LEVERAGE CLASSIFIER: ANOTHER LOOK AT SUPPORT VECTOR MACHINE

Yixin Han[1], Jun Yu[2], Nan Zhang[3], Cheng Meng[4], Ping Ma[5], Wenxuan Zhong[5], and Changliang Zou[1]

[1]*School of Statistics and Data Science, LPMC & KLMDASR, Nankai University, Tianjin, P.R. China*
[2]*School of Mathematics and Statistics, Beijing Institute of Technology, Beijing, P.R.China*
[3]*School of Data Science, Fudan University, Shanghai, P.R.China*
[4]*Institute of Statistics and Big Data, Renmin University, Beijing, P.R.China*
[5]*Department of Statistics, University of Georgia, Athens, GA, USA*

Supplementary Material

This supplementary material contains the proofs of technical results and some additional simulation results.

Appendix A: Useful Lemma

The following Lemma is a multivariate extension of the martingale central limit theorem, see Lemma 4 in Zhang et al. (2021) for details.

Lemma S.1 (Multivariate version of martingale CLT). *Let $\{\boldsymbol{\eta}_{ki}, i = 1, \ldots, N_k\}$ be a martingale difference sequence in $\mathbb{R}^p$ relative to the filtration $\{\mathcal{F}_{ki}, i = 0, 1, \ldots, N_k\}$ and let $\boldsymbol{Z}_k \in \mathbb{R}^p$ be an $\mathcal{F}_{k0}$-measurable random vector for $k = 1, 2, 3, \ldots$. Denote $\boldsymbol{R}_k = \sum_{i=1}^{N_k} \boldsymbol{\eta}_{ki}$. Assume the following conditions hold.*

(i) $\lim_{k\to\infty} \sum_{i=1}^{N_k} \mathbb{E}\left(\|\boldsymbol{\eta}_{ki}\|^4\right) = 0$.

(ii) $\lim_{k\to\infty} \mathbb{E}\left\{\| \sum_{i=1}^{N_k} \mathbb{E}\left(\boldsymbol{\eta}_{ki}\boldsymbol{\eta}_{ki}^\top \mid \mathcal{F}_{k,i-1}\right) - \mathbf{B}_k\|^2\right\} = 0$ *for some sequence of positive-definite matrices* $\{\mathbf{B}_k\}_{k=1}^{\infty}$ *with* $\sup_k \lambda_{\max}(\mathbf{B}_k) < \infty$, *say that the largest eigenvalue is uniformly bounded.*

(iii) For a probability distribution $\boldsymbol{L}_0$, $*$ denotes convolution and $\boldsymbol{L}(\cdot)$ denotes the law of random variables, $\boldsymbol{L}(\boldsymbol{Z}_k)*\mathcal{N}(\mathbf{0},\mathrm{B}_k)\to\boldsymbol{L}_0$, where the convergence is in distribution.

Then we have

$$\boldsymbol{L}(\boldsymbol{Z}_k+\boldsymbol{R}_k)\to\boldsymbol{L}_0.$$

Appendix B: Proof of Theorem 1

Proof. Denote

$$L_n(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{N\pi_i^*}\left[1-Y_i^*f(\boldsymbol{X}_i^*,\boldsymbol{\beta})\right]_+, L_N(\boldsymbol{\beta})=\frac{1}{N}\sum_{j=1}^{N}\left[1-Y_jf(\boldsymbol{X}_j,\boldsymbol{\beta})\right]_+,$$

$$l_{\lambda,n}(\boldsymbol{\beta})=\frac{1}{n}\sum_{i=1}^{n}\frac{1}{N\pi_i^*}\left[1-Y_i^*f(\boldsymbol{X}_i^*,\boldsymbol{\beta})\right]_++\frac{\lambda}{2}\|\boldsymbol{\beta}_1\|^2.$$

The proof can be divided into the following intermediate parts.

First, we consider the influence of a fixed λ. For a fixed $\boldsymbol{\theta}=(1,\boldsymbol{\theta}_1^\top)^\top\in\mathbb{R}^{p+1}$, define

$$\Lambda_n(\boldsymbol{\theta})=n\left\{l_{\lambda,n}\left(\boldsymbol{\beta}^\dagger+\frac{\boldsymbol{\theta}}{\sqrt{n}}\right)-l_{\lambda,n}\left(\boldsymbol{\beta}^\dagger\right)\right\},\quad T_n(\boldsymbol{\theta})=\mathbb{E}\left\{\Lambda_n(\boldsymbol{\theta})\right\}.$$

Observe that

$$\begin{aligned}\Lambda_n(\boldsymbol{\theta})=&\sum_{i=1}^{n}\frac{1}{N\pi_i^*}\left\{\left[1-Y_i^*f\left(\boldsymbol{X}_i^*,\boldsymbol{\beta}^\dagger+\frac{\boldsymbol{\theta}}{\sqrt{n}}\right)\right]_+-\left[1-Y_i^*f(\boldsymbol{X}_i^*,\boldsymbol{\beta}^\dagger)\right]_+\right\}\\&+n\frac{\lambda}{2}\left(\|\boldsymbol{\beta}_1^\dagger+\frac{\boldsymbol{\theta}_1}{\sqrt{n}}\|^2-\|\boldsymbol{\beta}_1^\dagger\|^2\right),\end{aligned}$$

and $\mathbb{E}\left\{L_n(\boldsymbol{\beta})\right\}=\mathbb{E}\left[\mathbb{E}\left\{L_n(\boldsymbol{\beta})\mid\mathcal{D}_N\right\}\right]=L(\boldsymbol{\beta})=\mathbb{E}\left[1-Yf(\boldsymbol{X},\boldsymbol{\beta})\right]_+$. Under Assumption 3, we assume $\boldsymbol{\beta}_1^\dagger\neq0$ without loss of generality. By Lemma 3 in Koo et al. (2008), we have

$$T_n(\boldsymbol{\theta}) = n\left\{L\left(\boldsymbol{\beta}^{\dagger} + \frac{\boldsymbol{\theta}}{\sqrt{n}}\right) - L(\boldsymbol{\beta}^{\dagger})\right\} + \frac{\lambda}{2}\left(\|\boldsymbol{\theta}_1\|^2 + 2\sqrt{n}\boldsymbol{\theta}_1^{\top}\boldsymbol{\beta}_1^{\dagger}\right),$$
$$= \frac{1}{2}\boldsymbol{\theta}^{\top}\mathbf{H}(\breve{\boldsymbol{\beta}})\boldsymbol{\theta} + \frac{\lambda}{2}\left(\|\boldsymbol{\theta}_1\|^2 + 2\sqrt{n}\boldsymbol{\theta}_1^{\top}\boldsymbol{\beta}_1^{\dagger}\right),$$

by applying Taylor expansion of $L(\boldsymbol{\beta})$ around $\boldsymbol{\beta}^{\dagger}$, where $\breve{\boldsymbol{\beta}} = \boldsymbol{\beta}^{\dagger} + (\boldsymbol{\theta}/\sqrt{n})t$ for some $0 < t < 1$.

