exponential random variables conditionally on the labelled data. We apply the Bernstein’s inequality (Lemma 17-h) conditionally on the labelled data to obtain
\[
\left\| \hat{\ell}^t(\beta_0; \hat{\gamma}) - \mathbb{E}\{\hat{\ell}^t(\beta_0; \hat{\gamma}) \mid \mathcal{L}\} \right\| = O_p\left(\sqrt{(1 - \rho)\log(p)/N}\right),
\]
\[
\left\| \mathbb{E}\{\ell^t(\beta_0; \gamma_0) \mid \mathcal{L}\} \right\| = O_p\left(\rho\log(p)/N\right).
\]
This establishes the order for \(\lambda_\beta\),
\[
\lambda_\beta \gtrsim \sqrt{(1 - \rho)\log(p)/N + \rho\log(p)/N \times \sqrt{\log(p)/N}}. \quad (A.19)
\]
By Lemma 19 from Negahban et al. (2010), we have that the probability of restricted strong convexity event converges to one,
\[
\mathbb{P}(\Omega) \geq 1 - \kappa_{\text{rc}} 3\kappa_{\text{rc}}^4 N \to 1.
\]
Setting \(\lambda_\beta \sim \sqrt{\log(p)/N}\) for optimal \(L_2\) estimation, we achieve the stated conclusion
\[
\| \hat{\beta} - \beta_0 \|_2 = O_p\left(\sqrt{s_\beta \log(p)/N} + (1 - \rho)\sqrt{s_r \log(p + q)/n}\right),
\]
\[
\sqrt{\log(p)/N} \| \hat{\beta} - \beta_0 \|_1 = O_p\left(\sqrt{s_r \log(p)/N} + (1 - \rho)^2 s_r \frac{\log(p + q)}{n}\right),
\]
by applying the rates from Lemma 21 and (A.19). For optimal \(L_1\) estimation, we set a larger penalty \(\lambda_\beta' \sim \sqrt{\log(p)/N} \vee \sqrt{s_r \log(p + q)/(s_\beta n)} \gtrsim \lambda_\beta\) to achieve
\[
\| \hat{\beta} - \beta_0 \|_1 = O_p\left(\sqrt{s_\beta \log(p)/N} + (1 - \rho)^2 \frac{s_r s_\beta \log(p + q)}{n}\right).
\]
B2 Proof of Corollary 5

Under Assumption 1b, we have
\[
|g(\hat{\beta}^T x_{new}) - g(\beta_0^T x_{new})| \leq M|((\hat{\beta} - \beta_0)^T x_{new})|.
\] (A.20)

Since \(x_{new}\) satisfies Assumption 2a, we have
\[
\|((\hat{\beta} - \beta_0)^T x_{new})\|_{\psi_2} \leq \|\hat{\beta} - \beta_0\|_{2\sigma_{max}}/\sqrt{2}.
\]

The tail distribution is regulated by the sub-Gaussian norm by Lemma 17-a,
\[
\mathbb{P}_{new}\left(|(\hat{\beta} - \beta_0)^T x_{new}| \geq t \mid \mathcal{F}\right) \leq 2\exp\left(-t\sqrt{2}/\left\{\|\hat{\beta} - \beta_0\|_{2\sigma_{max}}\right\}\right).
\] (A.21)

Combining (A.20) and (A.21), we obtain
\[
\mathbb{P}_{new}\left(|g(\hat{\beta}^T x_{new}) - g(\beta_0^T x_{new})| \geq tM|\hat{\beta} - \beta_0\|_{2\sigma_{max}}/\sqrt{2} \mid \mathcal{F}\right) \leq 2e^{-t/2}.
\]

Thus,
\[
|g(\hat{\beta}^T x_{new}) - g(\beta_0^T x_{new})| = O_p\left(\|\hat{\beta} - \beta_0\|_2\right).
\]

B3 Proof of Theorem 7

Our proof is organized in five parts. In Part 1, we establish the consistency of the cross-fitting estimator for precision matrix, namely \(\|\hat{u}^{(k)} - u_0\|_2 = o_p(1)\) with \(\hat{u}^{(k)}\) and \(u_0\) defined in (20) and (19), respectively. In Part 2, we show that the debiased estimator can be approximated by the empirical process
\[
\sqrt{n} \left(x_{std}^T \hat{\beta} - x_{std}^T \beta_0\right) = - (1 - \rho)\sqrt{n} u_0^T \ell_{emp}(\gamma_0) - \sqrt{n} u_0^T \ell\hat{\beta}(\beta_0; \gamma_0) + o_p(1)
\]
\[
= - n^{-1/2} \left[ \sum_{i=1}^{n} u_0^T X_i \{(1 - \rho) \cdot g(\gamma_0^T W_i) + \rho \cdot g(\beta_0^T X_i) - Y_i\} + \rho \sum_{i>n} X_i \{g(\beta_0^T X_i) - g(\gamma_0^T W_i)\}\right] + o_p(1)
\]

As long as the asymptotic variance \(V_{SAS}\) defined in (29) is bounded and bounded away from zero, we have the asymptotic normality of the leading term from the Central Limit Theorem
\[
-n^{-1/2} V_{SAS}^{-1/2} \left[ \sum_{i=1}^{n} u_0^T X_i \{(1 - \rho) \cdot g(\gamma_0^T W_i) + \rho \cdot g(\beta_0^T X_i) - Y_i\} + \rho \sum_{i>n} X_i \{g(\beta_0^T X_i) - g(\gamma_0^T W_i)\}\right] \sim N(0, 1).
\]

In Part 3, we deal with the asymptotic variance \(V_{SAS}\) and the consistency of the variance estimator \(\hat{V}_{SAS}\) defined in (23). In Part 4, we reach the conclusion of the theorem based on,
\[
(1 - \rho) \|\hat{\gamma}^{(k)} - \gamma_0\|_2 + \|\hat{u}^{(k)} - u_0\|_2
\]
\[
+ \sqrt{n} \|\hat{\beta}^{(k)} - \beta_0\|_2 \left(\|\hat{\beta}^{(k)} - \beta_0\|_2 + \|\hat{u}^{(k)} - u_0\|_2\right) = o_p(1),
\] (A.22)

for all \(1 \leq k \leq K\). Following Part 4, we show in Part 5 that (28) implies (A.22).
PART 1: CONSISTENCY OF ESTIMATED PRECISION MATRIX

The definitions of \( \mathbf{u}_0 \) and \( \hat{\mathbf{u}}^{(k)} \) are given in (19) and (20). In this part, we show

\[
\|\hat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 = O_p \left( \sqrt{(s_u + s_\beta) \log(p)/(N - N_k)} + (1 - \rho) \sqrt{s_u \log(p + q)/(n - n_k)} \right) = O_p \left( \sqrt{(s_u + s_\beta)/N + (1 - \rho) \sqrt{s_u \log(p + q)/n}} \right).
\]

Since we set the number of folds \( K \leq 10 \) to be finite, the estimation rate applies for \( \hat{\mathbf{u}}^{(k)} \) for all \( k = 1, \ldots, K \).

We denote the components in the quadratic loss function of (20) and their derivatives as

\[
m^{(k,k')}(\mathbf{u}; \beta) = \frac{1}{N_k} \sum_{i \in \mathcal{I}_k \cup \mathcal{J}_{k'}} \frac{1}{2} g'((\beta^\top \mathbf{X}_i)(\mathbf{X}_i^\top \mathbf{u})^2 - \mathbf{u}^\top \mathbf{x}_{\text{std}},
\]

\[
\hat{m}^{(k,k')}(\mathbf{u}; \beta) = \frac{\partial}{\partial \mathbf{u}} m^{(k,k')}(\mathbf{u}; \beta), \quad \hat{m}^{(k,k')}(\beta) = \frac{\partial}{\partial \beta} \hat{m}^{(k,k')}(\mathbf{u}; \beta)
\]

for \( k' \in \{1, \ldots, K\} \setminus \{k\} \). We may express (20) as

\[
\hat{\mathbf{u}}^{(k)} = \underset{\mathbf{u} \in \mathbb{R}^p}{\arg\min} \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} m^{(k,k')}(\mathbf{u}; \hat{\beta}^{(k,k')}) + \lambda_u \| \mathbf{u} \|_1,
\]

Similar to the proof of Theorem 1, we establish the estimation rate for \( \hat{\mathbf{u}} \) through an oracle inequality.

**Lemma 13** Under Assumptions 1, 2, we establish On event

\[
\Omega^{(k)} = \bigcap_{k' \neq k} \left\{ \Delta^\top \hat{\mathbf{m}}^{(k,k')} \left( \hat{\beta}^{(k,k')} \right) \Delta \geq \kappa_{\text{nc},1} \| \Delta \|_2 \left\| \| \Delta \|_2 - \kappa_{\text{nc},2} \sqrt{\log(p)/N_{k'}} \| \Delta \|_1 \right\}, \forall \| \Delta \|_2 \leq 1 \right\},
\]

setting \( \lambda_u \propto \sqrt{\log(p)/N} \) such that

\[
\lambda_u \geq 3 \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} \left\{ \| \hat{\mathbf{m}}^{(k,k')} \left( \mathbf{u}_0; \hat{\beta}^{(k,k')} \right) - \mathbb{E} \left\{ \hat{\mathbf{m}}^{(k,k')} \left( \mathbf{u}_0; \hat{\beta}^{(k,k')} \right) | \mathcal{F}_{k'} \right\} \|_\infty 
\]

\[
+ \kappa_{\text{nc},1}^* \kappa_{\text{nc},2}^* \sqrt{\log(p)/N_{k'}} \right\},
\]

we have the oracle inequality for estimation error of \( \hat{\beta} \),

\[
\|\hat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_2 \leq \max \left\{ 14 \sqrt{\kappa_u/\kappa_{\text{nc},1}^*}, 7 M \sigma^3 \sup_{k' \neq k} \| \beta_0 - \hat{\beta}^{(k,k')} \|_2 / \kappa_{\text{nc},1}^* \right\},
\]

\[
\|\hat{\mathbf{u}}^{(k)} - \mathbf{u}_0\|_1 \leq \max \left\{ 84 s_u \lambda_u / \kappa_{\text{nc},1}^*, 21 M \sigma^6 \sup_{k' \neq k} \| \beta_0 - \hat{\beta}^{(k,k')} \|_2^2 / (\kappa_{\text{nc},1}^* \lambda_u) \right\}.
\]

The constants \( \kappa_{\text{nc},1}^*, \kappa_{\text{nc},2}^* \) are the restrictive strong convexity parameters specified in Lemma 20.
The proof of Lemma 13 repeats the proof of the oracle inequality for Theorem 1, so we put the detail to Section C.

To use Lemma 13 for the estimation rate of \( \mathbf{u} \), we only need to verify two conditions. First, the event \( \Omega^{(k)} \) occurs with probability tending to one. Second, the oracle choice of \( \lambda_a \) is of order \( \sqrt{\log(p)/N} \).

Repeating Theorem 1 for each \( \beta^{(k,k')} \), we have under (28)

\[
\left\| \beta^{(k,k')} - \beta_0 \right\|_2 = o_p(1).
\]

Then by Lemma 20, the sets whose intersection forms \( \Omega^{(k)} \) each occurs with probability tending to one. Since we set the number of fold finite \( K \leq 10 \), we can take union bound to obtain that \( \Omega^{(k)} \) occurs with probability tending to one.

