SUPPLEMENTARY MATERIAL

S1 Proofs for results in Section 2

S1.1 Proof of Proposition 2.1

We prove (i) in Proposition 2.1 for both Algorithm 1 and 2. First we define $Q, \widetilde{Q}, H$ as follows:

$$
\begin{aligned}
Q(\theta;\theta^m) &:= n^{-1}E_{\theta^m}[\log L_f(\theta)|z_1^n, x_1^n] \\
\widetilde{Q}(\theta;\theta^m) &:= -Q(\theta;\theta^m) + P_\lambda(\theta) \\
H(\theta;\theta^m) &:= n^{-1}E_{\theta^m}[\log \mathbb{P}_\theta(y_1^n|z_1^n, x_1^n)|z_1^n, x_1^n].
\end{aligned}
$$

Note that for any θ^m, $\mathcal{F}_n(\theta) = \widetilde{Q}(\theta;\theta^m) + H(\theta;\theta^m)$ holds and $H(\theta^m;\theta^m) \geq H(\theta;\theta^m)$ by Jensen's inequality. Also since θ^{m+1} is a minimizer of $\widetilde{Q}(\theta;\theta^m)$, we have

$$
\mathcal{F}_n(\theta^{m+1}) = \widetilde{Q}(\theta^{m+1};\theta^m) + H(\theta^{m+1};\theta^m) \leq \widetilde{Q}(\theta^m;\theta^m) + H(\theta^m;\theta^m) = \mathcal{F}_n(\theta^m). \tag{S1}
$$

To show that the inequality is strict, it suffices to show that if $\theta^m \notin \mathcal{S}$, θ^m is not a stationary point of $\widetilde{Q}$. Since $\theta^m \notin \mathcal{S}$, there exists θ' such that

$$
\nabla\mathcal{F}_n(\theta^m)^T(\theta' - \theta^m) < 0, \forall \nabla\mathcal{F}_n(\theta^m) \in \partial\mathcal{F}_n(\theta^m) \tag{S2}
$$

Since θ^m is a maximizer of $H(\cdot;\theta^m)$, $\nabla H(\theta^m;\theta^m) = 0$. Then $\partial\mathcal{F}_n(\theta^m) = \partial\widetilde{Q}(\theta^m;\theta^m)$. Thus by (S2), θ^m is not a stationary point of $\widetilde{Q}(\cdot;\theta^m)$.

For Algorithm 2 (PUlasso algorithm), since $\overline{Q}$ is a surrogate function of Q which satisfies following two properties

$$
\overline{Q}(\theta^m;\theta^m) = Q(\theta^m;\theta^m), \quad \overline{Q}(\theta;\theta^m) \leq Q(\theta;\theta^m), \forall\theta \tag{S3}
$$

and θ^{m+1} is a minimizer of $-\overline{Q}(\theta;\theta^m) + P_\lambda(\theta)$, we have

$$\begin{aligned}
\mathcal{F}_n(\theta^m) &= -Q(\theta^m;\theta^m) + P_\lambda(\theta^m) + H(\theta^m;\theta^m) \\
&= -\overline{Q}(\theta^m;\theta^m) + P_\lambda(\theta^m) + H(\theta^m;\theta^m) \\
&\geq -\overline{Q}(\theta^{m+1};\theta^m) + P_\lambda(\theta^{m+1}) + H(\theta^m;\theta^m) \\
&\geq -Q(\theta^{m+1};\theta^m) + P_\lambda(\theta^{m+1}) + H(\theta^{m+1};\theta^m) = \mathcal{F}_n(\theta^{m+1})
\end{aligned}$$

The strict inequality follows from the fact that $\nabla Q(\theta^m;\theta^m) = \nabla\overline{Q}(\theta^m;\theta^m)$.