Define $\mathbf{D}_{ij}(\boldsymbol{\alpha}) = \mathbf{H}(\boldsymbol{\beta}^{\dagger}+\boldsymbol{\alpha})_{ij} - \mathbf{H}(\boldsymbol{\beta}^{\dagger})_{ij}$ for $0 \le i, j \le p+1$. By Assumption 1, $\mathbf{H}(\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$. Then, for any $\varepsilon_1 > 0$, there exist $\delta_1 > 0$ such that $\mathbf{D}_{ij}(\boldsymbol{\alpha}) < \varepsilon_1$ if $\|\boldsymbol{\alpha}\| < \delta_1$ for all $0 \le i, j \le p+1$. Thus, for sufficiently large n such that $\|(\boldsymbol{\theta}/\sqrt{n})t\| < \delta_1$

$$\left|\boldsymbol{\theta}^{\top}\left(\mathbf{H}(\breve{\boldsymbol{\beta}}) - \mathbf{H}(\boldsymbol{\beta}^{\dagger})\right)\boldsymbol{\theta}\right| \le \sum_{i,j} |\boldsymbol{\theta}_i||\boldsymbol{\theta}_j| \left|\mathbf{D}_{ij}\left(\frac{\boldsymbol{\theta}}{\sqrt{n}}t\right)\right| \le 2\varepsilon_1\|\boldsymbol{\theta}\|^2,$$

then $\boldsymbol{\theta}^{\top}\mathbf{H}(\breve{\boldsymbol{\beta}})\boldsymbol{\theta}/2 = \boldsymbol{\theta}^{\top}\mathbf{H}(\boldsymbol{\beta}^{\dagger})\boldsymbol{\theta}/2 + o(1)$ as $n \to \infty$. Combining the assumption that $\lambda = o(n^{-1/2})$, we have

$$T_n(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^{\top}\mathbf{H}(\boldsymbol{\beta}^{\dagger})\boldsymbol{\theta} + o(1).$$

Next, we would like to provide an expansion of $\Lambda_n(\boldsymbol{\theta})$ under Assumptions 1–3. Let $\boldsymbol{W}_n = -n^{-1}\sum_{i=1}^{n}(N\pi_i^*)^{-1}\xi_i^* Y_i^* \widetilde{\boldsymbol{X}}_i^*$, where $\xi_i^* = \mathbb{I}\left(Y_i^* f(\boldsymbol{X}_i^*, \boldsymbol{\beta}^{\dagger}) \le 1\right)$. If we define

$$R_{i,n}(\boldsymbol{\theta}) = \frac{1}{N\pi_i^*}\left\{\left[1 - Y_i^* f\left(\boldsymbol{X}_i^*, \boldsymbol{\beta}^{\dagger} + \frac{\boldsymbol{\theta}}{\sqrt{n}}t\right)\right]_+ - \left[1 - Y_i^* f\left(\boldsymbol{X}_i^*, \boldsymbol{\beta}^{\dagger}\right)\right]_+ + \xi_i^* Y_i^* f\left(\boldsymbol{X}_i^*, \frac{\boldsymbol{\theta}}{\sqrt{n}}\right)\right\},$$
$$R_{j,N}(\boldsymbol{\theta}) = \left[1 - Y_j f\left(\boldsymbol{X}_j, \boldsymbol{\beta}^{\dagger} + \frac{\boldsymbol{\theta}}{\sqrt{n}}t\right)\right]_+ - \left[1 - Y_j f\left(\boldsymbol{X}_j, \boldsymbol{\beta}^{\dagger}\right)\right]_+ + \xi_j Y_j f\left(\boldsymbol{X}_j, \frac{\boldsymbol{\theta}}{\sqrt{n}}\right),$$

where $i = 1, \ldots, n$ and $j = 1, \ldots, N$. Recall that $\mathbb{E}\{(N\pi_i^*)^{-1}\xi_i^* Y_i^* \widetilde{\boldsymbol{X}}_i^*\} = \boldsymbol{S}(\boldsymbol{\beta}^{\dagger}) = 0$.

Recall the definitions of $T_n(\boldsymbol{\theta})$ and $\boldsymbol{W}_n$, we have

$$\begin{aligned}\Lambda_n(\boldsymbol{\theta}) =& \sum_{i=1}^{n} \frac{1}{N\pi_i^*}\left[1 - Y_i^* f\left(\boldsymbol{X}_i^*, \boldsymbol{\beta}^\dagger + \frac{\boldsymbol{\theta}}{\sqrt{n}}\right)\right]_+ - nL\left(\boldsymbol{\beta}^\dagger + \frac{\boldsymbol{\theta}}{\sqrt{n}}\right) \\ &- \sum_{i=1}^{n} \frac{1}{N\pi_i^*}\left[1 - Y_i^* f\left(\boldsymbol{X}_i^*, \boldsymbol{\beta}^\dagger\right)\right]_+ + nL\left(\boldsymbol{\beta}^\dagger\right) + \frac{\lambda}{2}\left(\|\boldsymbol{\theta}_1\|^2 + 2\sqrt{n}\boldsymbol{\theta}_1^\top \boldsymbol{\beta}_1^\dagger\right) \\ &+ \sum_{i=1}^{n} \frac{1}{N\pi_i^*}\xi_i^* Y_i^* (\widetilde{\boldsymbol{X}}_i^*)^\top \frac{\boldsymbol{\theta}}{\sqrt{n}} - \sum_{i=1}^{n} \frac{1}{N\pi_i^*}\xi_i^* Y_i^* (\widetilde{\boldsymbol{X}}_i^*)^\top \frac{\boldsymbol{\theta}}{\sqrt{n}} \\ =& T_n(\boldsymbol{\theta}) + \sqrt{n}\boldsymbol{W}_n^\top \boldsymbol{\theta} + \sum_{i=1}^{n}\left[R_{i,n}(\boldsymbol{\theta}) - \mathbb{E}\left\{R_{i,n}(\boldsymbol{\theta})\right\}\right]. \end{aligned} \tag{S.1}$$

Recall that $[\cdot]_+$ denotes the hinge loss. We define $\varphi = \mathbb{I}\left(a \le 1\right)$ and $D = [1 - z]_+ - [1 - a]_+ + \varphi(z - a)$. Then we have

$$\begin{aligned} D &= (1 - z)\mathbb{I}(a > 1, z \le 1) + (z - 1)\mathbb{I}(a < 1, z > 1) \\ &\le |z - a|\,\mathbb{I}(a > 1, z \le 1) + |z - a|\,\mathbb{I}(a < 1, z > 1) \\ &= |z - a|\left\{\mathbb{I}(a > 1, z \le 1) + \mathbb{I}(a < 1, z > 1)\right\} \\ &\le |z - a|\,\mathbb{I}\left(|1 - a| \le |z - a|\right). \end{aligned} \tag{S.2}$$