We may write

\[
\tilde{m}^{(k,k')}(\mathbf{u}_0; \hat{\beta}^{(k,k')}) - \mathbb{E}\left\{ \tilde{m}^{(k)}(\mathbf{u}_0; \hat{\beta}^{(k,k')}) \mid \mathcal{D}_k \right\} = \frac{1}{N_{k'}} \sum_{i \in \mathcal{I}_{k'} \cup \mathcal{J}_{k'}} g'(\hat{\beta}^{(k,k')})^\top X_i X_i^\top \mathbf{u}_0 - \mathbb{E}_{i \in \mathcal{I}_{k'} \cup \mathcal{J}_{k'}} \left\{ g'(\hat{\beta}^{(k,k')})^\top X_i X_i^\top \mathbf{u}_0 \mid \mathcal{D}_k \right\}.
\]

Each element in (A.24) is an empirical process. Under Assumptions 1b and 2a, we can show that each summand is a sub-exponential random variable by Lemma 17-e, 17-f.

Hence, we can apply the Bernstein’s inequality to show that

\[
\left\| g'(\hat{\beta}^{(k,k')})^\top X_i X_i^\top \mathbf{u}_0 \right\|_{\psi_1} \leq M \left\| X_i, j \right\|_{\psi_2} \left\| X_i^\top \mathbf{u}_0 \right\|_{\psi_2} \leq M \sigma_{\max} \left\| \mathbf{u}_0 \right\|_2 / 2.
\]

Using the fact that \( N_{k'} \sim N \), we obtain that the oracle \( \lambda_a \) is of order \( O_p\left(\sqrt{\log(p)/N}\right) \).

Therefore, we can apply Lemma 13 to obtain

\[
\left\| \mathbf{u}^{(k)} - \mathbf{u}_0 \right\|_2 = O_p \left( \sqrt{s_\beta \log(p)/N} + \sup_{k' \neq k} \left\| \beta^{(k,k')} - \beta_0 \right\|_2 \right)
= O_p \left( \sqrt{(s_\beta + s_u) \log(p)/N} + (1 - \rho) \sqrt{s_\beta \log(p + q)/n} \right).
\]

PART 2: Asymptotic approximation

Under Assumption 2b-i, we also have the tightness of \( \left\| \mathbf{u}^{(k)} \right\|_2 \) from the bound of \( \left\| \mathbf{u}_0 \right\|_2 \)

\[
\left\| \mathbf{u}_0 \right\|_2 \leq \left\| \Sigma_0^{-1/2} \mathbf{x}_{\text{str}} \right\|_2 \leq \sigma_m^{-2}, \quad \left\| \mathbf{u}^{(k)} \right\|_2 \leq \left\| \mathbf{u}_0 \right\|_2 + \left\| \mathbf{u}^{(k)} - \mathbf{u}_0 \right\|_2 = O_p(1).
\]

Define the scores of in-fold data as

\[
\ell^{(k)}(\beta; \gamma) = \frac{1}{N_k} \left[ \sum_{i \in \mathcal{J}_k} X_i \{ g(\beta^\top X_i) - g(\gamma^\top W_i) \} + \sum_{i \in \mathcal{I}_k} X_i \{ g(\beta^\top X_i) - Y_i \} \right].
\]

35
\[ \ell_{\text{imp}}^{(k)}(\gamma) = \frac{1}{n_k} \sum_{i \in I_k} X_i \{ g(\gamma^T W_i) - Y_i \}. \] (A.26)

Since \( \hat{x}_{\text{std}}^T \beta \) is the average over \( K \) (at most 10) cross-fitted estimators, it suffices to study one of the cross-fitted estimators,

\[ \hat{x}_{\text{std}}^T \beta^{(k)} = x_{\text{std}}^T \beta - \ell^{(k)}(\beta^{(k)}; \gamma^{(k)}) - (1 - \rho) \ell_{\text{imp}}^{(k)}(\gamma^{(k)}), \]

(\( k = 1 \ldots K \)) (A.27)

We denote the expected Hessian matrices of losses in (A.26) as

\[ \mathbb{H}(\beta) = \mathbb{E} \{ g'(\beta^T X_i) X_i^T \}, \quad \Sigma_o = \mathbb{H}(\beta_o), \]

\[ \mathbb{H}_{\text{imp}}(\gamma) = \mathbb{E} \{ g'(\gamma^T W_i) W_i^T \}, \quad \Sigma_{\text{imp}} = \mathbb{H}_{\text{imp}}(\gamma_0). \] (A.28)

Our analysis of the approximation error is based on the first order Mean Value Theorem identity,

\[ \mathbb{E} \{ \ell_{\text{imp}}^{(k)}(\beta^{(k)}; \hat{\gamma}^{(k)}) | \mathcal{F} \} = (1 - \rho) \mathbb{E} \{ \ell_{\text{imp}}^{(k)}(\hat{\gamma}^{(k)}) | \mathcal{F} \} \]

\[ = \mathbb{E} \{ \ell_{\text{imp}}^{(k)}(\beta_o, \gamma_0) \} + \mathbb{H}(\beta)(\hat{\beta}^{(k)} - \beta_o) - (1 - \rho) \mathbb{H}_{\text{imp}}(\hat{\gamma})(\hat{\gamma}^{(k)} - \gamma_0) \]

\[ + (1 - \rho) \mathbb{E} \{ \ell_{\text{imp}}^{(k)}(\gamma_0) \} + (1 - \rho) \mathbb{H}_{\text{imp}}(\hat{\gamma})(\hat{\gamma}^{(k)} - \gamma_0) \]

(\( k = 0 \)) (A.29)

for some \( \hat{\beta} \) on the path from \( \beta(\hat{\gamma}) \) to \( \beta_o \) and some \( \hat{\gamma} \) on the path from \( \hat{\gamma} \) to \( \gamma_0 \). The conditional expectation notation is declared at Definition 11. Based on (A.29), we analyze the approximation error for \( \sqrt{n} \left( x_{\text{std}}^T \beta^{(k)} - x_{\text{std}}^T \beta_o \right) \) through the following decomposition,

\[ \sqrt{n} \left( x_{\text{std}}^T \beta^{(k)} - x_{\text{std}}^T \beta_o \right) = \sqrt{n} x_{\text{std}}^T (\hat{\beta}^{(k)} - \beta_o) - \sqrt{n} u_0^T \ell^{(k)}(\beta_o; \gamma_0) + \sqrt{n} (1 - \rho) u_0^T \ell_{\text{imp}}^{(k)}(\gamma_0) \]

\[ + \sqrt{n} u_0^T \ell^{(k)}(\beta_o; \gamma_0) + (1 - \rho) u_0^T \ell_{\text{imp}}^{(k)}(\gamma_0) + \sqrt{n} \tilde{u}^{(k)} T \ell_{\text{imp}}^{(k)}(\beta^{(k)}; \gamma^{(k)}) | \mathcal{F}_k \]

\[ + \sqrt{n} (1 - \rho) \tilde{u}^{(k)} T \mathbb{E} \{ \ell_{\text{imp}}^{(k)}(\gamma^{(k)}) | \mathcal{F}_k \} - \sqrt{n} \tilde{u}^{(k)} T \mathbb{H}(\beta)(\hat{\beta}^{(k)} - \beta_o) \]

\[ = \sqrt{n} \left\{ x_{\text{std}} - \mathbb{H}(\beta) u_0 \right\}^T (\hat{\beta}^{(k)} - \beta_o) + \sqrt{n} \left( u_0 - \tilde{u}^{(k)} \right)^T \mathbb{H}(\beta)(\hat{\beta}^{(k)} - \beta_o) \]

\[ + \sqrt{n} \tilde{u}^{(k)} T \left[ \mathbb{E} \{ \ell_{\text{imp}}^{(k)}(\beta^{(k)}; \gamma^{(k)}) | \mathcal{F}_k \} - \{ \ell_{\text{imp}}^{(k)}(\beta^{(k)}; \gamma^{(k)}) - \ell_{\text{imp}}^{(k)}(\beta_o; \gamma_0) \} \right] \]

\[ + \sqrt{n} (1 - \rho) \tilde{u}^{(k)} T \left[ \mathbb{E} \{ \ell_{\text{imp}}^{(k)}(\gamma^{(k)}) | \mathcal{F}_k \} - \{ \ell_{\text{imp}}^{(k)}(\gamma^{(k)}) - \ell_{\text{imp}}^{(k)}(\gamma_0) \} \right] \]
Here we state the rates for $T_1-T_5$,

\[ T_1 = O_p \left( \sqrt{n} \left\| \beta^{(k)} - \beta_0 \right\| \right), \quad T_2 = O_p \left( \sqrt{n} \left\| \hat{\beta}^{(k)} - \beta_0 \right\| \right), \]

\[ T_3 = O_p \left( \rho \left\| \beta^{(k)} - \beta_0 \right\| + \sqrt{\rho(1-\rho)} \left\| \gamma^{(k)} - \gamma_0 \right\| \right), \]

\[ T_4 = O_p \left( (1-\rho) \left\| \hat{\gamma}^{(k)} - \gamma_0 \right\| \right), \quad T_5 = O_p \left( \left\| \hat{\beta}^{(k)} - \beta_0 \right\| \right). \]

With the assumed estimation rate in (A.22), we have

\[ T_1 + T_2 + T_3 + T_4 + T_5 = o_p(1). \]

Thus, we have shown

\[
\sqrt{n} \left( \hat{x}_{\text{std}}^T \beta - x_{\text{std}}^T \beta_0 \right) = \frac{1}{K} \sum_{k=1}^{K} \sqrt{n} \left( \hat{x}_{\text{std}}^T \beta - x_{\text{std}}^T \beta_0 \right)
= \frac{1}{K} \sum_{k=1}^{K} -\sqrt{n} u_k^T \hat{\epsilon}^{(k)}(\beta_0; \gamma_0) - \sqrt{n} (1-\rho) u_k^T \hat{\epsilon}_{\text{imp}}(\gamma_0) + o_p(1)
= -\sqrt{n} u_k^T \hat{\epsilon}^{(k)}(\beta_0; \gamma_0) - \sqrt{n} (1-\rho) u_k^T \hat{\epsilon}_{\text{imp}}(\gamma_0) + o_p(1).
\]

Using the indicator $R_i = I(i \leq n)$, we can alternatively write

\[
\hat{x}_{\text{std}}^T \beta - x_{\text{std}}^T \beta_0 = \frac{1}{N} \sum_{i=1}^{N} \frac{R_i}{\rho} u_i^T X_i \{ g(\gamma_0^T W_i) - g(\beta_0^T X_i) \} + o_p \left( (\rho n)^{-1/2} \right).
\]

We provide the details of $T_1-T_5$ in Section C2.

**PART 3: VARIANCE ESTIMATION**

Finally, we show that asymptotic variance $V_{SAS}$ defined in (29) is bounded from infinity and zero with the consistent estimator $\hat{V}_{SAS}$ defined in (23).