Now we address (ii) and (iii) in Proposition 2.1. Using the same argument as in Wu (1983), we appeal to the global convergence theorem stated below as Theorem S1.1 in Zangwill (1969) with $\Gamma = \mathcal{S}, \alpha = \mathcal{F}_n$, and letting A be a mapping from θ^m to θ^{m+1} defined by Algorithm 1 or 2. As stated in Wu (1983), condition (iii) in Theorem S1.1 follows from the continuity of $-Q(\theta,\theta') + P_\lambda(\theta)$ or $-\bar{Q}(\theta;\theta') + P_\lambda(\theta)$ in both θ, θ'. Therefore, if we show that $\widetilde{\Theta_0}$ is compact, both (ii) and (iii) follow from the fact that $(\theta^m)_{m=0}^\infty$ lie in a compact set. Since $\widetilde{\Theta_0} \subseteq \mathbb{R}^p$ it suffices to show that $\widetilde{\Theta_0}$ is closed and bounded in $\mathbb{R}^p$. $\widetilde{\Theta_0}$ is bounded since $\mathcal{F}_n(\theta) \to \infty$ whenever $\|\theta\|_2 \to \infty$ since $\|\theta\|_{\mathcal{G},2,1} \geq \min_j w_j \|\theta\|_2 \to \infty$. For closedness of the set, consider $(\theta_k)_{k\geq 1}$ such that $\theta_k \in \widetilde{\Theta_0}$ and $\theta_k \to \theta'$. We have $\mathcal{F}_n(\theta_k) \leq \mathcal{F}_n(\theta_{null})$ for all k. Then by the continuity of $\mathcal{F}_n$, $\mathcal{F}_n(\theta') \leq \mathcal{F}_n(\theta_{null})$ thus $\theta' \in \widetilde{\Theta_0}$.

Theorem S1.1 (Global Convergence Theorem, Zangwill (1969))**.** *Let the sequence $\{x_k\}_{k=0}^\infty$ be generated by $x_{k+1} \in A(x_k)$, where A is a point-to-set map on X. Let a solution set $\Gamma \in X$ be given, and suppose that:*

(i) The sequence $\{x_k\}_{k=0}^\infty \subset S$ for $S \subset X$ a compact set.

(ii) There is a continuous function α on X such that (a) if $x \notin \Gamma$, then $\alpha(y) < \alpha(x)$ for all $y \in A(x)$. (b) if $x \in \Gamma$, then $\alpha(y) \leq \alpha(x)$ for all $y \in A(x)$.

(iii) The mapping A is closed at all points of $X \setminus \Gamma$.

Then all the limit points of any convergent subsequence of $(x_k)_{k=0}^{\infty}$ *are in the solution set* Γ *and* $\alpha(x_k)$ *converges monotonically to* $\alpha(x)$ *for some* $x \in \Gamma$.

S2 Proofs for results in Section 3

S2.1 Derivation of the log-likelihood in the form of GLMs

$$\begin{aligned}\log L(\theta; x, z, s=1) &= \log\left(\prod_i \mathbb{P}_\theta(z_i|x_i, s_i=1)\right) \\ &= \sum_i z_i \log \mathbb{P}_\theta(z_i=1|x_i, s_i=1) + (1-z_i)\log \mathbb{P}_\theta(z_i=0|x_i, s_i=1) \\ &= \sum_i z_i \log \frac{\mathbb{P}_\theta(z_i=1|x_i, s_i=1)}{\mathbb{P}_\theta(z_i=0|x_i, s_i=1)} + \log \mathbb{P}_\theta(z_i=0|x_i, s_i=1).\end{aligned}$$

From Lemma 2.1, we have $\mathbb{P}_\theta(z=1|x, s=1) = \dfrac{\frac{n_l}{\pi n_u} e^{\theta^T x}}{1 + (1+\frac{n_l}{\pi n_u}) e^{\theta^T x}}$. Then,

$$\log \frac{\mathbb{P}_\theta(z=1|x, s=1)}{\mathbb{P}_\theta(z=0|x, s=1)} = \log \frac{\frac{n_l}{\pi n_u} e^{\theta^T x}}{1+e^{\theta^T x}} = \log \frac{n_l}{\pi n_u} + \theta^T x - \log(1+e^{\theta^T x}).$$

and,

$$\begin{aligned}\log \mathbb{P}_\theta(z=0|x, s=1) &= -\log\left(\frac{1+(1+\frac{n_l}{\pi n_u})e^{\theta^T x}}{1+e^{\theta^T x}}\right) = -\log\left(1 + \frac{\frac{n_l}{\pi n_u} e^{\theta^T x}}{1+e^{\theta^T x}}\right) \\ &= -\log\left(1 + e^{\log \frac{n_l}{\pi n_u} + \theta^T x - \log(1+e^{\theta^T x})}\right).\end{aligned}$$

Therefore we obtain,

$$\log\left(\prod_i \mathbb{P}_\theta(z_i|x_i, s_i=1)\right) = \sum_i z_i \eta_i - \log(1+e^{\eta_i})$$

where $\eta_i = \log \frac{n_l}{\pi n_u} + \theta^T x - \log(1+e^{\theta^T x})$.

S2.2 Useful inequalities and technical lemmas

In this section, we provide some results that will be useful for our proofs. First we state the symmetrization inequality, which shows relationships between empirical and Rademacher processes.