Let $z_i = Y_i^* f(\boldsymbol{X}_i^*, \boldsymbol{\beta}^\dagger + \boldsymbol{\theta}/\sqrt{n})$ and $a_i = Y_i^* f(\boldsymbol{X}_i^*, \boldsymbol{\beta}^\dagger)$ in (S.2), we have

$$|R_{i,n}(\boldsymbol{\theta})| \le \frac{1}{N\pi_i^*}\left|\frac{f(\boldsymbol{X}_i^*, \boldsymbol{\theta})}{\sqrt{n}}\right| U_i\left(\left|\frac{f(\boldsymbol{X}_i^*, \boldsymbol{\theta})}{\sqrt{n}}\right|\right), \tag{S.3}$$

where $U_i(t) = \mathbb{I}\left(\left|1 - Y_i^* f(\boldsymbol{X}_i^*, \boldsymbol{\beta}^\dagger)\right| \le t\right)$ with respect to the i-th subsample point for

$t \in \mathbb{R}$. By (S.3), for each fixed $\boldsymbol{\theta}$ we obtain

$$\begin{aligned}\mathbb{E}\left[\sum_{i=1}^{n}\left\{R_{i,n}(\boldsymbol{\theta})-\mathbb{E}\left(R_{i,n}(\boldsymbol{\theta})\right)\right\}\right]^2 &= \mathbb{E}\left\{\mathbb{E}\left[\sum_{i=1}^{n}\left\{R_{i,n}(\boldsymbol{\theta})-\mathbb{E}\left(R_{i,n}(\boldsymbol{\theta})\right)\right\}\right]^2 \middle| \mathcal{D}_N\right\} \\ &= \frac{n}{N^2}\sum_{j=1}^{N}\mathbb{E}\left[\frac{1}{\pi_j}\left\{R_{j,N}(\boldsymbol{\theta})-\mathbb{E}\left(R_{j,N}(\boldsymbol{\theta})\right)\right\}^2\right] \\ &\leq \frac{n}{N^2}\sum_{j=1}^{N}\mathbb{E}\left\{\frac{1}{\pi_j}R_{i,N}^2(\boldsymbol{\theta})\right\} \\ &\leq \frac{n}{N^2}\sum_{j=1}^{N}\mathbb{E}\left\{\frac{1}{\pi_j}\left(1+\|\boldsymbol{X}_j\|^2\right)\frac{\|\boldsymbol{\theta}\|^2}{n}U_j\left(\sqrt{1+\|\boldsymbol{X}_j\|^2}\frac{\|\boldsymbol{\theta}\|}{\sqrt{n}}\right)\right\} \\ &\leq \frac{\|\boldsymbol{\theta}\|^2}{N^2}\sum_{j=1}^{N}\mathbb{E}\left\{\frac{1}{\pi_j}\left(1+\|\boldsymbol{X}_j\|^2\right)U_j\left(\sqrt{1+\|\boldsymbol{X}_j\|^2}\frac{\|\boldsymbol{\theta}\|}{\sqrt{n}}\right)\right\}.\end{aligned}$$

By Assumption 1 implies that $\mathbb{E}(\|\boldsymbol{X}\|^4) < \infty$, there exists c_1 such that

$$\mathbb{E}\left\{(1+\|\boldsymbol{X}\|^4)\mathbb{I}\left(\|\boldsymbol{X}\| > c_1\right)\right\} < \varepsilon_2/2,$$

for any $\varepsilon_2 > 0$. Let $U(t) = \mathbb{I}\left(\left|1 - Yf(\boldsymbol{X}, \boldsymbol{\beta}^{\dagger})\right| \leq t\right)$ for $t \in \mathbb{R}$. By Assumption 4 and holder inequality, we have

$$\begin{aligned}&\frac{1}{N^2}\sum_{j=1}^{N}\mathbb{E}\left\{\frac{1}{\pi_j}\left(1+\|\boldsymbol{X}_j\|^2\right)U_j\left(\sqrt{1+\|\boldsymbol{X}_j\|^2}\frac{\|\boldsymbol{\theta}\|}{\sqrt{n}}\right)\right\} \\ \leq& \frac{1}{N^2}\sum_{j=1}^{N}\mathbb{E}\left\{\frac{1}{\pi_j}\left(1+\|\boldsymbol{X}_j\|^2\right)\mathbb{I}\left(\|\boldsymbol{X}_j\| > c_1\right)\right\} + \frac{1}{N^2}\sum_{j=1}^{N}\mathbb{E}\left\{\frac{1+c_1^2}{\pi_j}U\left(\sqrt{1+c_1^2}\frac{\|\boldsymbol{\theta}\|}{\sqrt{n}}\right)\right\} \\ \leq& \sqrt{\mathbb{E}\left(\frac{1}{N^3}\sum_{j=1}^{N}\frac{1}{\pi_j^2}\right)}\sqrt{\mathbb{E}\left\{\frac{1}{N}\sum_{j=1}^{N}\left(1+\|\boldsymbol{X}_j\|^2\right)^2\mathbb{I}\left(\|\boldsymbol{X}_j\| > c_1\right)\right\}} \\ &+ (1+c_1^2)\sqrt{\mathbb{E}\left(\frac{1}{N^3}\sum_{j=1}^{N}\frac{1}{\pi_j^2}\right)}\sqrt{\frac{1}{N}\sum_{j=1}^{N}\mathrm{P}\left\{U\left(\sqrt{1+c_1^2}\|\boldsymbol{\theta}\|/\sqrt{n}\right) = 1\right\}},\end{aligned}$$

By Assumption 1, the conditional distribution of $\boldsymbol{X}$ given Y is not degenerate, which implies $\lim_{t\to 0}\mathrm{P}\left(U(t) = 1\right) = 0$. We can take a large c_2 such that

$$\mathrm{P}\left\{U\left(\sqrt{1+c_1^2}\|\boldsymbol{\theta}\|/\sqrt{n}\right) = 1\right\} < \varepsilon_2/\left\{2(1+c_1^2)\right\},$$

for $n > c_2$. By Assumption 4, it proves that $\mathbb{E}\left[\sum_{i=1}^n \left\{R_{i,n}(\boldsymbol{\theta}) - \mathbb{E}\left(R_{i,n}(\boldsymbol{\theta})\right)\right\}\right]^2 \to 0$.