By the Cauchy-Schwartz inequality, we have a bound for the variance

\[
V_{SAS} = \mathbb{E} \left[ (u_0^T X_i)^2 \{ (1-\rho) \cdot g(\gamma_0^T W_i) + \rho \cdot g(\beta_0^T X_i) - Y_i \}^2 \right]
+ \rho (1-\rho) \mathbb{E} \left[ (u_0^T X_i)^2 \{ g(\gamma_0^T W_i) - g(\beta_0^T X_i) \}^2 \right]
\leq \sqrt{\mathbb{E} \left[ (u_0^T X_i)^4 \right] \mathbb{E} \left[ \{ (1-\rho) \cdot g(\gamma_0^T W_i) + \rho \cdot g(\beta_0^T X_i) - Y_i \}^4 \right]}
+ \rho (1-\rho) \sqrt{\mathbb{E} \left[ (u_0^T X_i)^4 \right] \mathbb{E} \left[ \{ g(\gamma_0^T W_i) - g(\beta_0^T X_i) \}^4 \right]}
\]

Under Assumptions 1a, 2a, we have the sub-Gaussian and sub-exponential variables

\[
\| u_0^T X_i \|_{\psi_2} \leq \| u_0 \|_{\psi_2} \frac{2 \sigma_{\max}}{\sqrt{2}} \leq \sigma_{\min}^{-2} \sigma_{\max} / \sqrt{2},
\]

37
\[
\| (1 - \rho) \cdot g(\gamma_t^T W_i) + \rho \cdot g(\beta_0^T X_i) - Y_i \|_2 \leq 2(\nu_1 \lor \nu_2), \\
\| g(\gamma_t^T W_i) - g(\beta_0^T X_i) \|_2 \leq 2 \{ \| g(\gamma_t^T W_i) - Y_i \|_2 \lor \| g(\beta_0^T X_i) - Y_i \|_2 \} \leq 2(\nu_1 \lor \nu_2). 
\]

By the bound for the moments of sub-Gaussian and sub-exponential random variables stated in Lemma 17-b, we have
\[
V_{\text{SAS}} \leq 8\sqrt{2}\sigma_{\text{min}}^{-4}\sigma_{\text{max}}^2(\nu_1 \lor \nu_2)^2.
\]

Under Assumptions 1-b, 2a, 2b-i and 2-c, we have a lower bound for \( V_{\text{SAS}} \),
\[
\begin{align*}
V_{\text{SAS}} &\geq u_0^T \mathbb{E}[X_i X_i^T \{(1 - \rho) \cdot g(\gamma_t^T W_i) + \rho \cdot g(\beta_0^T X_i) - Y_i\}^2] u_0 \\
&\geq \|u_0\|^2_2 \rho_{\text{min}}^{-4} \\
&\geq M^{-2}\sigma_{\text{max}}^{-4}\sigma_{\text{min}}^{-4}(\nu_3 \|X_{\text{std}}\|_2^2,
\end{align*}
\]

which is bounded away from zero.

We analyze the estimation error of variance \( \hat{V}_{\text{SAS}} - V_{\text{SAS}} \) through the decomposition,
\[
\begin{align*}
\hat{V}_{\text{SAS}} - V_{\text{SAS}} \\
= \sum_{k=1}^K n_k \left( \frac{1}{n_k} \sum_{i \in I_k} (\widehat{\mu}^{(k)^T} X_i)^2 \{(1 - \rho) \cdot g(\widehat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\widehat{\beta}^{(k)^T} X_i) - Y_i\}^2 \\
- \mathbb{E}_{i \in I_k} \left[ (\widehat{\mu}^{(k)^T} X_i)^2 \{(1 - \rho) \cdot g(\widehat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\widehat{\beta}^{(k)^T} X_i) - Y_i\}^2 \mid \mathcal{F}_k \right] \right) \\
+ \sum_{k=1}^K n_k \left( \mathbb{E}_{i \in I_k} \left[ (\widehat{\mu}^{(k)^T} X_i)^2 \{(1 - \rho) \cdot g(\widehat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\widehat{\beta}^{(k)^T} X_i) - Y_i\}^2 \mid \mathcal{F}_k \right] \\
- \mathbb{E} \left[ (u_0^T X_i)^2 \{(1 - \rho) \cdot g(\gamma_t^T W_i) + \rho \cdot g(\beta_0^T X_i) - Y_i\}^2 \right] \right) \\
+ \rho(1 - \rho) \sum_{k=1}^K \frac{N_k - n_k}{N - n} \left( \frac{K}{N_k - n_k} \sum_{i \in J_k} (\widehat{\mu}^{(k)^T} X_i)^2 \left\{ g(\widehat{\beta}^{(k)^T} X_i) - g(\widehat{\gamma}^{(k)^T} W_i) \right\}^2 \\
- \mathbb{E}_{i \in I_k} \left[ (\widehat{\mu}^{(k)^T} X_i)^2 \left\{ g(\widehat{\beta}^{(k)^T} X_i) - g(\widehat{\gamma}^{(k)^T} W_i) \right\}^2 \mid \mathcal{F}_k \right] \right) \\
+ \rho(1 - \rho) \sum_{k=1}^K \frac{N_k - n_k}{N - n} \left( \mathbb{E}_{i \in I_k} \left[ (\widehat{\mu}^{(k)^T} X_i)^2 \left\{ g(\widehat{\beta}^{(k)^T} X_i) - g(\widehat{\gamma}^{(k)^T} W_i) \right\}^2 \mid \mathcal{F}_k \right] \\
- \mathbb{E} \left[ (u_0^T X_i)^2 \left\{ g(\beta_0^T X_i) - g(\gamma_t^T W_i) \right\}^2 \right] \right)
\end{align*}
\]

Here we state the rates for \( T_1', T_2', T_3', T_4' \),
\[
T_1' = O_p \left( n^{-1/2} \right), \\
T_2' = O(p \left( \| \hat{\mu} - u_0 \|_2 + (1 - \rho) \| \hat{\gamma} - \gamma_0 \|_2 + \rho \| \hat{\beta} - \beta_0 \|_2 \right), \\
T_3' = O(p (1 - \rho) N^{-1/2}), \\
T_4' = O(p (1 - \rho) \left\{ \| \hat{\mu} - u_0 \|_2 + \| \hat{\gamma} - \gamma_0 \|_2 + \| \hat{\beta} - \beta_0 \|_2 \right\}.
\]

38
With the assumed estimation rate in (A.22), we have

$$T_1' + T_2' + T_3' + T_4' = o_p(1).$$

We provide the details of $T_1', T_2'$ in Section C2.

**PART 4: CONCLUSION WITH ESTIMATION RATES**

From the approximation in Part 2 and the boundedness and non-degeneracy of $V_{SAS}$ in Part 3, we have shown the asymptotic normality of the cross-fitted debiased estimator

$$\sqrt{n}V_{SAS}^{-1/2} \left( \hat{\beta} - \beta_0 \right) \Rightarrow N(0, 1).$$

Together with the consistency of $\hat{V}_{SAS}$ in Part 3, we have

$$\sqrt{n}\hat{V}_{SAS}^{-1/2} \left( \hat{\beta} - \beta_0 \right) \Rightarrow N(0, 1).$$

**PART 5: SUFFICIENT DIMENSION CONDITION**

We have established the rate of estimation for $\hat{\gamma}, \hat{\beta}$ and $\hat{u}$ from Lemma 21, Theorem 1 and Part 4 of this proof above. Since we only keep one fold of the data away for the cross-fitted estimators, they follow the same rates of estimation,

$$\|\hat{\gamma}^{(k)} - \gamma_0\|_2 = O_p \left( \sqrt{s_y \log(p + q)/n} \right),$$

$$\|\hat{\beta}^{(k)} - \beta_0\|_2 = O_p \left( \sqrt{s_{\beta} \log(p)/N + (1 - \rho)\sqrt{s_y \log(p + q)/n}} \right),$$

$$\|\hat{u}^{(k)} - u_0\|_2 = O_p \left( \sqrt{(s_{\beta} + s_y) \log(p)/N + (1 - \rho)\sqrt{s_y \log(p + q)/n}} \right).$$

Applying the rates of estimation, we show dimension assumption (28) is sufficient for (A.22).

**B4 Efficiency of SAS Inference**

**Relative Efficiency to Supervised Learning**

**Proof** [Proof of Proposition 9] We prove the Proposition by direct calculation

$$V_{SL} - V_{SAS}$$

$$= \mathbb{E}(\mathbb{E}(u_i'X_i)^2(Y - g(\beta_0'X_i))^2) - \mathbb{E}(\mathbb{E}(u_i'X_i)^2(Y - (1 - \rho)\mathbb{E}(Y|S_i, X_i)\rho \cdot g(\beta_0'X_i))^2)$$

$$= \mathbb{E}(\mathbb{E}(u_i'X_i)^2((1 - \rho)^2g(\beta_0'X_i)^2 - 2(1 - \rho)g(\beta_0'X_i)\mathbb{E}(Y|S_i, X_i) + (1 - \rho^2)\mathbb{E}(Y|S_i, X_i)^2))$$

$$= (1 - \rho)^2\mathbb{E}(\mathbb{E}(u_i'X_i)^2(\mathbb{E}(Y|S_i, X_i) - g(\beta_0'X_i))^2)$$

$$+ 2\rho(1 - \rho)\mathbb{E}(\mathbb{E}(u_i'X_i)^2(\mathbb{E}(Y|S_i, X_i)^2 + g(\beta_0'X_i)^2)).$$

The last expression is the sum of expectations of complete squares, so it must be non-negative. Thus, we have shown that the SAS asymptotic variance is no greater than the supervised learning variance. The equality holds only if 1) $\rho = 1$ all samples are labelled; 2) or $\rho = 0$ and $u_i'X_i(\mathbb{E}(Y|S_i, X_i) - g(\beta_0'X_i)) = 0$ almost surely. \[\square\]
Proof [Proof of Proposition 10] The proof follows the flow of Section D.2 in Kallus and Mao (2020). The semi-parametric model for the observed data is

\[
\mathcal{M}_{obs} = \left\{ f_{X,Y,S,R}(x,y,s,r) = f_X(x) f_{S|X}(s|x) \left\{ \rho f_{Y|S,X}(y|s,x) \right\}^r (1 - \rho)^{1-r} : f_X, f_{S|X}, f_{Y|S,X} \text{ are arbitrary pdf/pmf}, \right\}.
\]  

(A.32)

We consider the parametric sub-model

\[
\mathcal{M}_{par} = \left\{ f_{X,Y,S,R}(x,y,s,r; \zeta) = f_X(x; \zeta) f_{S|X}(s|x; \zeta) \left\{ \rho f_{Y|S,X}(y|s,x; \zeta) \right\}^r \times (1 - \rho)^{1-r} : \zeta \in \mathbb{R}^d \right\}.
\]  

(A.33)

The score vector of the parametric sub-model is

\[
\Psi(X,Y,S,R) = \left. \frac{\partial \log \left\{ f_{X,Y,S,R}(X,Y,S,R; \zeta) \right\}}{\partial \zeta} \right|_{\zeta = \zeta_0} = \left. \frac{\partial \log \left\{ f_X(X; \zeta) \right\}}{\partial \zeta} \right|_{\zeta = \zeta_0} + \left. \frac{\partial \log \left\{ f_{S|X}(S|X; \zeta) \right\}}{\partial \zeta} \right|_{\zeta = \zeta_0} + R \left. \frac{\partial \log \left\{ f_{Y|S,X}(Y|S,X; \zeta) \right\}}{\partial \zeta} \right|_{\zeta = \zeta_0} = \Psi_X(X) + \Psi_S(S,X) + R\Psi_Y(Y,S,X).
\]  

(A.34)

Next, we decompose the Hilbert space of mean zero finite variance random variables measurable to \( \sigma\{X,S,R,Y,R\} \), denoted as \( \mathcal{H} \). The model tangent space spanned by the score (A.34) is a linear sub-space of \( \mathcal{H} \),

\[
\Lambda = \Lambda_X \oplus \Lambda_S \oplus \Lambda_Y,
\]

\[
\Lambda_X = \bigcup_{\mathcal{M}_{par}} \text{span}\{\Psi_X(X)\} = \{ h(X) \in \mathcal{H} : \mathbb{E}[h(X)] = 0 \},
\]

\[
\Lambda_S = \bigcup_{\mathcal{M}_{par}} \text{span}\{\Psi_S(S,X)\} = \{ h(S,X) \in \mathcal{H} : \mathbb{E}[h(S,X) | X = 0] = 0 \},
\]

\[
\Lambda_Y = \bigcup_{\mathcal{M}_{par}} \text{span}\{R\Psi_Y(Y,S,X)\} = \{ Rh(Y,S,X) \in \mathcal{H} : \mathbb{E}[h(Y,S,X) | S,X] = 0 \}.
\]  