Theorem S2.1. *(Symmetrization theorem[van der Vaart and Wellner (1996)]) Let $U_1, \ldots, U_n$ be independent random variables with values in $\mathcal{U}$ and (ϵ_i) be an i.i.d. sequence of Rademacher variables, which take values ± 1 each with probability 1/2. Let Γ be a class of real-valued functions on $\mathcal{U}$. then*

$$E\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n}\{\gamma(U_i)-E(\gamma(U_i))\}\right|\right)\leq 2E\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^{n}\epsilon_i\gamma(U_i)\right|\right).$$

The next theorem is Ledoux-Talagrand contraction theorem. The stated version is Theorem 2.2 in Koltchinskii (2011), which allows T be any subset in $\mathbb{R}^n$, thus slightly more general than the original theorem in Ledoux and Talagrand (1991) where T needs to be bounded.

Theorem S2.2. *(Contraction theorem[Ledoux and Talagrand (1991)]) Let $T \subset \mathbb{R}^n$ and let $\varphi_i : \mathbb{R} \to \mathbb{R}$, $i = 1, \ldots, n$ be contractions which satisfy $|\varphi_i(s) - \varphi_i(t)| \leq |s - t|, s, v \in \mathbb{R}$ and $\varphi_i(0) = 0$. Let (ϵ_i) be independent Rademacher random variables. Then*

$$E\left(\sup_{t\in T}\left|\sum_{i=1}^{n}\epsilon_i\varphi_i(t_i)\right|\right)\leq 2E\left(\sup_{t\in T}\left|\sum_{i=1}^{n}\epsilon_i t_i\right|\right).$$

Finally, we state the bounded differences inequality, also sometimes called as Hoeffding-Azuma inequality.

Theorem S2.3. *(Bounded difference inequality[McDiarmid (1989)]) Let $X_1, \ldots, X_n$ be arbitrary independent random variables on set A and $\varphi : A^n \to \mathbb{R}$ satisfy the bounded difference assumption: there exists constants $c_i, i = 1, \ldots, n$ such that for all $i = 1, \ldots, n$*

and all $x_1, x_2, \ldots, x_i, x_i', \ldots, x_n$,

$$|\varphi(x_1, \ldots, x_i, \ldots, x_n) - \varphi(x_1, \ldots, x_i', \ldots, x_n)| \leq c_i$$

Then $\forall t > 0$,

$$\mathbb{P}\left(\varphi(X_1, \ldots, X_n) - E[\varphi(X_1, \ldots, X_n)] \geq t\right) \leq \exp(-2t^2 / \sum_{i=1}^{n} c_i^2)$$

Now we state and prove some useful results about sub-Gaussian and sub-exponential random variables.

Lemma S2.4. *Let* $v, u \in \mathbb{R}^p$ *and* $(g_1, \ldots, g_J)$ *be a partition of* $(1, \ldots, p)$. *For* $\mathcal{G} = ((g_1, \ldots, g_J), (w_j)_1^J)$ *and* $\bar{\mathcal{G}} = ((g_1, \ldots, g_J), (w_j^{-1})_1^J)$ *such that all* g_j *are non-empty and* $w_j > 0$, $|v^T u| \leq \|v\|_{\mathcal{G},2,1} \|u\|_{\bar{\mathcal{G}},2,\infty}$.

Proof. We note $\|v\|_{\mathcal{G},2,1} = \sum_{j=1}^{J} w_j \|v_{g_j}\|_2$ and $\|u\|_{\bar{\mathcal{G}},2,\infty} := \max_{1 \leq j \leq J} \|w_j^{-1} u_{g_j}\|_2$. By Cauchy-Schwarz inequality, we have

$$|v^T u| \leq \sum_{j=1}^{J} |w_j v_{g_j}^T w_j^{-1} u_{g_j}| \leq \sum_{j=1}^{J} \|w_j v_{g_j}\|_2 \|w_j^{-1} u_{g_j}\|_2.$$

Taking the maximum of the second quantity,

$$|v^T u| \leq \max_{1 \leq j \leq J} \|w_j^{-1} u_{g_j}\|_2 \sum_{j=1}^{J} w_j \|v_{g_j}\|_2 = \|v\|_{\mathcal{G},2,1} \|u\|_{\bar{\mathcal{G}},2,\infty}.$$