By (S.1), for each fixed $\boldsymbol{\theta}$

$$\Lambda_n(\boldsymbol{\theta}) = \frac{1}{2}\boldsymbol{\theta}^\top \mathbf{H}(\boldsymbol{\beta}^\dagger)\boldsymbol{\theta} + \sqrt{n}\boldsymbol{W}_n{}^\top\boldsymbol{\theta} + o_P(1).$$

Last, we devote to giving the Bahadur representation of $\widetilde{\boldsymbol{\beta}}$. Let $\boldsymbol{\kappa}_n = -\sqrt{n}\mathbf{H}(\boldsymbol{\beta}^\dagger)^{-1}\boldsymbol{W}_n$ and $\boldsymbol{\Theta}$ be a convex open subset in $\mathbb{R}^{p+1}$. By Convexity Lemma in Pollard (1991), we have

$$\Lambda_n(\boldsymbol{\theta}) = \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\kappa}_n\right)^\top \mathbf{H}(\boldsymbol{\beta}^\dagger)\left(\boldsymbol{\theta} - \boldsymbol{\kappa}_n\right) - \frac{1}{2}\boldsymbol{\kappa}_n^\top \mathbf{H}(\boldsymbol{\beta}^\dagger)\boldsymbol{\kappa}_n + r_n(\boldsymbol{\theta}),$$

where for each compact set K of $\boldsymbol{\Theta}$, the aforementioned part is shown for every $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and then we have $\sup_{\boldsymbol{\theta}\in K}|r_n(\boldsymbol{\theta})| \to 0$ in probability. Lemma S.4 shows that $\boldsymbol{\kappa}_n$ is asymptotically normal which will be proved in the next section, then there exists a compact set $K \in \mathcal{B}_\rho$ with probability close to one, where $\mathcal{B}_\rho$ is a closed ball with center $\boldsymbol{\kappa}_n$ and radius ρ. Let $\Delta_n = \sup_{\boldsymbol{\theta}\in\mathcal{B}_\rho}|r_n(\boldsymbol{\theta})|$. Then we have

$$\Delta_n \to 0 \ \text{ in probability.} \tag{S.4}$$

Next, we discuss the behavior of $\Lambda_n(\boldsymbol{\theta})$ outside the closed ball $\mathcal{B}_\rho$. Consider $\boldsymbol{\theta} = \boldsymbol{\kappa}_n + \gamma\boldsymbol{e}$, with $\gamma > \rho$ and the unit vector $\boldsymbol{e}$. A boundary point $\boldsymbol{\theta}^\dagger = \boldsymbol{\kappa}_n + \rho\boldsymbol{e}$. Under Assumptions 1–3 and a similar discussion in Lemma 5 of Koo et al. (2008), there exists a constant c_3 such that $\boldsymbol{\beta}^\top\mathbf{H}(\boldsymbol{\beta}^\dagger)\boldsymbol{\beta} \geq c_3\|\boldsymbol{\beta}\|^2$. Then, by the convexity of $\Lambda_n(\boldsymbol{\theta})$ and the definition of Δ_n, we have

$$\begin{aligned}
\frac{\rho}{\gamma}\Lambda_n(\boldsymbol{\theta}) + \left(1 - \frac{\rho}{\gamma}\right)\Lambda_n(\boldsymbol{\kappa}_n) &\geq \Lambda_n\left(\frac{\rho}{\boldsymbol{\gamma}}\boldsymbol{\theta} + \left(1 - \frac{\rho}{\gamma}\right)\boldsymbol{\kappa}_n\right) \\
&= \Lambda_n(\boldsymbol{\theta}^{\dagger}) \\
&\geq \frac{1}{2}\left(\boldsymbol{\theta} - \boldsymbol{\kappa}_n\right)^{\top}\mathbf{H}(\boldsymbol{\beta}^{\dagger})\left(\boldsymbol{\theta} - \boldsymbol{\kappa}_n\right) - \frac{1}{2}\kappa_n^{\top}\mathbf{H}(\boldsymbol{\beta}^{\dagger})\boldsymbol{\kappa}_n - \Delta_n \\
&\geq \frac{c_3}{2}\rho^2 + \Lambda_n(\boldsymbol{\kappa}_n) - 2\Delta_n,
\end{aligned}$$

which implies that

$$\inf_{\|\boldsymbol{\theta}-\boldsymbol{\kappa}_n\|>\rho}\Lambda_n(\boldsymbol{\theta}) \geq \Lambda_n(\boldsymbol{\kappa}_n) + \left(\frac{c_3}{2}\rho^2 - 2\Delta_n\right).$$

By (S.4), we can take Δ_n such that $2\Delta_n < c_3\rho^2/2$ with probability tending to one. Thus $\inf_{\|\boldsymbol{\theta}-\boldsymbol{\kappa}_n\|>\rho}\Lambda_n(\boldsymbol{\theta}) \geq \Lambda_n(\boldsymbol{\kappa}_n)$. This implies the minimum of $\Lambda_n(\boldsymbol{\theta})$ cannot occur at any $\boldsymbol{\theta}$ with $\|\boldsymbol{\theta} - \boldsymbol{\kappa}_n\| > \rho$. Hence for each $\rho > 0$ and let $\widetilde{\boldsymbol{\theta}}_n = \sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\dagger})$, we have $\mathrm{P}(\|\widetilde{\boldsymbol{\theta}}_n - \boldsymbol{\kappa}_n\| > \rho) \to 0$. Thus

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\dagger}) = -\sqrt{n}\mathbf{H}(\boldsymbol{\beta}^{\dagger})^{-1}\boldsymbol{W}_n + o_P\left(1\right).$$

The theorem follows the above arguments. □

Appendix C: Proof of asymptotic normality

Recall that

$$\boldsymbol{M} = \sum_{i=1}^{n}\boldsymbol{M}_i = \sum_{i=1}^{n}\frac{1}{nN\pi_i^*}\xi_i^* Y_i^*\widetilde{\boldsymbol{X}}_i^* - \sum_{i=1}^{n}\left(\frac{1}{nN}\sum_{j=1}^{N}\xi_j Y_j\widetilde{\boldsymbol{X}}_j\right), \tag{S.5}$$

$$\boldsymbol{Q} = \frac{1}{N}\sum_{j=1}^{N}\xi_j Y_j\widetilde{\boldsymbol{X}}_j,\quad \boldsymbol{T} = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{N\pi_i^*}\xi_i^* Y_i^*\widetilde{\boldsymbol{X}}_i^*,\quad \mathbf{B}_N = \mathbf{V}_T^{-1/2}\mathbf{V}_M\mathbf{V}_T^{-1/2},$$

where $\mathbf{V}_T$ and $\mathbf{V}_M$ are the variances of $\boldsymbol{T}$ and $\boldsymbol{M}$.