(A.35)

The orthogonal space of model tangent space \( \Lambda \) is

\[
\Lambda^\perp = \{ h(R,S,X) \in \mathcal{H} : \mathbb{E}[h(R,S,X) | S,X] = 0 \}, \mathcal{H} = \Lambda \oplus \Lambda^\perp.
\]  

(A.36)

Now, we verify that the supervised learning influence function

\[
\phi_{SL}(\theta; \beta) = \frac{R}{\rho} u_s^T X \{ Y - g(\beta^T X) \}
\]
is indeed an influence function for \( x_{\text{std}}^T \beta \) by showing

\[
\mathbb{E}\{\phi_{\text{sl}}(\theta_0; \beta_0) \Psi(X, Y, S, R)\} = x_{\text{std}} \frac{d}{d\zeta} \beta(\zeta) \bigg|_{\zeta=\zeta_0}.
\]

Since \( \beta(\zeta) \) is an implicit function of \( \zeta \) through the moment condition

\[
\mathbb{E}_{\zeta}[X\{g(\beta(\zeta)^T X) - Y\} = 0,
\]

we solve for its derivative by differentiating the moment condition

\[
\frac{d}{d\zeta} \mathbb{E}_{\zeta}[X\{g(\beta(\zeta)^T X) - Y\}]\bigg|_{\zeta=\zeta_0} = 0
\]

and

\[
\frac{d}{d\zeta} \mathbb{E}_{\zeta}[X\{g(\beta_0^T X) - Y\} + \mathbb{E}_{\zeta_0} \{XX^T g'(\beta_0^T X)\} \frac{d}{d\zeta} \beta(\zeta)\bigg|_{\zeta=\zeta_0} = 0
\]

Then, we verify that the supervised learning influence function is valid

\[
\frac{d}{d\zeta} x_{\text{std}}^T \beta(\zeta)\bigg|_{\zeta=\zeta_0} = -\mathbb{E}_{\zeta_0} \left[ \frac{R}{\rho} u_0^T X\{g(\beta_0^T X) - Y\} \Psi(X, Y, S, R) \right]
\]

\[
= \mathbb{E}\{\phi_{\text{sl}}(\theta_0; \beta_0) \Psi(X, Y, S, R)\}.
\]

Finally, we derive the efficient influence function by subtract from \( \phi_{\text{sl}} \) its projection onto \( \Lambda^\perp = \Lambda_R \). Let \( \Pi[h(D) \mid \Lambda] \) be the projection of \( h(D) \in \mathcal{H} \) to the space \( \Lambda \). We can easily calculate the projection of \( \phi_{\text{sl}} \) onto \( \Lambda_R \),

\[
\Pi[\phi_{\text{sl}}(\theta_0; \beta_0) \mid \Lambda_R] = \mathbb{E}\{\phi_{\text{sl}}(\theta_0; \beta_0) \mid R, S, X\} - \mathbb{E}\{\phi_{\text{sl}}(\theta_0; \beta_0) \mid S, X\}
\]

\[
= \frac{R}{\rho} u_0^T X\{\mathbb{E}(Y \mid S, X) - g(\beta^T X)\} - \frac{R}{\rho} u_0^T X\{\mathbb{E}(Y \mid S, X) - g(\beta^T X)\}.
\]

The efficient influence function is thus obtained

\[
\phi_{\text{eff}}(\theta_0; \beta_0) = \phi_{\text{sl}}(\theta_0; \beta_0) - \Pi[\phi_{\text{sl}}(\theta_0; \beta_0) \mid \Lambda_R]
\]

\[
= \frac{R}{\rho} u_0^T X\{Y - g(\beta^T X)\} - \frac{R}{\rho} u_0^T X\{\mathbb{E}(Y \mid S, X) - g(\beta^T X)\}
\]

\[
+ u_0^T X\{\mathbb{E}(Y \mid S, X) - g(\beta^T X)\}
\]

\[
= \frac{R}{\rho} u_0^T X\{Y - \mathbb{E}(Y \mid S, X)\} + u_0^T X\{\mathbb{E}(Y \mid S, X) - g(\beta^T X)\}.
\]

\[\blacksquare\]
Appendix C. Auxiliary Results

C1 General

Lemma 14 Under Assumptions 1a, 1b, 2a, the residuals of the imputed loss are sub-Gaussian random variables,

\[ \| g(\beta^T X_i) - g(\gamma^T W_i) \|_{\psi_2} \leq 4 \max \{ \nu_1, \nu_2, M \| \beta - \beta_o \| 2 \sigma_{\text{max}} / \sqrt{2}, M \| \gamma - \gamma_o \| 2 \sigma_{\text{max}} / \sqrt{2} \} \]

Similarly,

\[ \| Y_i - g(\gamma^T W_i) \|_{\psi_2} \leq 2 \max \{ \nu_1, M \| \gamma - \gamma_o \| 2 \sigma_{\text{max}} / \sqrt{2} \}, \]
\[ \| g(\gamma^T W_i) - g(\gamma_o^T W_i) \|_{\psi_2} \leq M \| \gamma - \gamma_o \| 2 \sigma_{\text{max}} / \sqrt{2}, \]
\[ \| g(\beta^T X_i) - g(\beta_o^T X_i) \|_{\psi_2} \leq M \| \beta - \beta_o \| 2 \sigma_{\text{max}} / \sqrt{2}, \]
\[ \| \rho \cdot g(\beta^T X_i) + (1 - \rho) \cdot g(\gamma^T W_i) - Y \|_{\psi_2} \leq 4 \max \{ \nu_1, \nu_2, \rho M \| \beta - \beta_o \| 2 \sigma_{\text{max}} / \sqrt{2}, \rho M \| \gamma - \gamma_o \| 2 \sigma_{\text{max}} / \sqrt{2} \} \]
\[ \| g(\beta^T X_i) - g(\beta_o^T X_i) - g(\gamma^T W_i) + g(\gamma_o^T W_i) \|_{\psi_2} \leq \sqrt{2} M \sigma_{\text{max}} \max \{ \| \beta - \beta_o \|, \| \gamma - \gamma_o \| \} \]

Proof [Proof of Lemma 14] To establish the sub-exponential tail, we consider the following decomposition

\[ g(\beta^T X_i) - g(\gamma^T W_i) = \{ g(\beta_o^T X_i) - Y_i \} - \{ g(\gamma_o^T W_i) - Y_i \} \]
\[ + \{ g(\beta^T X_i) - g(\beta_o^T X_i) \} - \{ g(\gamma^T W_i) - g(\gamma_o^T W_i) \}. \]

(A.37)

According to Assumption 1a, the first two terms on the right-hand side of (A.37) are sub-Gaussian,

\[ \| g(\beta_o^T X_i) - Y_i \|_{\psi_2} \leq \nu_1, \| g(\gamma_o^T W_i) - Y_i \|_{\psi_2} \leq \nu_2. \]

According to Assumption 1b, the latter two terms on the right-hand side of (A.37) are bounded by

\[ \| g(\beta^T X_i) - g(\beta_o^T X_i) \|_{\psi_2} \leq M \| (\beta - \beta_o)^T X_i \|_{\psi_2} \leq M \| (\beta - \beta_o)^T X_i \|_{\psi_2} \leq M \| (\gamma - \gamma_o)^T W_i \|. \]

Under Assumption 2a, \((\beta - \beta_o)^T X_i\) and \((\gamma - \gamma_o)^T W_i\) are sub-Gaussian random variables,

\[ \| (\beta - \beta_o)^T X_i \|_{\psi_2} \leq \| \beta - \beta_o \| 2 \sigma_{\text{max}} / \sqrt{2}, \]
\[ \| (\gamma - \gamma_o)^T W_i \|_{\psi_2} \leq \| \gamma - \gamma_o \| 2 \sigma_{\text{max}} / \sqrt{2}. \]

By Lemma 17-e,

\[ \| g(\beta^T X_i) - g(\beta_o^T X_i) \|_{\psi_2} \leq M \| (\beta - \beta_o)^T X_i \|_{\psi_2} \leq M \| \beta - \beta_o \| 2 \sigma_{\text{max}} / \sqrt{2} \]
\[ \| g(\gamma^T W_i) - g(\gamma_o^T W_i) \|_{\psi_2} \leq \| (\gamma - \gamma_o)^T W_i \|_{\psi_2} \leq M \| \gamma - \gamma_o \| 2 \sigma_{\text{max}} / \sqrt{2}. \]

Finally, we apply Lemma 17-d

\[ \| g(\beta^T X_i) - g(\gamma^T W_i) \|_{\psi_2} \leq 4 \max \{ \| g(\beta_o^T X_i) - Y_i \|_{\psi_2}, \| g(\gamma_o^T W_i) - Y_i \|_{\psi_2}, \]
\[ \| g(\beta^T X_i) - g(\beta_o^T X_i) \|_{\psi_2}, \| g(\gamma^T W_i) - g(\gamma_o^T W_i) \|_{\psi_2} \} \]
\[ \leq 4 \max \{ \nu_1, \nu_2, M \| \beta - \beta_o \| 2 \sigma_{\text{max}} / \sqrt{2}, M \| \gamma - \gamma_o \| 2 \sigma_{\text{max}} / \sqrt{2} \}. \]

Therefore, we have reached the conclusion.

We may obtain the rest of bounds following the same derivation.
C2 Inference

Analysis of Estimated Precision Matrix

Proof [Proof of Lemma 13]

The definition of the cross-fitted loss functions \( m^{(k,k')} \) and their derivatives can be found at (A.23). By the definition of \( \hat{u}^{(k)} \), we have

\[
\sum_{k' \neq k} \frac{N_{k'}}{N - N_k} m^{(k,k')} \left( \hat{u}^{(k)}; \hat{\beta}^{(k,k')} \right) + \lambda_u \| \hat{u}^{(k)} \|_1 \leq \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) + \lambda_u \| u_0 \|_1.
\]

Denote the standardized estimation error as \( \delta = (\hat{u}^{(k)} - u_0)/\| \hat{u}^{(k)} - u_0 \|_2 \). Due to convexity of the loss function, we have for \( t = \| \hat{u}^{(k)} - u_0 \|_2 \wedge 1 \)

\[
\sum_{k' \neq k} \frac{N_{k'}}{N - N_k} m^{(k,k')} \left( u_0 + t\delta; \hat{\beta}^{(k,k')} \right) + \lambda_u \| u_0 + t\delta \|_1 \leq \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) + \lambda_u \| u_0 \|_1.
\]

By the triangle inequality \( \| u_0 \|_1 - \| u_0 + t\delta \|_1 \leq t\| \delta \|_1 \), we have from (A.38)

\[
\sum_{k' \neq k} \frac{N_{k'}}{N - N_k} \left( m^{(k,k')} \left( u_0 + t\delta; \hat{\beta}^{(k,k')} \right) - m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) \right) \leq t\lambda_u \| \delta \|_1
\]

(A.39)

Because the loss functions \( m^{(k)} \) are quadratic functions of \( u \), we can apply the restricted strong convexity event \( \Omega^{(k)} \) to obtain

\[
m^{(k,k')} \left( u_0 + t\delta; \hat{\beta}^{(k,k')} \right) - m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) = t^2 \delta^T m^{(k,k')} \left( \hat{\beta}^{(k,k')} \right) \delta \\
\geq t^2 \kappa_{\text{rec},1}^* - 2t \kappa_{\text{rec},1}^* \kappa_{\text{rec},2}^* \sqrt{\log(p)/N_{k'}} \| \delta \|_1.
\]

(A.40)