□

Lemma S2.5. *Let* $x \in \mathbb{R}^p$ *such that* $x^T v \sim subG(\|v\|_2^2 \sigma_x^2)$ *for any fixed* $v \in \mathbb{R}^p$ *and* $E[x] = 0$. *For any* $i \in (1, \ldots, p)$, $k \geq 1$,

$$E[|x_i|^k] \leq k (2\sigma_x^2)^{k/2} \Gamma(k/2).$$

Proof. Taking $v = e_i$ where e_i is an ith coordinate vector, we have $E(\exp(tv^T x)) = E[\exp(tx_i)] \leq \exp(t^2\sigma_x^2/2)$ for $t \in \mathbb{R}$. Then following a standard argument for sub-Gaussian random variables,

$$\begin{aligned} E[|x_i|^k] &= \int_{s=0}^{\infty} \mathbb{P}(|x_i| \geq s^{1/k}) ds \\ &\leq 2\int_{s=0}^{\infty} \exp(-s^{2/k}/2\sigma_x^2) ds \\ &= k(2\sigma_x^2)^{k/2} \int_{s=0}^{\infty} e^{-u} u^{k/2-1} du = k(2\sigma_x^2)^{k/2}\Gamma(k/2) \end{aligned}$$

where the third inequality comes from the change of variable $u = s^{2/k}/2\sigma_x^2$. □

The next lemma concerns distribution of $x \circ x = [x_1^2, \ldots, x_s^2]$ for independent sub-Gaussian $(x_i)_{i=1}^s$.

Lemma S2.6. *Let $x \in \mathbb{R}^s$ such that $x^T v \sim subG(\|v\|_2^2\sigma_x^2)$ for any fixed $v \in \mathbb{R}^s$ and $E[x] = 0$. Also, assume $(x_i)_{i=1}^s$ are independent. Then we have $v^T(x \circ x) \sim subExp(\nu, b)$ with $\nu = 16\sigma_x^2\|v\|_2$, $b = 16\sigma_x^2\|v\|_\infty$ for any fixed $v \in \mathbb{R}^s$.*

Proof. Let $z := x \circ x - E[x \circ x]$. For any given $v \in \mathbb{R}^s$ and $t > 0$,

$$\begin{aligned} E[\exp(tv^T z)] &= E[\exp(tv_1 z_1 + \ldots tv_s z_s)] \\ &= \prod_{i=1}^{s} E[\exp(tv_i z_i)] \end{aligned}$$

where we use independence. Then by Taylor series expansion,

$$\begin{aligned} E[\exp(tv^T z)] &= \prod_{i=1}^{s} E\left(1 + tv_i z_i + \frac{t^2(v_i z_i)^2}{2} + \ldots\right) \\ &= \prod_{i=1}^{s} \left(1 + \sum_{k=2}^{\infty} \frac{t^k E\left(v_i(x_i^2 - E[x_i^2])\right)^k}{k!}\right) \end{aligned}$$

By Jensen's inequality, we have,

$$E(v_i x_i^2 - E[v_i x_i^2])^k \leq |v_i|^k 2^{k-1}(E[x_i^{2k}] + E[x_i^2]^k),$$

and by applying Jensen's inequality again, we get

$$E[\exp(tv^T z)] \leq \prod_{i=1}^{s} \left(1 + \sum_{k=2}^{\infty} \frac{t^k |v_i|^k 2^k E[x_i^{2k}]}{k!}\right). \tag{S4}$$

We let $t_i = t|v_i|$. By Lemma S2.5, we have,

$$E[x_i^{2k}] \leq (2k)(2\sigma_x^2)^k \Gamma(k) = 2(k!)(2\sigma_x^2)^k \tag{S5}$$

Substituting (S5) into (S4),

$$\begin{aligned} E[\exp(tv^T z)] &\leq \prod_{i=1}^{s} \left(1 + \sum_{k=2}^{\infty} t_i^k 8^k (\sigma_x^2)^k\right) \\ &= \prod_{i=1}^{s} \left(1 + (8t_i\sigma_x^2)^2 \sum_{k=0}^{\infty} (8t_i\sigma_x^2)^k\right) \\ &\leq \prod_{i=1}^{s} \left(1 + 128 t_i^2 \sigma_x^4\right) \end{aligned}$$

if $t|v_i| \leq 1/(16\sigma_x^2)$, for all i. By the fact that $1 + 128t_i^2\sigma_x^4 \leq \exp(128t_i^2\sigma_x^4)$

$$E[\exp(tv^T z)] \leq \prod_{i=1}^{s} \exp(128 t_i^2 \sigma_x^4) = \exp(\sum_{i=1}^{s} 128 t^2 v_i^2 \sigma_x^4) = \exp(128 t^2 \|v\|_2^2 \sigma_x^4)$$

for $t \leq 1/(16\sigma_x^2 \max_i |v_i|)$. Therefore $v^T x \circ x \sim \text{subExp}(\nu, b)$ with $\nu = 16\sigma_x^2\|v\|_2$, $b = 16\sigma_x^2\|v\|_\infty$). □

Also, we have a lemma about maximum of sum of variables with sub-exponential tails.