Lemma S.2. *$\{\boldsymbol{M}_i, i = 1, \ldots, n\}$ in (S.5) is a martingale difference sequence relative to the filtration $\{\mathcal{F}_{N,i}, i = 1, \ldots, n\}$.*

Proof. The $\mathcal{F}_{n,i}$-measurability follows from the definition of $\boldsymbol{M}_i$ and the definition of the filtration $\{\mathcal{F}_{N,i}, i = 1, \ldots, n\}$. Moreover, we have

$$\begin{aligned}\mathbb{E}\left\{\boldsymbol{M}_i \mid \mathcal{F}_{N,i-1}\right\} &= \mathbb{E}_{Y|\boldsymbol{X}}\left\{\frac{1}{nN\pi_i^*}\xi_i^* Y_i^* \widetilde{\boldsymbol{X}}_i^*\right\} - \frac{1}{nN}\sum_{j=1}^{N}\xi_j Y_j \widetilde{\boldsymbol{X}}_j \\ &= \frac{1}{nN}\sum_{i=1}^{N}\xi_i Y_i \widetilde{\boldsymbol{X}}_i - \frac{1}{nN}\sum_{j=1}^{N}\xi_j Y_j \widetilde{\boldsymbol{X}}_j \\ &= 0,\end{aligned}$$

where $\mathbb{E}_{Y|\boldsymbol{X}}$ is the expectation with respect to sampling randomness or the conditional expectation of Y given $\boldsymbol{X}_1^N$ with $\boldsymbol{X}_1^N = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$. Then $\{\boldsymbol{M}_i, i = 1, \ldots, n\}$ is a martingale difference sequence. □

Lemma S.3. *Suppose Assumptions 1 and 4 hold. Let $\mathbf{V}_T$ and $\mathbf{V}_Q$ denote the variances of $\boldsymbol{T}$ and $\boldsymbol{Q}$. For any $\boldsymbol{t} \in \mathbb{R}^{p+1}$, we have*

$$\left|\mathbb{E}\left\{\exp\left(i\boldsymbol{t}^\top \mathbf{V}_T^{-1/2}\boldsymbol{Q}\right)\right\} - \mathbb{E}\left\{\exp\left(i\boldsymbol{t}^\top \mathbf{V}_T^{-1/2}\mathbf{V}_Q^{1/2}\boldsymbol{A}_0\right)\right\}\right| \to 0,$$

as $N \to \infty$, where $\boldsymbol{A}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+1})$.

Proof. Note $\boldsymbol{Q}$ is a sum of i.i.d mean zero random vectors, $\xi_j Y_j \widetilde{\boldsymbol{X}}_j$. The Linderberg-Feller conditions are satisfied by Assumption 1 and Assumption 4, then we have

$$\mathbf{V}_Q^{-1/2}\boldsymbol{Q} \to \mathcal{N}\left(\mathbf{0}, \mathbf{I}_{p+1}\right). \tag{S.6}$$

Furthermore, for any $\boldsymbol{\varsigma} \in \mathbb{R}^{p+1}$ and as $N \to \infty$

$$\left|\mathbb{E}\left\{\exp\left(i\boldsymbol{\varsigma}^\top \mathbf{V}_Q^{-1/2}\boldsymbol{Q}\right)\right\} - \mathbb{E}\left\{\exp\left(i\boldsymbol{\varsigma}^\top \mathbf{A}_0\right)\right\}\right| \to 0.$$

Let $\boldsymbol{\varsigma} = \mathbf{V}_Q^{1/2}\mathbf{V}_T^{-1/2}\boldsymbol{t}^\top$. For any fixed $\boldsymbol{t}$, we need to verify the following condition to prove this lemma

$$\sup_N \|\boldsymbol{\varsigma}\| < \infty.$$

We note that $\|\boldsymbol{\varsigma}\| \le \sigma_{\max}\left(\mathbf{V}_Q^{1/2}\mathbf{V}_T^{-1/2}\right) \cdot \|\boldsymbol{t}\|$, where $\sigma_{\max}(\cdot)$ denotes the maximum eigenvalue of the corresponding matrix. Hence it is enough to show $\sigma_{\max}(\mathbf{V}_Q^{1/2}\mathbf{V}_T^{-1/2}) \le 1$. Since the covariance matrix $\mathbf{V}_Q$ and $\mathbf{V}_T$ are positive-defined, the following equation holds

$$\mathbf{V}_Q^{1/2}\mathbf{V}_T^{-1/2} = \mathbf{V}_T^{1/4}\left(\mathbf{V}_T^{-1/4}\mathbf{V}_Q^{1/2}\mathbf{V}_T^{-1/4}\right)\mathbf{V}_T^{-1/4},$$

thus $\mathbf{V}_Q^{1/2}\mathbf{V}_T^{-1/2}$ is similar to $\mathbf{V}_T^{-1/4}\mathbf{V}_Q^{1/2}\mathbf{V}_T^{-1/4}$. It only needs to show $\sigma_{\max}(\mathbf{V}_T^{-1/4}\mathbf{V}_Q^{1/2}\mathbf{V}_T^{-1/4}) \le 1$, which is equal to show

$$\mathbf{I}_{p+1} - \mathbf{V}_T^{-1/4}\mathbf{V}_Q^{1/2}\mathbf{V}_T^{-1/4} = \mathbf{V}_T^{-1/4}\left(\mathbf{V}_T^{1/2} - \mathbf{V}_Q^{1/2}\right)\mathbf{V}_T^{-1/4} > 0,$$

that is equivalent to show $\mathbf{V}_T^{1/2} - \mathbf{V}_Q^{1/2}$ is positive-defined.

Recall that $\boldsymbol{M} = \boldsymbol{T} - \boldsymbol{Q}$ and by Lemma S.1, we have $\mathbf{V}_T - \mathbf{V}_Q = \mathbf{V}_M > 0$. Then by the Löwner-Heinz theorem in Zhan (2004), we get $\mathbf{V}_T^{1/2} - \mathbf{V}_Q^{1/2} > 0$ which completes the proof of this lemma. $\square$

Lemma S.4. *Suppose Assumptions 1 and 4 hold. Then we have*

$$\mathbf{V}_T^{-1/2}\boldsymbol{T} \to \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+1}).$$

Proof. Recall the conditions in Lemma S.1 with

$$\boldsymbol{\eta}_{ki} = \boldsymbol{\eta}_{Ni}, \boldsymbol{Z}_k = \mathbf{V}_T^{-1/2}\boldsymbol{Q}, \mathbf{B}_k = \mathbf{B}_N, \boldsymbol{L}_0 \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+1}).$$

By Lemma S.2, $\{M_i, i = 1, \ldots, n\}$ is a martingale difference sequence, then the first two conditions in Lemma S.2 are easily satisfied by Assumption 1. It suffices to show the third condition in Lemma S.1 holds.