Applying (A.40) to (A.39), we have with large probability

\[
\sum_{k' \neq k} \frac{N_{k'}}{N - N_k} \left( t\delta^T m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) + t^2 \kappa_{\text{rec},1}^* - 2t \kappa_{\text{rec},1}^* \kappa_{\text{rec},2}^* \sqrt{\log(p)/N_{k'}} \| \delta \|_1 \right) \leq t\lambda_u \| \delta \|_1
\]

where \( \| \delta \|_2 = 1 \) from definition. Thus, we have reach

\[
t \kappa_{\text{rec},1}^* \leq \lambda_u \| \delta \|_1 - \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} \left( \delta^T m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) - \kappa_{\text{rec},1}^* \kappa_{\text{rec},2}^* \sqrt{\log(p)/N_{k'}} \| \delta \|_1 \right)
\]

(A.41)

The target parameter \( u_0 \) can be identify by \( \mathbb{E} \left\{ m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) \mid \mathcal{D}'_{k'} \right\} = 0 \). We use the fact to do a careful analysis of \( \delta^T m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) \) by the decomposition

\[
\left\| \delta^T m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) \right\| = \delta^T \left[ m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) - \mathbb{E} \left\{ m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) \mid \mathcal{D}'_{k'} \right\} \right]
\]

\[
+ \delta^T \left\{ \mathbb{E} \left\{ m^{(k,k')} \left( u_0; \hat{\beta}^{(k,k')} \right) \mid \mathcal{D}'_{k'} \right\} - \mathbb{E} \{ m^{(k,k')} \left( u_0; \beta_0 \right) \mid \mathcal{D}'_{k'} \} \right\}
\]

43
Thus, we have
\[ \| \mathbf{m}^{(k,k')} (\mathbf{u}_0; \hat{\beta}^{(k,k')}) - \mathbb{E} \{ \mathbf{m}^{(k,k')} (\mathbf{u}_0; \beta^{(k,k')}) | \mathcal{D}_k \} \|_{\infty} \] 
\[ + \left\| \mathbb{E}_{i \in \mathcal{I}_k \cup \mathcal{J}_k} [\mathbf{v}^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{u}_0 \{ g(\beta_0^T \mathbf{x}_i) - g(\hat{\beta}^{(k,k')}^T \mathbf{x}_i) \} | \mathcal{D}_k] \right\|_2. \] (A.42)

We establish the rate for $L_2$-norm of the population score at $\mathbf{u}_0$ through analyzing
\[ \sup_{\| \mathbf{v} \|_2 = 1} \mathbb{E}_{i \in \mathcal{I}_k \cup \mathcal{J}_k} [\mathbf{v}^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{u}_0 \{ g(\beta_0^T \mathbf{x}_i) - g(\hat{\beta}^{(k,k')}^T \mathbf{x}_i) \} | \mathcal{D}_k], \]
whose bound can be derived from Assumptions 1b, 2a, 2a, the Cauchy-Schwartz inequality and Lemma 17-b,
\[ \mathbb{E}_{i \in \mathcal{I}_k \cup \mathcal{J}_k} [\mathbf{v}^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{u}_0 \{ g(\beta_0^T \mathbf{x}_i) - g(\hat{\beta}^{(k,k')}^T \mathbf{x}_i) \} | \mathcal{D}_k] \]
\[ \leq M \mathbb{E}_{i \in \mathcal{I}_k \cup \mathcal{J}_k} [\| \mathbf{v}^T \mathbf{X}_i \mathbf{X}_i^T \mathbf{u}_0 \{ (\beta_0 - \hat{\beta}^{(k,k')})^T \mathbf{x}_i \} | \mathcal{D}_k] \]
\[ \leq M \left[ \mathbb{E} \{ \| \mathbf{v}^T \mathbf{X} \|_2 \} \mathbb{E} \{ \| \mathbf{v}^T \mathbf{X} \|_2 \} \right]^{1/4} \sqrt{\mathbb{E}_{i \in \mathcal{I}_k \cup \mathcal{J}_k} \left\{ \| (\beta_0 - \hat{\beta}^{(k,k')})^T \mathbf{x}_i \|_2 \right\}^2 | \mathcal{D}_k] \]
\[ \leq M \sigma_{\max}^3 \| \mathbf{v} \|_2 \| \mathbf{u}_0 \|_2 \| \beta_0 - \hat{\beta}^{(k,k')} \|_2. \]

Hence, we have shown
\[ \left\| \mathbb{E}_{i \in \mathcal{I}_k \cup \mathcal{J}_k} \left[ \mathbf{X}_i \mathbf{X}_i^T \mathbf{u}_0 \left\{ g(\beta_0^T \mathbf{x}_i) - g(\hat{\beta}^{(k,k')}^T \mathbf{x}_i) \right\} | \mathcal{D}_k \right] \right\|_2 \leq M \sigma_{\max}^3 \| \mathbf{u}_0 \|_2 \| \beta_0 - \hat{\beta}^{(k,k')} \|_2. \] (A.43)

By the bound for (A.42) through (A.43) and the definition of $\lambda_0$, we have the bound from (A.41)
\[ t \kappa_{\text{rec,1}}^* \leq 2 \lambda_0 \| \delta \|_1 + M \sigma_{\max}^3 \| \mathbf{u}_0 \|_2 \sup_{k' \neq k} \| \beta_0 - \hat{\beta}^{(k,k')} \|_2. \] (A.44)

Hence, we can reach an immediate bound for estimation error from (A.44) without considering the sparsity of $\mathbf{u}_0$. We shall proceed to derive a sharper bound that involves the sparsity of $\mathbf{u}_0$. We separately analyze two cases.

**Case 1:**
\[ M \sigma_{\max}^3 \| \mathbf{u}_0 \|_2 \sup_{k=1, \ldots, K} \| \beta_0 - \hat{\beta}^{(k,k')} \|_2 \geq \| \delta \|_1 \lambda_0 / 3 \]

In this case, the estimation error is dominated by $\beta_0 - \hat{\beta}^{(k,k')}$. We simply have from (A.44)
\[ t \kappa_{\text{rec,1}}^* \leq 7 M \sigma_{\max}^3 \| \mathbf{u}_0 \|_2 \sup_{k=1, \ldots, K} \| \beta_0 - \hat{\beta}^{(k,k')} \|_2, \]
\[ t \kappa_{\text{rec,1}}^* \| \delta \|_1 \lambda_0 / 3 \leq 7 M \sigma_{\max}^3 \| \mathbf{u}_0 \|_2 \sup_{k=1, \ldots, K} \| \beta_0 - \hat{\beta}^{(k,k')} \|_2^2. \]

Thus, we have
\[ \| \hat{\mathbf{u}}^{(k)} - \mathbf{u}_0 \|_2 \leq 7 M \sigma_{\max}^3 \| \mathbf{u}_0 \|_2 \sup_{k' \neq k} \| \beta_0 - \hat{\beta}^{(k,k')} \|_2 / \kappa_{\text{rec,1}}^*, \]
\[
\|\hat{u}^{(k)} - u_0\|_1 \leq 21M \sigma_{\text{max}}^2 \|u_0\|_2^2 \sup_{k' \neq k} \|\beta_0 - \hat{\beta}^{(k,k')}\|_2^2 / (\kappa_{\text{nc,1}}^* \lambda_u). \quad (A.45)
\]

**Case 2:**
\[
M \sigma_{\text{max}}^3 \|u_0\|_2 \sup_{k' \neq k} \|\beta_0 - \hat{\beta}^{(k,k')}\|_2 \leq \|\delta\|_1 \lambda_u / 3 \quad (A.46)
\]

In this case, the estimation error is comparable to the situation that we have the true \( \beta_0 \) for the Hessian. Thus, the sparsity of \( u_0 \) may affect the estimation error.

Following the typical approach to establish the cone condition for \( \delta \), we analyze the symmetrized Bregman’s divergence,
\[
(\hat{u}^{(k)} - u_0)^T \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} \left\{ m^{(k,k')} (\hat{u}^{(k)}; \hat{\beta}^{(k,k')}) - m^{(k,k')} (u_0; \beta^{(k,k')}) \right\}
\]
\[
= \|\hat{u}^{(k)} - u_0\|_2 \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} \delta^T \left\{ m^{(k,k')} (\hat{u}^{(k)}; \hat{\beta}^{(k,k')}) - m^{(k,k')} (u_0; \beta^{(k,k')}) \right\}. \quad (A.47)
\]
Due to the convexity of the quadratic loss \( m^{(k,k')} (\cdot; \hat{\beta}^{(k,k')}) \), the symmetrized Bregman’s divergence (A.47) is nonnegative through a mean-value theorem,
\[
(\hat{u}^{(k)} - u_0)^T \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} \left\{ m^{(k,k')} (\hat{u}^{(k)}; \hat{\beta}^{(k,k')}) - m^{(k,k')} (u_0; \beta^{(k,k')}) \right\}
\]
\[
= \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} \sum_{\hat{u} \in I_k \cup I_{k'}} g' (\beta^{(k,k')}^T X_i)^2 \left\{ (\hat{u}^{(k)} - u_0)^T X_i \right\}^2 \geq 0.
\]

Denote the indices set of nonzero coefficient in \( u_0 \) as \( \mathcal{O}_u = \{ j : u_{0,j} \neq 0 \} \). We denote the \( \delta_{\mathcal{O}_u} \) and \( \delta_{\mathcal{O}_u^c} \) as the sub-vectors for \( \delta \) at positions in \( \mathcal{O}_u \) and at positions not in \( \mathcal{O}_u \), respectively. The solution \( \hat{u}^{(k)} \) satisfies the KKT condition
\[
\left\| \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} m^{(k,k')} (\hat{u}^{(k)}; \hat{\beta}^{(k,k')}) \right\|_\infty \leq \lambda_u,
\]
\[
\sum_{k' \neq k} \frac{N_{k'}}{N - N_k} m^{(k,k')} (\hat{u}^{(k)}; \hat{\beta}^{(k,k')})_j = -\lambda_u \text{sign}(\hat{u}^{(k)})_j, j : \hat{u}^{(k)}_j \neq 0.
\]
From the KKT condition and the definitions of \( \delta \) and \( \mathcal{O}_u \), we have
\[
\delta_j \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} m^{(k,k')} (\hat{u}^{(k)}; \hat{\beta}^{(k,k')})_j \leq |\delta_j| \lambda_u, j \in \mathcal{O}_u;
\]
\[
\delta_j \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} m^{(k,k')} (\hat{u}^{(k)}; \hat{\beta}^{(k,k')})_j = -\frac{\hat{u}^{(k)}_j \lambda_u \text{sign}(\hat{u}^{(k)})_j}{\|\hat{u}^{(k)} - u_0\|_2} = -\lambda_u |\delta_j|, j \in \mathcal{O}_u^c. \quad (A.48)
\]
Applying the (A.48) to (A.47), we have the upper bound,
\[
\delta^T \sum_{k' \neq k} \frac{N_{k'}}{N - N_k} \left\{ m^{(k,k')} (\hat{u}^{(k)}; \hat{\beta}^{(k,k')}) - m^{(k,k')} (u_0; \beta^{(k,k')}) \right\}
\]
which is our oracle inequality.