Lemma S2.7. *Consider $(u_j)_{j=1}^J$ where $u_j \in \mathbb{R}^{m_j}$ such that $\mathbb{1}^T u_j \sim subExp(\nu_j, b)$ with $E[u_j] = 0$ for $1 \leq j \leq J$. We let $m := \max_j m_j$. Also, assume $\exists \nu_* > 0$ such that $\nu_j \leq \nu_* \sqrt{m}$ for all j and $\exists c > 0$ such that $b \leq c\nu_*$. Then we have,*

$$E[\max_{1 \leq j \leq J} \mathbb{1}^T u_j] \leq c\nu_*(\log J + m/(2c^2)).$$

In particular, when $c = 1$, $E[\max_{1 \leq j \leq J} \mathbb{1}^T u_j] \leq \nu_*(\log J + m/2)$.

Proof. For $|t| \leq 1/b$ we have,

$$E[\exp(t\mathbb{1}^T u_j)] \leq \exp(t^2\nu_j^2/2) \leq \exp(mt^2\nu_*^2/2) \tag{S6}$$

Then,

$$\begin{aligned} E[\max_{1\leq j\leq J} \mathbb{1}^T u_j] &= \frac{1}{t} E\left(\log e^{\max_{1\leq j\leq J} t(\mathbb{1}^T u_j)}\right) \\ &\leq \frac{1}{t}\log E\left(e^{\max_{1\leq j\leq J} t(\mathbb{1}^T u_j)}\right) \\ &= \frac{1}{t}\log E\left(\max_{1\leq j\leq J} e^{t(\mathbb{1}^T u_j)}\right). \end{aligned}$$

where the second inequality comes from Jensen's. Using a union bound,

$$\begin{aligned} \frac{1}{t}\log E\left(\max_{1\leq j\leq J} e^{t(\mathbb{1}^T u_j)}\right) &\leq \frac{1}{t}\log\left(\sum_{j=1}^{J} E\left(e^{t(\mathbb{1}^T u_j)}\right)\right) \\ &\leq \frac{1}{t}\log\left(Je^{mt^2\nu_*^2/2}\right). \end{aligned} \tag{S7}$$

where the last inequality uses (S6). Since $1/(c\nu_*) \leq 1/b$ by assumption, the inequality (S7) holds for $t = 1/(c\nu_*)$. Plugging $t = 1/(c\nu_*)$ into (S7), we obtain,

$$E[\max_{1\leq j\leq J} \mathbb{1}^T u_j] \leq c\nu_*(\log J + m/(2c^2))$$

as claimed. □

Finally, in Lemma S2.8 and S2.9, we provide expectation and probability tail bounds of a dual ℓ_1/ℓ_2 norm of a sub-Gaussian vector.

Lemma S2.8. *Let $\mathcal{G} = ((g_1, \ldots, g_J), (w_j)_1^J)$. Consider a random vector $v \in \mathbb{R}^p$ such that for each j and any fixed $u \in \mathbb{R}^{|g_j|}$, $u^T v_{g_j} \sim subG(\sigma^2\|u\|_2^2)$ with $E[v_{g_j}] = 0$ and $u^T(v_{g_j} \circ v_{g_j}) \sim subExp(\nu\|u\|_2, \nu\|u\|_\infty)$. Then,*

$$E[\|v\|_{\bar{\mathcal{G}},2,\infty}] \leq c\sqrt{\log J + m}$$

for $c = (\min_{1\leq j\leq J} w_j)^{-1}\sqrt{\max(\nu, 8\sigma^2)}$, where we define $\bar{\mathcal{G}} = ((g_1, \ldots, g_J), (w_j^{-1})_1^J)$ and $m := \max_j |g_j|$, the largest group size.