By (S.6) in Lemma S.3, we have $\mathbf{V}_Q^{-1/2}\boldsymbol{Q} \to \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+1})$. Next, we devote ourselves to verifying the third condition in Lemma S.1. Let $\mathbf{V}_M$ be the variance of $\boldsymbol{M}$. For any $\boldsymbol{t} \in \mathbb{R}^{p+1}$, we have the following characteristic function

$$\begin{aligned}
&\mathbb{E}\left\{\exp\left(i\boldsymbol{t}^\top \mathbf{V}_T^{-1/2}\boldsymbol{Q}\right)\right\} \cdot \exp\left(-\frac{1}{2}\boldsymbol{t}^\top \mathbf{V}_T^{-1/2}\mathbf{V}_M \mathbf{V}_T^{-1/2}\boldsymbol{t}\right)\\
=&\left\{\exp\left(i\boldsymbol{t}^\top \mathbf{V}_T^{-1/2}\mathbf{V}_Q \mathbf{V}_T^{-1/2}\boldsymbol{t}\right) + o(1)\right\} \cdot \exp\left(-\frac{1}{2}\boldsymbol{t}^\top \mathbf{V}_T^{-1/2}\mathbf{V}_M \mathbf{V}_T^{-1/2}\boldsymbol{t}\right)\\
=&\left\{\exp\left(i\boldsymbol{t}^\top \mathbf{V}_T^{-1/2}\mathbf{V}_Q \mathbf{V}_T^{-1/2}\boldsymbol{t}\right)\right\} \cdot \exp\left(-\frac{1}{2}\boldsymbol{t}^\top \mathbf{V}_T^{-1/2}\mathbf{V}_M \mathbf{V}_T^{-1/2}\boldsymbol{t}\right) + o(1)\\
=&\exp\left(-\frac{1}{2}\boldsymbol{t}^\top \boldsymbol{t}\right) + o(1),
\end{aligned}$$

where the first equality holds by Lemma S.3. And the third condition in Lemma S.1 is satisfied. Then by Lemma S.1 and (S.6) we have

$$\mathbf{V}_T^{-1/2}\boldsymbol{Q} + \mathbf{V}_T^{-1/2}\boldsymbol{M} = \mathbf{V}_T^{-1/2}\boldsymbol{T} \to \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+1}).$$

□

Proof of Theorem 2. By Theorem 1 and Lemma S.4, we have

$$\sqrt{n}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\dagger) = -\sqrt{n}\mathbf{H}(\boldsymbol{\beta}^\dagger)^{-1}\boldsymbol{T} + o_p(1).$$

It follows that

$$\mathbf{V}_T^{-1/2}\mathbf{H}(\boldsymbol{\beta}^\dagger)(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^\dagger) + o_p(1) = -\mathbf{V}_T^{-1/2}\boldsymbol{T}.$$

By Lemma S.4, we have

$$\mathbf{V}^{-1/2}(\widetilde{\boldsymbol{\beta}} - \boldsymbol{\beta}^{\dagger}) \to \mathcal{N}(\mathbf{0}, \mathbf{I}_{p+1}),$$

where $\mathbf{V} = \mathbf{H}(\boldsymbol{\beta}^{\dagger})^{-1}\mathbf{V}_T\mathbf{H}(\boldsymbol{\beta}^{\dagger})^{-1}$. □

Appendix D: Proof of Theorem 3

Proof of Theorem 3. Recall that $\boldsymbol{X}_1^N = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_N)$ and $Y_1^N = (Y_1, \ldots, Y_N)$, then $\mathcal{D}_N = \left\{\boldsymbol{X}_1^N, Y_1^N\right\}$. Let $\mathrm{var}(Y \mid \boldsymbol{X})$ be the conditional variance of Y given $\boldsymbol{X}$. First we calculate $\mathrm{var}(\boldsymbol{T} \mid \boldsymbol{X}_1^N)$. We have

$$\mathrm{var}(\boldsymbol{T} \mid \boldsymbol{X}_1^N) = \mathbb{E}_{Y|\boldsymbol{X}}\left\{\mathrm{var}(\boldsymbol{T} \mid \mathcal{D}_N)\right\} + \mathrm{var}_{Y|\boldsymbol{X}}\left\{\mathbb{E}(\boldsymbol{T} \mid \mathcal{D}_N)\right\}.$$

Some algebra yields

$$\begin{aligned}
\mathrm{var}_{Y|\boldsymbol{X}}\left\{\mathbb{E}(\boldsymbol{T} \mid \mathcal{D}_N)\right\} &= \mathrm{var}_{Y|\boldsymbol{X}}\left(\frac{1}{N}\sum_{j=1}^{N}\xi_j Y_j \widetilde{\boldsymbol{X}}_j\right) \\
&= \frac{1}{N^2}\sum_{j=1}^{N}\mathbb{E}_{Y|\boldsymbol{X}}\left(\xi_j^2 Y_j^2 \widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^{\top}\right) - \frac{1}{N^2}\sum_{j=1}^{N}\left\{\mathbb{E}_{Y|\boldsymbol{X}}(\xi_j Y_j \widetilde{\boldsymbol{X}}_j)\right\}^2 \\
&= \frac{1}{N^2}\sum_{j=1}^{N}\mathbb{E}_{Y|\boldsymbol{X}}\left(\xi_j \widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^{\top}\right) - \frac{1}{N^2}\sum_{j=1}^{N}\left\{\mathbb{E}_{Y|\boldsymbol{X}}(\xi_j Y_j \widetilde{\boldsymbol{X}}_j)\right\}^2,
\end{aligned} \tag{S.7}$$

where the third equality holds by the fact that $\xi_j^2 = \xi_j$ and $Y_j^2 = 1$. Next

$$\begin{aligned}
\mathbb{E}_{Y|\boldsymbol{X}}\left\{\mathrm{var}(\boldsymbol{T} \mid \mathcal{D}_N)\right\} &= \frac{1}{nN^2}\sum_{j=1}^{N}\mathbb{E}_{Y|\boldsymbol{X}}\left\{\pi_j\left(\frac{1}{\pi_j^2}\xi_j^2 Y_j^2 \widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^{\top}\right)\right\} - \frac{1}{nN}\sum_{j=1}^{N}\left\{\mathbb{E}_{Y|\boldsymbol{X}}(\xi_j Y_j \widetilde{\boldsymbol{X}}_j)\right\}^2 \\
&= \frac{1}{nN^2}\sum_{j=1}^{N}\mathbb{E}_{Y|\boldsymbol{X}}\left\{\frac{1}{\pi_j}\xi_j \widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^{\top}\right\} - \frac{1}{nN}\sum_{j=1}^{N}\left\{\mathbb{E}_{Y|\boldsymbol{X}}(\xi_j Y_j \widetilde{\boldsymbol{X}}_j)\right\}^2.
\end{aligned} \tag{S.8}$$