Thus, we obtain the rate for estimation error

\[ \lambda \]

Then, we apply (A.42), the definition of \( \lambda_u \) and (A.46),

\[ 0 \leq \lambda_u \| \delta_{O_u} \|_1 - \lambda_u \| \delta_{C_u} \|_1 + \frac{2}{3} \lambda_u \| \delta \|_1, \quad \text{and} \quad \lambda_u \| \delta_{O_u} \|_1 \leq 5 \lambda_u \| \delta_{C_u} \|_1. \]

Therefore, we can bound the \( L_1 \) norm of \( \delta \) by the cone property,

\[ \| \delta \|_1 \leq 6 \lambda_u \| \delta_{O_u} \|_1 \leq 6 \sqrt{s_u} \| \delta \|_2 = 6 \sqrt{s_u}. \] (A.49)

Now, we apply the cone condition (A.49) and the case condition (A.46) to the bound (A.44),

\[ tr^{*}_{sc, 1} \leq 14 \sqrt{s_u} \lambda_u, \quad tr^{*}_{sc, 1} \| \delta \|_1 \leq 84 s_u \lambda_u \]

Thus, we obtain the rate for estimation error

\[ \| \hat{u}^{(k)} - u_0 \|_2 \leq 14 \sqrt{s_u} \lambda_u / \kappa_{sc, 1}^*, \quad \| \hat{u}^{(k)} - u_0 \|_1 \leq 84 s_u \lambda_u / \kappa_{sc, 1}^*. \] (A.50)

**Conclusion:**

Since Case 1 and Case 2 are the complement of each other, one of them must occur. Thus, the bound of estimation error is controlled by the larger bound in the two cases,

\[ \| \hat{u}^{(k)} - u_0 \|_2 \leq \max \left\{ 14 \sqrt{s_u} \lambda_u / \kappa_{sc, 1}^*, 7 M \sigma^3_{\max} \| u_0 \|_2 \sup_{k' \neq k} \left\| \beta_0 - \beta^{(k, k')} \right\|_2 / \kappa_{sc, 1}^* \right\}, \]

\[ \| \hat{u}^{(k)} - u_0 \|_1 \leq \max \left\{ 84 s_u \lambda_u / \kappa_{sc, 1}^*, 21 M \sigma^6_{\max} \| u_0 \|_2 \sup_{k' \neq k} \left\| \beta_0 - \beta^{(k, k')} \right\|_2 / \kappa_{sc, 1}^* \lambda_u \right\}, \]

which is our oracle inequality.

**Analysis for Terms T_1-T_5 in Part 1**

To show

\[ T_1 = \sqrt{n} \left\{ \mathbf{x}_{\text{std}} - \mathbb{H}(\hat{\beta}) u_0 \right\} (\hat{\beta}^{(k)} - \beta_0) = O_p \left( \sqrt{n} \| \hat{\beta}^{(k)} - \beta_0 \|_2^2 \right), \]

we rewrite the term as a conditional expectation

\[ T_1 = \sqrt{n} u_0^T \left\{ \Sigma_0 - \mathbb{H}(\tilde{\beta}) \right\} (\hat{\beta}^{(k)} - \beta_0) \]
\[ = \sqrt{n} \mathbb{E}_{i \in J_k} \left[ u_i' X_i (\hat{\beta}^{(k)} - \beta_0)^T X_i \{ g' (\hat{\beta}^T X_i) - g' (\beta^{(k)}^T X_i) \} \mid \mathcal{D}_k \right]. \]

Under Assumptions 1b, 2a, we derive the bound for the expectation using the Cauchy-Schwarz inequality and Lemma 17-b,

\[
|T_1| \leq M \sqrt{n} \mathbb{E}_{i \in J_k} \left[ (u_i' X_i)^2 \mid \mathcal{D}_k \right] \frac{\| (\hat{\beta}^{(k)} - \beta_0)^T X_i \|^4}{\mathbb{E}_{i \in J_k} \left[ \{ (\hat{\beta}^{(k)} - \beta_0)^T X_i \}^4 \mid \mathcal{D}_k \right]} \leq M \sqrt{n} \| u_i \|_2 \| \hat{\beta}^{(k)} - \beta_0 \|_2 \sigma_{\max}^3.
\]

(A.51)

Since \( \| u_i \|_2 \) is bounded according to (A.25), we have established in

\[ |T_1| = O_p \left( \sqrt{n} \| \hat{\beta}^{(k)} - \beta_0 \|_2 \right) \]

as declared.

To show

\[ T_2 = \sqrt{n} (u_i - \hat{u}^{(k)})^T \mathbb{H} (\hat{\beta}^{(k)} - \beta_0) = O_p \left( \sqrt{n} \| \hat{\beta}^{(k)} - \beta_0 \|_2 \| \hat{u}^{(k)} - u_i \|_2 \right), \]

we rewrite the term as a conditional expectation

\[ T_2 = \sqrt{n} \mathbb{E}_{i \in J_k} \left[ (u_i - \hat{u}^{(k)})^T X_i (\hat{\beta}^{(k)} - \beta_0)^T X_i \mid \mathcal{D}_k \right]. \]

Similar to (A.51), we derive the bound for the expectation under Assumptions 1b, 2a through the Cauchy-Schwarz inequality and Lemma 17-b, 17-f,

\[ |T_2| \leq M \sqrt{n} \mathbb{E}_{i \in J_k} \left[ \| (u_i - \hat{u}^{(k)})^T X_i (\hat{\beta}^{(k)} - \beta_0)^T X_i \| \mid \mathcal{D}_k \right] \leq 2M \sqrt{n} \| u_i - \hat{u}^{(k)} \| X_i (\hat{\beta}^{(k)} - \beta_0)^T X_i \|_{\mathbb{P}_1} \leq 2M \sqrt{n} \| u_i - \hat{u}^{(k)} \| X_i \| (\hat{\beta}^{(k)} - \beta_0)^T X_i \|_{\mathbb{P}_2} \leq M \sqrt{n} \| \hat{u}^{(k)} - u_i \|_2 \| \hat{\beta}^{(k)} - \beta_0 \|_2 \sigma_{\max}^2. \]

This bound immediately implies

\[ T_2 = O_p \left( \sqrt{n} \| \hat{\beta} - \beta_0 \|_2 \| \hat{u}^{(k)} - u_i \|_2 \right). \]

To show

\[ T_3 = \sqrt{n} \hat{u}^{(k)} \left[ \mathbb{E} \{ \ell^{(k)} (\hat{\beta}^{(k)}; \hat{\gamma}^{(k)}) \mid \mathcal{D}_k \} - \{ \ell^{(k)} (\hat{\beta}^{(k)}; \hat{\gamma}^{(k)}) - \ell^{(k)} (\beta_0; \gamma_0) \} \right] = O_p \left( \rho \| \hat{\beta}^{(k)} - \beta_0 \|_2 + \sqrt{\rho (1 - \rho)} \| \hat{\gamma}^{(k)} - \gamma_0 \|_2 \right), \]

47
we rewrite the term as two empirical processes with diminishing summands

\[
T_3 = -\sqrt{n} \frac{1}{N_k} \sum_{i \in J_k} \left( \hat{u}^{(k)\top} X_i \{g(\hat{\beta}^{(k)\top} X_i) - g(\beta_0^\top X_i) - g(\hat{\gamma}^{(k)\top} W_i) + g(\gamma_0^\top W_i) \} \\
- \mathbb{E}_{i \in \mathcal{J}_k} \left[ \hat{u}^{(k)\top} X_i \{g(\hat{\beta}^{(k)\top} X_i) - g(\beta_0^\top X_i) - g(\hat{\gamma}^{(k)\top} W_i) + g(\gamma_0^\top W_i) \} | \mathcal{D}_k^c \right] \right)
\]

We have used the identity \( \mathbb{E}\{\hat{\ell}^{(k)}(\hat{\beta}_i, \gamma_0) | \mathcal{D}_k^c \} = 0 \) above. Using Lemmas 14, 17-h and Assumptions (2a) and (2a), we show that each summand is sub-exponential

\[
\|\hat{u}^{(k)\top} X_i \{g(\hat{\beta}^{(k)\top} X_i) - g(\beta_0^\top X_i) - g(\hat{\gamma}^{(k)\top} W_i) + g(\gamma_0^\top W_i) \}\|_{\psi_1} \\
\leq \|\hat{u}^{(k)\top} X_i\|_{\psi_2} \|g(\hat{\beta}^{(k)\top} X_i) - g(\beta_0^\top X_i) - g(\hat{\gamma}^{(k)\top} W_i) + g(\gamma_0^\top W_i)\|_{\psi_2} \\
\leq M\sigma_{\max} \|\hat{u}^{(k)}\|_2 (\|\hat{\beta}^{(k)} - \beta_0\|_2 + \|\hat{\gamma}^{(k)} - \gamma_0\|_2), \\
\|\hat{u}^{(k)\top} X_i \{g(\hat{\beta}^{(k)\top} X_i) - g(\beta_0^\top X_i)\}\|_{\psi_1} \leq M\sigma_{\max} \|\hat{u}^{(k)}\|_2 \|\hat{\beta}^{(k)} - \beta_0\|_2/2.
\]

Applying the Bernstein’s inequality, we obtain

\[
T_3 = O_p \left( \|\hat{u}^{(k)}\|_2 \left\{ \sqrt{\rho} \|\hat{\beta}^{(k)} - \beta_0\|_2 + \sqrt{\rho(1 - \rho)} \|\hat{\gamma}^{(k)} - \gamma_0\|_2 \right\} \right).
\]

We achieve the stated rate with the tightness of \( \|\hat{u}^{(k)}\|_2 \) from (A.25).

To show

\[
T_4 = \sqrt{n} (1 - \rho) \hat{u}^{(k)\top} \left[ \mathbb{E}\{\hat{\ell}^{(k)}(\hat{\gamma}^{(k)}) | \mathcal{D}_k^c \} - \{\hat{\ell}^{(k)}(\hat{\gamma}^{(k)}) - \hat{\ell}^{(k)}(\gamma_0) \} \right] \\
= O_p \left( (1 - \rho) \|\hat{\gamma}^{(k)} - \gamma_0\|_2 \right),
\]

we rewrite the term as the empirical process with diminishing summands

\[
T_4 = -\sqrt{n} (1 - \rho) \frac{1}{N_k} \sum_{i \in I_k} \left( \hat{u}^{(k)\top} X_i \{g(\hat{\gamma}^{(k)\top} W_i) - g(\gamma_0^\top W_i) \} \\
- \mathbb{E}_{i \in \mathcal{J}_k} \left[ \hat{u}^{(k)\top} X_i \{g(\hat{\gamma}^{(k)\top} W_i) - g(\gamma_0^\top W_i) \} | \mathcal{D}_k^c \right] \right)
\]

We have used the identity \( \mathbb{E}\{\hat{\ell}^{(k)}(\gamma_0) | \mathcal{D}_k^c \} = 0 \) above. Similar to the analysis of \( T_3 \), we show that each summand is sub-exponential

\[
\|\hat{u}^{(k)\top} X_i \{g(\hat{\gamma}^{(k)\top} W_i) - g(\gamma_0^\top W_i) \}\|_{\psi_1} \leq M\sigma_{\max} \|\hat{u}^{(k)}\|_2 \|\hat{\gamma}^{(k)} - \gamma_0\|_2/2.
\]

Applying the Bernstein’s inequality, we obtain

\[
T_4 = O_p \left( (1 - \rho) \|\hat{u}^{(k)}\|_2 \|\hat{\gamma}^{(k)} - \gamma_0\|_2 \right).
\]
We achieve the stated rate with the tightness of \( \|\hat{u}^{(k)}\|_2 \) from (A.25).