Proof. First we let $m_j = |g_j|$. By Holder's inequality, we have,

$$E[\max_{1\le j\le J}\|w_j^{-1}v_{g_j}\|_2] \le E[\max_{1\le j\le J}\|w_j^{-1}v_{g_j}\|_2^2]^{1/2} = E[\max_{1\le j\le J} w_j^{-2}(v_{g_j,1}^2 + \dots v_{g_j,m_j}^2)]^{1/2}$$

Then,

$$\begin{aligned} E[\max_{1\le j\le J} w_j^{-2}(v_{g_j,1}^2 + \dots v_{g_j,m_j}^2)] &\le (\max_{1\le j\le J} w_j^{-2})E[\max_{1\le j\le J}(v_{g_j,1}^2 + \dots v_{g_j,m_j}^2)] \\ &= (\max_{1\le j\le J} w_j^{-2})E[\max_{1\le j\le J}\sum_{i=1}^{m_j}(u_{g_j,i} + E[v_{g_j,i}^2]))] \\ &\le (\max_{1\le j\le J} w_j^{-2})\left(E[\max_{1\le j\le J}\mathbb{1}^T u_{g_j}] + 4m\sigma^2\right) \end{aligned}$$

where $u_{g_j} \overset{d}{:=} v_{g_j}\circ v_{g_j} - E[v_{g_j}\circ v_{g_j}]$ and the last inequality uses Lemma S2.5 and $m_j \le m$, for all j. By assumption, we have, $\mathbb{1}^T u_{g_j} \sim \text{subExp}(\nu\sqrt{m_j},\nu)$ and $E[u_{g_j}] = 0$. Then, by Lemma S2.7,

$$\begin{aligned} LHS &\le (\max_{1\le j\le J} w_j^{-2})[\nu(\log J + m/2) + 4m\sigma^2] \\ &\le (\max_{1\le j\le J} w_j^{-2})\max(\nu, 8\sigma^2)(\log J + m). \end{aligned}$$

Since $\max_{1\le j\le J} w_j^{-2} = 1/(\min_{1\le j\le J} w_j)^2$, defining $c = (\min_{1\le j\le J} w_j)^{-1}\sqrt{\max(\nu, 8\sigma^2)}$, we obtain

$$\|v\|_{\bar{\mathcal{G}},2,\infty} \le c\sqrt{\log J + m}$$

as desired.

□

Lemma S2.9. *Let $\mathcal{G} = ((g_1,\dots,g_J),(w_j)_1^J)$. Consider a random vector $v \in \mathbb{R}^p$ such that for each j and for any fixed $u \in \mathbb{R}^{|g_j|}$, $u^T v_{g_j} \sim subG(\sigma^2\|u\|_2^2)$ with $E[v_{g_j}] = 0$ and $u^T(v_{g_j}\circ v_{g_j}) \sim subExp(\nu\|u\|_2, \nu\|u\|_\infty)$. Then,*

$$\mathbb{P}\left(\|v\|_{\bar{\mathcal{G}},2,\infty} \ge \delta\right) \le J\exp\left(-\frac{1}{2}\min(C_\delta^2/\nu^2, C_\delta/\nu)\right)$$

where we define $C_\delta := (\min_j w_j^2)\delta^2/m - 4\sigma^2$, $\bar{\mathcal{G}} = ((g_1, \ldots, g_J), (w_j^{-1})_1^J)$, *and* $m := \max_j |g_j|$, *the largest group size.*

Proof. By the union bound, we have

$$\mathbb{P}\left(\max_{1\le j\le J}\|w_j^{-1}v_{g_j}\|_2 \ge \delta\right) \le \sum_{j=1}^{J}\mathbb{P}\left(\|w_j^{-1}v_{g_j}\|_2^2 \ge \delta^2\right).$$

Defining $u_{g_j} \overset{d}{:=} v_{g_j} \circ v_{g_j} - E[v_{g_j} \circ v_{g_j}]$ and $m_j := |g_j|$.

$$\begin{aligned}\mathbb{P}\left(\max_{1\le j\le J}\|w_j^{-1}v_{g_j}\|_2 \ge \delta\right) &\le \sum_{j=1}^{J}\mathbb{P}\left(\sum_{k=1}^{m_j} v_{g_j,k}^2 \ge w_j^2\delta^2\right)\\ &\le \sum_{j=1}^{J}\mathbb{P}\left(\mathbb{1}^T u_{g_j} \ge (\min_j w_j^2)\delta^2 - 4m_j\sigma^2\right)\end{aligned}$$

where the last inequality uses Lemma S2.5. By assumption, we have $\mathbb{1}^T u_{g_j} \sim \text{subExp}(\nu\sqrt{m_j}, \nu)$ and $E[u_{g_j}] = 0$. We use Bernstein type inequality to bound the probability. More concretely for any $s > 0$ such that $|s| \le 1/\nu$, we have,

$$\begin{aligned}\mathbb{P}\left(\mathbb{1}^T u_{g_j} \ge (\min_j w_j^2)\delta^2 - 4m_j\sigma^2\right) &\le \mathbb{P}\left(s\mathbb{1}^T u_{g_j} \ge smC_\delta\right)\\ &\le \exp(-smC_\delta)E\left[\exp\left(s\mathbb{1}^T u_{g_j}\right)\right]\\ &\le \exp(-smC_\delta + s^2 m\nu^2/2).\end{aligned}$$