In view of (S.7) and (S.8), we get

$$\begin{aligned}\operatorname{var}(\boldsymbol{T} \mid \boldsymbol{X}_1^N) =& \frac{1}{nN^2}\sum_{j=1}^N \mathbb{E}_{Y|\boldsymbol{X}}\left(\frac{1}{\pi_j}\xi_j\widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^\top\right) + \frac{1}{N^2}\sum_{j=1}^N \mathbb{E}_{Y|\boldsymbol{X}}\left(\xi_j\widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^\top\right)\\ &- \frac{1}{N}\sum_{j=1}^N\left\{\mathbb{E}_{Y|\boldsymbol{X}}\left(\xi_j Y_j\widetilde{\boldsymbol{X}}_j\right)\right\}^2\left(\frac{1}{N}+\frac{1}{n}\right).\end{aligned}$$

Next we calculate $\mathbf{V}_T$ through

$$\mathbf{V}_T = \mathbb{E}\left\{\operatorname{var}(\boldsymbol{T} \mid \boldsymbol{X}_1^N)\right\} + \operatorname{var}\left\{\mathbb{E}(\boldsymbol{T} \mid \boldsymbol{X}_1^N)\right\}.$$

A simple calculation shows that

$$\begin{aligned}\mathbb{E}(T \mid \boldsymbol{X}_1^N) = \mathbb{E}\left\{\mathbb{E}(\boldsymbol{T} \mid \boldsymbol{X}_1^N, Y_1^N)\right\} &= \frac{1}{N}\sum_{j=1}^N \mathbb{E}_{Y|\boldsymbol{X}}\left(\xi_j Y_j\widetilde{\boldsymbol{X}}_j\right),\\ \operatorname{var}\left\{\mathbb{E}(T \mid \boldsymbol{X}_1^N)\right\} &= \frac{1}{N^2}\sum_{j=1}^N \mathbb{E}_{Y|\boldsymbol{X}}\left(\xi_j\widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^\top\right) - \frac{1}{N^2}\sum_{j=1}^N\left\{\mathbb{E}_{Y|\boldsymbol{X}}\left(\xi_j Y_j\widetilde{\boldsymbol{X}}_j\right)\right\}^2.\end{aligned}$$

Therefore, we have

$$\mathbf{V}_T = \frac{1}{nN^2}\sum_{j=1}^N \mathbb{E}_{Y|\boldsymbol{X}}\left(\frac{1}{\pi_j}\xi_j\widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^\top\right) + \mathbf{C},$$

where $\mathbf{C} = 2N^{-2}\sum_{j=1}^N \mathbb{E}_{Y|\boldsymbol{X}}\left(\xi_j\widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^\top\right) - N^{-1}\sum_{j=1}^N\left\{\mathbb{E}_{Y|\boldsymbol{X}}\left(\xi_j Y_j\widetilde{\boldsymbol{X}}_j\right)\right\}^2(2N^{-1}+n^{-1})$ is a constant matrix that does not depend on $\boldsymbol{\pi}$.

Let $\operatorname{tr}(\mathbf{A})$ denotes the trace of matrix $\mathbf{A}$. We minimize $\operatorname{tr}(\mathbf{V}_T)$ to obtain the A-optimality subsampling probability

$$\begin{aligned}\operatorname{tr}(\mathbf{V}_T) &= \frac{1}{nN^2}\sum_{j=1}^N \operatorname{tr}\left\{\mathbb{E}_{Y|\boldsymbol{X}}\left(\frac{1}{\pi_j}\xi_j\mathbf{H}(\boldsymbol{\beta}^\dagger)^{-1}\widetilde{\boldsymbol{X}}_j\widetilde{\boldsymbol{X}}_j^\top\mathbf{H}(\boldsymbol{\beta}^\dagger)^{-1}\right)\right\} + \operatorname{tr}(\mathbf{C})\\ &= \frac{1}{nN^2}\mathbb{E}_{Y|\boldsymbol{X}}\left\{\sum_{j=1}^N\pi_j\sum_{j=1}^N\left(\frac{1}{\pi_j}\xi_j\|\mathbf{H}(\boldsymbol{\beta}^\dagger)^{-1}\widetilde{\boldsymbol{X}}_j\|^2\right)\right\} + \operatorname{tr}(\mathbf{C})\\ &\geq \frac{1}{nN^2}\left\{\sum_{j=1}^N \mathrm{P}\left(Y_j f(\boldsymbol{X}_j,\boldsymbol{\beta}^\dagger)\leq 1\right)\|\mathbf{H}(\boldsymbol{\beta}^\dagger)^{-1}\widetilde{\boldsymbol{X}}_j\|\right\}^2 + \operatorname{tr}(\mathbf{C}),\end{aligned}$$

where the last inequality follows from the Cauchy-Schwarz inequality, and the equality holds if and only if

$$\pi_j^{\mathrm{A}} \propto \mathbb{I}\left(Y_j f(\boldsymbol{X}_j, \boldsymbol{\beta}^\dagger) \leq 1\right) \|\mathbf{H}(\boldsymbol{\beta}^\dagger)^{-1}\widetilde{\boldsymbol{X}}_j\|.$$

Note that $\mathbf{H}(\boldsymbol{\beta}^\dagger)^{-1}\mathrm{var}(\boldsymbol{T} \mid \boldsymbol{X}_1^N)\mathbf{H}(\boldsymbol{\beta}^\dagger)^{-1}$ depends on subsampling probability π only through $\mathrm{var}(\boldsymbol{T} \mid \boldsymbol{X}_1^N)$. Hence, by the similar argument for minimizing $\mathrm{tr}\left\{\mathrm{var}(\boldsymbol{T} \mid \boldsymbol{X}_1^N)\right\}$, we get the L-optimality subsampling probability

$$\pi_j^{\mathrm{L}} \propto \mathbb{I}\left(Y_j f(\boldsymbol{X}_j, \boldsymbol{\beta}^\dagger) \leq 1\right) \|\widetilde{\boldsymbol{X}}_j\|.$$