To show

\[
T_5 = \sqrt{n} (u_0 - \hat{u}^{(k)})^T \left\{ g^{(k)}(\beta_0; \gamma_0) + (1 - \rho) \hat{e}^{(k)}_\text{emp}(\gamma_0) \right\} = O_p \left( \|\hat{u}^{(k)} - u_0\|_2 \right),
\]

we rewrite the term as the empirical process with diminishing summands

\[
T_5 = -\sqrt{n} \frac{1}{N_k} \sum_{i \in J_k} (\hat{u}^{(k)} - u_0)^T X_i \{g(\beta_0^T X_i) - g(\gamma_0^T W_i)\}
\]

\[
-\sqrt{n} \frac{1}{n_k} \sum_{i \in I_k} (\hat{u}^{(k)} - u_0)^T X_i \{\rho \cdot g(\beta_0^T X_i) + (1 - \rho) \cdot g(\gamma_0^T W_i) - Y_i\}.
\]

The summands have zero mean because

\[
\mathbb{E}_{i \in J_k} \left[ (\hat{u}^{(k)} - u_0)^T X_i \{g(\beta_0^T X_i) - g(\gamma_0^T W_i)\} \mid \mathcal{D}_k \right] = (\hat{u}^{(k)} - u_0)^T \mathbb{E}[X_i \{g(\beta_0^T X_i) - g(\gamma_0^T W_i)\}] = 0,
\]

\[
\mathbb{E}_{i \in I_k} \left[ (\hat{u}^{(k)} - u_0)^T X_i \{\rho \cdot g(\beta_0^T X_i) + (1 - \rho) \cdot g(\gamma_0^T W_i) - Y_i\} \mid \mathcal{D}_k \right] = (\hat{u}^{(k)} - u_0)^T (\rho \mathbb{E}[X_i \{g(\beta_0^T X_i) - Y_i\}] + (1 - \rho) \mathbb{E}[X_i \{g(\gamma_0^T W_i) - Y_i\}]) = 0.
\]

Similar to the analysis of \( T_3 \), we show that each summand is sub-exponential

\[
\| (\hat{u}^{(k)} - u_0)^T X_i \{g(\beta_0^T X_i) - g(\gamma_0^T W_i)\} \|_{\psi_1} \leq \sqrt{2} \sigma_{\max}(\nu_1 \lor \nu_2) \|\hat{u}^{(k)} - u_0\|_2
\]

\[
\| (\hat{u}^{(k)} - u_0)^T X_i \{\rho \cdot g(\beta_0^T X_i) + (1 - \rho) \cdot g(\gamma_0^T W_i) - Y_i\} \|_{\psi_1} \leq \sqrt{2} \sigma_{\max}(\nu_1 \lor \nu_2) \|\hat{u}^{(k)} - u_0\|_2
\]

Applying the Bernstein’s inequality, we obtain

\[
T_5 = O_p \left( \|\hat{u}^{(k)} - u_0\|_2 \left\{ \sqrt{\rho(1 - \rho)} + 1 \right\} \right) = O_p \left( \|\hat{u}^{(k)} - u_0\|_2 \right).
\]

**Analysis for Terms \( T_1' - T_4' \) in Part 2**

Conditionally on the out-of-fold data, the term \( T_1' \) is the empirical average of i.i.d. mean zero random variables,

\[
T_1 = \sum_{k=1}^K \frac{n_k}{n} \left( \frac{1}{n_k} \sum_{i \in I_k} (\hat{u}^{(k)^T} X_i)^2 \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}^2 
\]

\[
- \mathbb{E}_{i \in I_k} \left[ (\hat{u}^{(k)^T} X_i)^2 \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}^2 \mid \mathcal{D}_k \right] \).
\]

We bound the variance of each summand by the Cauchy-Schwartz inequality and Lemmas 17-b, 17-d, 17-e,

\[
\text{Var}_{i \in I_k} \left[ (\hat{u}^{(k)^T} X_i)^2 \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}^2 \mid \mathcal{D}_k \right]
\]

49
Thus by the Tchebychev’s inequality, we obtain

\[
\E_{i \in I_k} \left[ (\hat{\mu}^{(k)^T} X_i)^4 \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}^4 \mid \mathcal{C}_k \right] \\
\leq \sqrt{\text{Var}_{i \in I_k} \{ (\hat{\mu}^{(k)^T} X_i)^8 \} \mathcal{C}_k} \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}^8 \mid \mathcal{C}_k \\
\leq \sqrt{48} \| \hat{\mu}^{(k)^T} X_i \|_2^8 48 \left( \rho \| g(\hat{\beta}^{(k)^T} X_i) - Y_i \|_2 \vee (1 - \rho) \| g(\hat{\gamma}^{(k)^T} W_i) - Y_i \|_2 \right)^8 \\
\]  

(A.52)

Under Assumption 1a, 1b, 2a, we have

\[
\| \hat{\mu}^{(k)^T} X_i \|_2 \leq (\| u_0 \|_2 + \| \hat{\mu}^{(k)} - u_0 \|_2) \sigma_{\max} / \sqrt{2} = O_p \left( 1 + \| \hat{\mu}^{(k)} - u_0 \|_2 \right).
\]

We apply Lemma 14 to obtain

\[
\| g(\hat{\beta}^{(k)^T} X_i) - Y_i \|_2 = O_p \left( 1 + \| \hat{\beta}^{(k)} - \beta_0 \|_2 \right),
\]

\[
\| g(\hat{\gamma}^{(k)^T} W_i) - Y_i \|_2 = O_p \left( 1 + \| \hat{\gamma}^{(k)} - \gamma_0 \|_2 \right).
\]

We have shown that the variance in (A.52) is of order

\[
O_p \left( 1 + \| \hat{\mu}^{(k)} - u_0 \|_2 + \rho^4 \| \hat{\beta}^{(k)} - \beta_0 \|_2 + (1 - \rho)^4 \| \hat{\gamma}^{(k)} - \gamma_0 \|_2 \right).
\]

Thus by the Tchebychev’s inequality, we obtain

\[
T'_1 = O_p \left( \left\{ 1 + \| \hat{\mu}^{(k)} - u_0 \|_2 + \rho \| \hat{\beta}^{(k)} - \beta_0 \|_2 + (1 - \rho) \| \hat{\gamma}^{(k)} - \gamma_0 \|_2 \right\} / \sqrt{n} \right)
\]

Applying the consistency of \( \hat{\gamma}^{(k)}, \hat{\beta}^{(k)} \) and \( \hat{\mu}^{(k)} \) from (A.22)

\[
T'_1 = O_p \left( n^{-1/2} \right) = o_p(1).
\]

To analyze \( T'_2 \), we consider the decomposition in which the estimators are replaced by the estimands one by one,

\[
T'_2 = \sum_{k=1}^K \sum_{i \in I_k} \frac{n_k}{n} \left( \left( \hat{\mu}^{(k)^T} X_i \right)^2 \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}^2 \mid \mathcal{C}_k \right) \\
- \E \left[ \left( \hat{\mu}^{(k)^T} X_i \right)^2 \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}^2 \right]
\]

\[
= \sum_{k=1}^K \frac{n_k}{n} \left( \hat{\mu}^{(k)^T} X_i \right)^2 \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}^2 \mid \mathcal{C}_k \right) \\
+ \sum_{k=1}^K \frac{n_k}{n} \left( \hat{\mu}^{(k)^T} X_i \right)^2 \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}^2 \mid \mathcal{C}_k \right)
\]

\[
+ \sum_{k=1}^K \frac{n_k}{n} \left( \hat{\mu}^{(k)^T} X_i \right)^2 \{ (1 - \rho) \cdot g(\hat{\gamma}^{(k)^T} W_i) + \rho \cdot g(\hat{\beta}^{(k)^T} X_i) - Y_i \}
\]

50
of Theorem 7, we have established

\[
\frac{\Sigma^2}{\|\Theta\|_2^2} \cdot [\sum_{k=1}^K \frac{n_k}{n} \mathbb{E}_{i \in T_k} \left( (1-\rho) \cdot (\gamma_0^T W_i) + \rho \cdot (\beta_0^T X_i - Y_i) \right)
\]
\[
\times \left[ (1-\rho) \cdot (\gamma_0^T W_i) + \rho \cdot (\beta_0^T X_i - Y_i) \right] | \mathcal{D}_k \]}
\]

Following the same calculation as in (A.52), we can bound the expectations

\[
T_2' = O_p \left( \|\hat{\nu}^{(k)} - \nu_0\|_2 + \rho \|\hat{\beta}^{(k)} - \beta_0\|_2 + (1-\rho)\|\hat{\gamma}^{(k)} - \gamma_0\|_2 \right)
\]

Applying the consistency of \(\hat{\gamma}, \hat{\beta}\) and \(\hat{\nu}\) from Lemma 21, Theorem 1 and Part 1 in the proof of Theorem 7, we have established

\[
T_2' = o_p(1).
\]

Repeating the analyses for \(T_1'\) and \(T_2'\), we can show

\[
T_3' = O_p \left( \rho \sqrt{(1-\rho)/N} \left\{ 1 + \|\hat{\nu}^{(k)} - \nu_0\|_2 + \|\hat{\beta}^{(k)} - \beta_0\|_2 + \|\hat{\gamma}^{(k)} - \gamma_0\|_2 \right\} \right) = o_p(1),
\]

\[
T_4' = O_p \left( \rho (1-\rho) \left\{ \|\hat{\nu}^{(k)} - \nu_0\|_2 + \|\hat{\beta}^{(k)} - \beta_0\|_2 + \|\hat{\gamma}^{(k)} - \gamma_0\|_2 \right\} \right) = o_p(1)
\]

Appendix D. Additional Technical Details

D1 Definitions

We adopt the following definition of sub-Gaussian and sub-exponential random variables.

**Definition 15 (Sub-Gaussian and Sub-Exponential Random Variables)** The sub-Gaussian parameter for a random variable \(V\) is defined as

\[
\|V\|_{\psi_2} = \inf \left\{ \sigma > 0 : \mathbb{E}(e^{V^2/\sigma^2}) \leq 2 \right\}.
\]

The random variable \(V\) is sub-Gaussian if \(\|V\|_{\psi_2}\) is finite. The sub-Gaussian parameter for a random vector \(U\) is defined as

\[
\|U\|_{\psi_2} = \sup_{\|V\|_2 = 1} \|V^T U\|_{\psi_2}.
\]

The sub-Gaussian parameter for a random variable \(V\) is defined as

\[
\|V\|_{\psi_1} = \inf \left\{ \nu > 0 : \mathbb{E}(e^{V/\nu}) \leq 2 \right\}.
\]

The random variable \(V\) is sub-exponential if \(\|V\|_{\psi_1}\) is finite. The more general Orlicz norm for \(\alpha \in (0, 1)\) is defined as

\[
\|V\|_{\psi_\alpha} = \inf \left\{ \nu > 0 : \mathbb{E}(e^{(V/\nu)^\alpha}) \leq 2 \right\}.
\]

Mimicking the (minimal) Restricted Eigenvalue condition on the minimal eigenvalue of matrix over a cone (Bickel et al., 2009), we define the maximal Restricted Eigenvalue in Definition 16.
**Definition 16 (Maximal Restricted Eigenvalue)** For a cone-set of the indices set \( O \subset \{1, \ldots, p\} \)

\[
C_\gamma(\xi, O) := \{v \in \mathbb{R}^{p+q+1} : \|v_O\|_1 \leq \xi \|v_O\|_1\},
\]

we define the maximal Restricted Eigenvalue of a matrix \( \Sigma \) as

\[
\text{RE}_{\text{max}}(\xi, O; \Sigma) = \sup_{v \in C_\gamma(\xi, O) \setminus \{0\}} \frac{\sqrt{v^T \Sigma v}}{\|v\|_2}.
\]

**D2 Statements of Existing Results**

The properties in Lemmas 17 and 18 are covered in Vershynin (2018) Chapter 2 and 4.