In the first and third inequality, the bound $m_j \le m$ was also used. Optimizing over $s > 0$, we take $s = \min\{C_\delta/\nu^2, 1/\nu\}$. Hence, we have,

$$\mathbb{P}\left(\max_{1\le j\le J}\|w_j^{-1}v_{g_j}\|_2 \ge \delta\right) \le J\exp\left(-\frac{m}{2}\min(C_\delta^2/\nu^2, C_\delta/\nu)\right)$$

□

S2.3 Proof for Proposition 3.1

The proof of this result follows similar lines to the proof of Theorem 1 in Loh and Wainwright (2013), which established the result with a different tolerance function and an additive penalty. Since θ^* is feasible, by the first order optimality condition, we have the following inequality

$$(\nabla \mathcal{L}_n(\hat{\theta}) + \nabla P_\lambda(\hat{\theta}))^T(\theta^* - \hat{\theta}) \geq 0.$$

Letting $\hat{\Delta} := \hat{\theta} - \theta^*$, since $\hat{\theta} \in \Theta_0$ by the setup of the problem, we can apply RSC condition to obtain

$$\alpha\|\hat{\Delta}\|_2^2 - \tau(\|\hat{\Delta}\|_{\mathcal{G},2,1}) \leq (-\nabla P_\lambda(\hat{\theta}) - \nabla \mathcal{L}_n(\theta^*))^T\hat{\Delta}. \tag{S8}$$

On the other hand, convexity of $P_\lambda(\theta)$ implies

$$P_\lambda(\theta^*) - P_\lambda(\hat{\theta}) \geq -\nabla P_\lambda(\hat{\theta})^T\hat{\Delta}. \tag{S9}$$

Combining (S8) with (S9), we obtain

$$\begin{aligned}\alpha\|\hat{\Delta}\|_2^2 - \tau(\|\hat{\Delta}\|_{\mathcal{G},2,1}) &\leq (-\nabla P_\lambda(\hat{\theta}) - \nabla \mathcal{L}_n(\theta^*))^T\hat{\Delta} \\ &\leq P_\lambda(\theta^*) - P_\lambda(\hat{\theta}) + \|\nabla \mathcal{L}_n(\theta^*)\|_{\bar{\mathcal{G}},2,\infty}\|\hat{\Delta}\|_{\mathcal{G},2,1}.\end{aligned}$$

by Lemma S2.4. Since $\tau(\|\hat{\Delta}\|_{\mathcal{G},2,1}) = \tau_1\dfrac{\log J + m}{n}\|\hat{\Delta}\|_{\mathcal{G},2,1}^2 + \tau_2\sqrt{\dfrac{\log J + m}{n}}\|\hat{\Delta}\|_{\mathcal{G},2,1}$,

$$\alpha\|\hat{\Delta}\|_2^2 \leq P_\lambda(\theta^*) - P_\lambda(\hat{\theta}) + \|\hat{\Delta}\|_{\mathcal{G},2,1}\left(\tau_1\frac{\log J + m}{n}\|\hat{\Delta}\|_{\mathcal{G},2,1} + \tau_2\sqrt{\frac{\log J + m}{n}} + \|\nabla \mathcal{L}_n(\theta^*)\|_{\bar{\mathcal{G}},2,\infty}\right),$$

By the choice of λ,

$$\tau_1\frac{\log J + m}{n}\|\hat{\Delta}\|_{\mathcal{G},2,1} + \tau_2\sqrt{\frac{\log J + m}{n}} + \|\nabla \mathcal{L}_n(\theta^*)\|_{\bar{\mathcal{G}},2,\infty} \leq \frac{\lambda}{2}.$$