$\square$

Appendix E: Additional simulation results

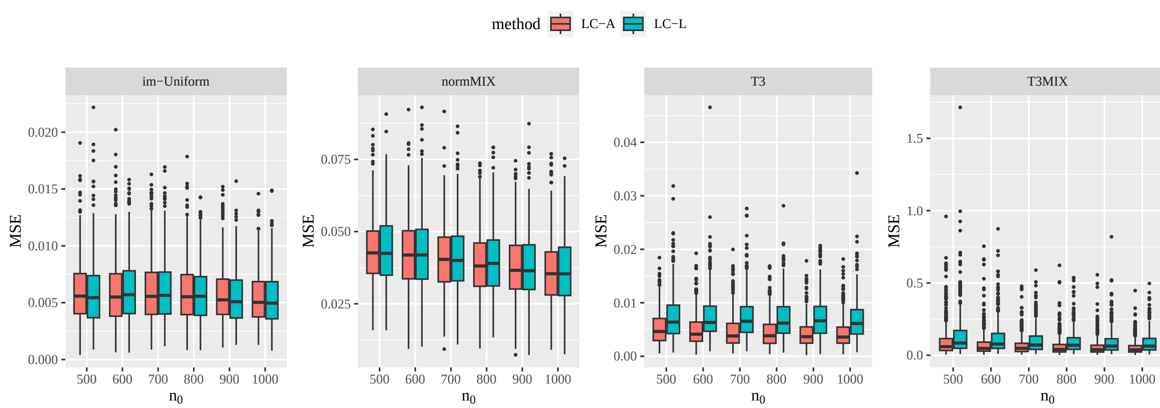

Figure S1: Comparison of MSE for approximating the full sample SVM estimator $\widehat{\boldsymbol{\beta}}$ with different pilot subsample sizes given $n = 1000$ under Scenarios I–IV.

To assess the impact of the pilot study in our proposed algorithm, we conduct the following boxplot by 500 replications on the four scenarios presented in Section 4. Figure S1 reveals that the MSE is not sensitive to the pilot subsample size n_0.

As n_0 increases, the boxplot shows a slight decrease in MSE, suggesting that a smaller pilot subsample size can reduce computational costs without significantly compromising accuracy.

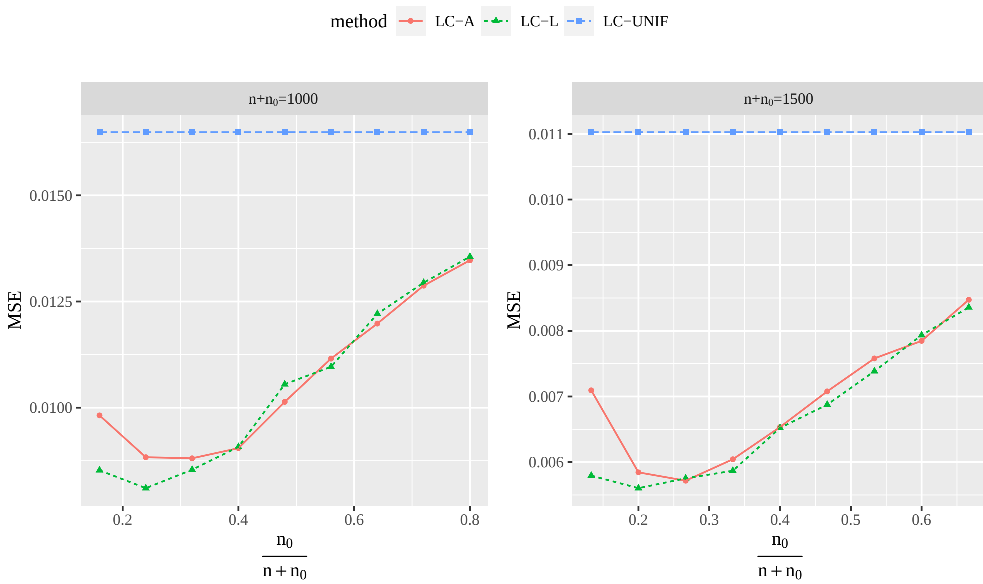

Figure S2: Comparison of mean squared errors (MSEs) for approximating the full sample SVM estimator $\widehat{\boldsymbol{\beta}}$ with different subsample size allocations under Scenario I.

Moreover, we fix the total subsample size of $n + n_0$ and vary the proportions of n and n_0. It provides practical guidelines on allocating subsamples in two steps. We evaluate both $\widehat{\boldsymbol{\pi}}^{\mathrm{A}}$ and $\widehat{\boldsymbol{\pi}}^{\mathrm{L}}$ and the results are presented in Figure S2 under Scenario I. It illustrates that the MSEs increase when n_0 is either too small or too large. This is because that if n_0 is too small, the pilot estimate is not accurate, and thus the optimal subsampling probabilities may not be well approximated; on the other hand, if n_0 is too large, there is not enough sampling budget to select informative subsample in subsequent steps. Figure S2 shows that our methods perform well when the ratio $n_0/(n + n_0)$ is around $(0.2, 0.4)$. Therefore, we use $n_0 = 500$ in our simulation studies with $N = 10^5$.

Bandwidth selection is a critical issue in nonparametric estimation. In Table S1, we compare the MSE and accuracy of LC-A with three bandwidth selectors: Silverman's rule of thumb (ROT, Silverman, 1986), Sheather and Jones method, (SJ, Sheather and Jones, 1991), and biased cross-validation, (BCV, Scott and Terrell, 1987). Clearly, The results demonstrate that the choice of bandwidth selector has a negligible impact on the empirical MSE and accuracy. To this end, we employ the commonly-used bandwidth selector, Silverman's rule of thumb (Silverman, 1986), in our numerical analysis.

Table S1: Comparison of MSE (10^{-2}) and prediction accuracy (%) for LC-A against different bandwidth selectors under Scenarios I–II when $n = 1000$.

		ROT		**SJ**		**BCV**	
Scenario	n_0	MSE	Accuracy	MSE	Accoracy	MSE	Accuracy
im-Uniform	300	0.68	95.54	0.92	94.52	0.65	94.56
	400	0.64	94.53	0.85	94.52	0.61	94.54
	500	0.60	94.53	0.75	94.52	0.60	94.53
normMIX	300	4.84	97.52	4.89	97.52	4.87	97.52
	400	4.49	97.53	4.63	97.53	4.56	97.53
	500	4.33	97.54	4.43	97.54	4.35	97.54

References

Koo, J.-Y., Lee, Y., Kim, Y., and Park, C. (2008). A Bahadur representation of the linear support vector machine. *Journal of Machine Learning Research*, **9**(7):1343–1368.

Pollard, D. (1991). Asymptotics for least absolute deviation regression estimators. *Econometric Theory*, **7**(2):186–199.

Scott, D. W. and Terrell, G. R. (1987). Biased and unbiased cross-validation in density estimation. *Journal of the American Statistical Association*, **82**(400):1131–1146.

Sheather, S. J. and Jones, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, **53**(3):683–690.

Silverman, B. W. (1986). *Density estimation for statistics and data analysis*. Champman & Hall.

Zhan, X. (2004). *Matrix Inequalities*. Springer.

Zhang, T., Ning, Y., and Ruppert, D. (2021). Optimal sampling for generalized linear models under measurement constraints. *Journal of Computational and Graphical Statistics*, **30**(1):106–114.