**Lemma 17 (Properties of sub-Gaussian and sub-exponential random variables)**

a) **Tail-probability:**

\[
P(|V| \geq x) \leq 2e^{-x/\|V\|_{\psi_1}},
\]

\[
P(|V| \geq x) \leq 2e^{-x^2/\|V\|_{\psi_2}^2};
\]

b) **Moments:** \( \mathbb{E}(|V|^r) \leq \min\{\kappa_{\psi,1} \|V\|_{\psi_1}^r, \kappa_{\psi,2} \|V\|_{\psi_2}^r\} \) with \( \kappa_{\psi,1} = r!2 \) and \( \kappa_{\psi,2} = \Gamma(r/2)r \), and \( \mathbb{E}(|V|) \leq \sqrt{\pi} \|V\|_{\psi_2} \);

c) **Hierarchy:** \( \|V\|_{\psi_1} \leq \|V\|_{\psi_2} \);

d) **Arbitrary addition:** \( \|\sum_{i=1}^m V_i\|_{\psi_2} \leq m \max_{i=1,\ldots,m} \|V_i\|_{\psi_2} \) and \( \|\sum_{i=1}^m V_i\|_{\psi_1} \leq m \max_{i=1,\ldots,m} \|V_i\|_{\psi_1} \);

e) **Multiplication with bounded random variable:** \( \|V_1 V_2\|_{\psi_2} \leq \|V_1\|_{\psi_2} K, \quad \|V_1 V_2\|_{\psi_1} \leq \|V_1\|_{\psi_1} K \) for \( |V_2| \leq K \) almost surely;

f) **Multiplication between sub-Gaussian random variables:** \( \|V_1 V_2\|_{\psi_1} \leq \|V_1\|_{\psi_2} \|V_2\|_{\psi_2} \), in particular, \( \|V_1\|_{\psi_1} \leq \|V_1\|_{\psi_2}/\sqrt{\log(2)} \);

g) **Hoeffding’s inequality:** \( V_1, \ldots, V_m \) are independent mean zero sub-Gaussian random variables. For \( t > 0 \),

\[
P\left( \left| \sum_{i=1}^m V_i \right| \geq t \right) \leq 4 \exp \left( -\frac{t^2}{\kappa_{\psi,3} \sum_{i=1}^m \|V_i\|_{\psi_2}^2} \right), \quad \kappa_{\psi,3} = 8.
\]

h) **Bernstein’s inequality:** \( V_1, \ldots, V_m \) are independent mean zero sub-exponential random variables. For \( t > 0 \), \( \kappa_{\psi,4} = 16 \) and \( \kappa_{\psi,5} = 4 \)

\[
P\left( \left| \sum_{i=1}^m V_i \right| \geq t \right) \leq 2 \exp \left[ -\min \left\{ t^2 \left( \kappa_{\psi,4} \sum_{i=1}^m \|V_i\|_{\psi_1}^2 \right)^{-1}, t \left( \kappa_{\psi,5} \max_{i=1,\ldots,m} \|V_i\|_{\psi_1}^{-1} \right) \right\} \right].
\]
Lemma 18 Let $V_1, \ldots, V_m$ be i.i.d sub-Gaussian vectors in $\mathbb{R}^p$ such that
\[ \|v^TV\|_2^2 \leq K^2 \mathbb{E}\{(v^TV)^2\} \]
for some $1 \leq K < \infty$. Then,
\[ \left\| \frac{1}{m} \sum_{i=1}^{m} V_i V_i^T - \mathbb{E}(VV^T) \right\|_2 = O_p \left( \frac{p}{m} + \sqrt{\frac{p}{m}} \right). \]

From Negahban et al. (2010) and Huang and Zhang (2012) among other literatures, we have the following results concerning the LASSO under the generalized linear models.

Lemma 19 Under Assumptions 1b, 2a and 2b,
\[ \mathbb{P}\left( \ell_{\text{imp}}(\gamma_0 + \Delta) - \ell_{\text{imp}}(\gamma_0) - \Delta^T \ell_{\text{imp}}(\gamma_0) \right) \geq \kappa_{\text{rsc},1} \|\Delta\|_2 \{\|\Delta\|_2 - \kappa_{\text{rsc},2} \sqrt{\log(p+q)/n} \|\Delta\|_1, \forall \|\Delta\|_2 \leq 1\} \geq 1 - \kappa_{\text{rsc},3} e^{-\kappa_{\text{rsc},4}n}; \]
\[ \mathbb{P}\left( \ell_{\text{PL}}(\beta_0 + \Delta) - \ell_{\text{PL}}(\beta_0) - \Delta^T \ell_{\text{PL}}(\beta_0) \right) \geq \kappa_{\text{rsc},1} \|\Delta\|_2 \{\|\Delta\|_2 - \kappa_{\text{rsc},2} \sqrt{\log(p)/N} \|\Delta\|_1, \forall \|\Delta\|_2 \leq 1\} \geq 1 - \kappa_{\text{rsc},3} e^{-\kappa_{\text{rsc},4}N}. \]

The negative log-likelihoods are defined in (8) and (10), and their gradients defined in (12). See Definition 11 for the definition of conditional expectation notation. The constants are all absolute.

The two inequalities in Lemma 19 are direct application of Negahban et al. (2010) Proposition 2 page 22. We can construct an auxiliary loss function to prove the following lemma.

Lemma 20 Under Assumptions 1b, 2a and 2b,
\[ \mathbb{P}\left( \frac{1}{Nk'} \sum_{i \in I_{k'} \cup J_{k'}} g' \left( \hat{\beta}^{(k,k')}^T X_i \right) (\Delta^T X_i)^2 \right) \geq 2 \kappa_{\text{rsc},1} \|\Delta\|_2^2 - \kappa_{\text{rsc},2}^* \sqrt{\log(p)/N} \|\Delta\|_1 \|\Delta\|_2, \forall \|\Delta\|_2 \leq 1 \right) \geq \mathbb{P}\left( \|\hat{\beta}^{(k)} - \beta_0\|_2 \leq \frac{\sigma_{\text{min}}^2}{2\sigma_{\text{max}}^2} \right) - \kappa_{\text{rsc},3}^* e^{-\kappa_{\text{rsc},4}N}. \]

The constants are all absolute.

Proof [Proof of Lemma 20] First, we show $\sqrt{g' \left( \hat{\beta}^{(k,k')}^T X_i \right) X_i}$ is a sub-Gaussian random vector whose second moment has all eigenvalues bounded away from infinity and zero. Under Assumptions 1b and 2a, we may apply Lemma 17-e,
\[ \left\| v^T \sqrt{g' \left( \hat{\beta}^{(k,k')}^T X_i \right) X_i} \right\|_{\psi_2} \leq \sqrt{M} \|v^TX_i\|_{\psi_2} \leq \sqrt{M} \sigma_{\text{max}} \|v\|_2 / \sqrt{2}. \]
Thus, $\sqrt{g'\left(\hat{\beta}^{(k,k')\top}X_i\right)}X_i$ is a sub-Gaussian random vector. Under Assumptions 1b and 2a, we can bound the maximal eigenvalue of its second moment,

$$v^T\mathbb{E}_{i \in I_{k'} \cup J_k'} \left\{ g'\left(\hat{\beta}^{(k,k')\top}X_i\right)X_iX_i^\top | \mathcal{D}_{k'}^c \right\} v \leq M\mathbb{E}\{v'^2\} \leq M\|v\|^2_2\sigma_{\text{max}}^2.$$

We derive the lower bound for the minimal eigenvalue of its second moment from Assumptions 1b, 2a, 2a, 2b-i, the Cauchy-Schwartz inequality and Lemma 17-b.

$$v^T\mathbb{E}_{i \in I_{k'} \cup J_k'} \left\{ g'\left(\beta_0^{(k,k')}X_i\right)X_iX_i^\top | \mathcal{D}_{k'}^c \right\} v - \mathbb{E}_{i \in I_{k'} \cup J_k'} \left[ (v'^2X_i)^2 \left\{ g(\beta_0^{(k,k')}X_i) - g(\hat{\beta}^{(k,k')\top}X_i) \right\} \right] \geq \|v\|^2_2\sigma_{\text{min}}^2 - M\mathbb{E}_{i \in I_{k'} \cup J_k'} \left[ (v'^2X_i)^2 \left\{ (\beta_0^{(k,k')} - \hat{\beta}^{(k,k')})^\top X_i \right\} \right] \geq \|v\|^2_2\left(\sigma_{\text{min}}^2 - M\sigma_{\text{max}}^3\|\hat{\beta}^{(k,k')} - \beta_0\|^2_2\right).$$

Whenever $\left\|\hat{\beta}^{(k)} - \beta_0\right\|_2 \leq \frac{\sigma_{\text{min}}^2}{2\sigma_{\text{max}}^3}$, we have

$$v^T\mathbb{E}_{i \in I_{k'} \cup J_k'} \left\{ g'\left(\hat{\beta}^{(k,k')\top}X_i\right)X_iX_i^\top | \mathcal{D}_{k'}^c \right\} v \geq \|v\|^2_2\sigma_{\text{min}}^2/2.$$

Second, we construct an auxiliary least square loss to apply Negahban et al. (2010). Let $\varepsilon_i$ be independent standard normal random variables. Construct the loss function

$$\mathcal{L}^{(k,k')}(v) = \frac{1}{N_{k'}} \sum_{i \in I_{k'} \cup J_k'} \left\{ \varepsilon_i + (v_0 - v)^\top \sqrt{g'\left(\hat{\beta}^{(k,k')\top}X_i\right)}X_i \right\}^2.$$

By the design, we have

$$\mathcal{L}^{(k,k')}(v_0 + \Delta) - \mathcal{L}^{(k,k')}(v_0) - \Delta^\top \frac{\partial}{\partial v} \mathcal{L}^{(k,k')}(v_0 + \Delta) = \frac{1}{N_{k'}} \sum_{i \in I_{k'} \cup J_k'} g'\left(\hat{\beta}^{(k,k')\top}X_i\right) \left(\Delta^\top X_i\right)^2.$$

We apply Proposition 2 in Negahban et al. (2010) for $\mathcal{L}^{(k,k')}(v)$ conditionally on out-of-fold data $\mathcal{D}_{k'}^c$ and the event $\left\{ \left\|\hat{\beta}^{(k,k')} - \beta_0\right\|_2 \leq \frac{\sigma_{\text{min}}^2}{2\sigma_{\text{max}}^3}\right\}$ to finish the proof.

Lemma 21 For a constant $\kappa_{\text{cone}}(n,p,q,\varepsilon_r) \approx \sqrt{s_n \log(p + q)/n}$, the event

$$\Omega_{\text{cone}} = \left\{ \left\|\hat{\ell}_{\text{imp}}(\gamma_0)\right\|_{\infty} = \left\| \frac{1}{n} \sum_{i=1}^n W_i \{ g(\gamma_0^\top W_i) - Y_i \} \right\|_{\infty} \leq \kappa_{\text{cone}}(n,p,q,\varepsilon_r) \right\}$$

54
occur with probability greater than $1 - \varepsilon_r$ under Assumptions 1a and 2a. Setting $\lambda_\gamma = 2$, we have on event $\Omega_{\text{cone}}$ that

$$\hat{\gamma} - \gamma_0 \in \mathcal{C}_\gamma(3, \text{supp}(\gamma_0)) = \left\{ v \in \mathbb{R}^{p+q+1} : \|v_{\mathcal{O}_\gamma}\|_1 \leq 3\|v_{\mathcal{O}_\gamma}\|_1 \right\},$$

where $\mathcal{O}_\gamma = \{ j : \gamma_j \neq 0 \}$ is the indices set for nonzero coefficient in $\gamma_0$. Moreover, we have

$$\|\hat{\gamma} - \gamma_0\|_2 = O_p \left( \sqrt{s_\gamma \log(p + q)/n} \right).$$

The concentration on the event $\Omega_{\text{cone}}$ is established by the union bound of element wise concentration, which is in turn obtained by the Bernstein inequality for sub-exponential random variables (Lemma 17-h). The rest of Lemma 21 follows Huang and Zhang (2012) Lemma 1 page 5 (page 1843 of the issue) and Negahban et al. (2010) Corollary 5 page 23.