Then by using the triangle inequality

$$\begin{aligned}\alpha\|\hat{\Delta}\|_2^2 &\le P_\lambda(\theta^*) - P_\lambda(\hat{\theta}) + \frac{\lambda}{2}\|\hat{\Delta}\|_{\mathcal{G},2,1} \\ &= \lambda\sum_{j\in S} w_j\|\theta^*_{g_j}\|_2 - \lambda\sum_{j\in S} w_j\|\hat{\theta}_{g_j}\|_2 - \lambda\sum_{j\in S^c} w_j\|\hat{\theta}_{g_j}\|_2 + \frac{\lambda}{2}\sum_{j=1}^{J} w_j\|\hat{\Delta}_{g_j}\|_2 \\ &\le \lambda\sum_{j\in S} w_j\|\hat{\Delta}_{g_j}\|_2 - \lambda\sum_{j\in S^c} w_j\|\hat{\theta}_{g_j}\|_2 + \frac{\lambda}{2}\sum_{j=1}^{J} w_j\|\hat{\Delta}_{g_j}\|_2\end{aligned}$$

where $S := \{j \in (1,\dots,J); \theta^*_{g_j} \neq 0\}$ where the last inequality comes from the triangle inequality. Since for $j \in S^c$, $\hat{\theta}_{g_j} = \hat{\theta}_{g_j} - \theta^*_{g_j}$,

$$\begin{aligned}\alpha\|\hat{\Delta}\|_2^2 &\le \lambda\sum_{j\in S} w_j\|\hat{\Delta}_{g_j}\|_2 - \lambda\sum_{j\in S^c} w_j\|\hat{\Delta}_{g_j}\|_2 + \frac{\lambda}{2}\sum_{j=1}^{J} w_j\|\hat{\Delta}_{g_j}\|_2 \\ &= \frac{3\lambda}{2}\sum_{j\in S} w_j\|\hat{\Delta}_{g_j}\|_2 - \frac{\lambda}{2}\sum_{j\in S^c} w_j\|\hat{\Delta}_{g_j}\|_2.\end{aligned}$$

In particular, we have

$$\sum_{j\in S^c} w_j\|\hat{\Delta}_{g_j}\|_2 \le 3\sum_{j\in S} w_j\|\hat{\Delta}_{g_j}\|_2 \tag{S10}$$

and

$$\alpha\|\hat{\Delta}\|_2^2 \le \frac{3\lambda}{2}\sum_{j\in S} w_j\|\hat{\Delta}_{g_j}\|_2. \tag{S11}$$

Then,

$$\alpha\|\hat{\Delta}\|_2^2 \le (\max_{j\in S} w_j)\frac{3\lambda}{2}(\sum_{j\in S}\|\hat{\Delta}_{g_j}\|_2^2)^{1/2}(\sum_{j\in S} 1)^{1/2} \le (\max_{j\in S} w_j)\frac{3\lambda}{2}\sqrt{|S|}\|\hat{\Delta}\|_2.$$

The ℓ_1/ℓ_2 upper bound follows from the ℓ_2-bound and

$$\|\hat{\Delta}\|_{\mathcal{G},2,1} = \sum_{j\in S} w_j\|\hat{\Delta}_{g_j}\|_2 + \sum_{j\in S^c} w_j\|\hat{\Delta}_{g_j}\|_2 \le 4(\max_{j\in S} w_j)\sum_{j\in S}\|\hat{\Delta}_{g_j}\|_2 \le 4(\max_{j\in S} w_j)\sqrt{|S|}\|\hat{\Delta}\|_2$$

.

S2.4 Proof of Lemma 3.1

Recalling $\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^n \left(-z_i f(\theta^T x_i) - A(f(\theta^T x_i))\right)$, we have

$$\nabla \mathcal{L}_n(\theta^*) = \frac{1}{n}\sum_{i=1}^n \left(-z_i + \mu(f(\theta^{*T} x_i))\right) \frac{1}{1+e^{\theta^{*T} x_i}} x_i,$$

where we define $A(\eta) = \log(1+e^\eta)$, $\mu(\eta) = A'(\eta) = e^\eta/(1+e^\eta)$ and $f(\theta^T x) = \log(n_\ell/\pi n_u) + \theta^T x - \log(1+e^{\theta^T x})$. For $1 \le i \le n$ and $1 \le j \le p$, define $V_{ij} := \left(-z_i + \mu(f(\theta^{*T} x_i))\right) \frac{1}{1+e^{\theta^{*T} x_i}} x_{ij}$. We note $\nabla \mathcal{L}_n(\theta^*)_j = \frac{1}{n}\sum_{i=1}^n V_{ij}$.

Considering the event, with $C := 36\sigma_x^2$,

$$\mathcal{E} = \left\{ \max_{1\le j\le p} \frac{1}{n}\sum_{i=1}^n x_{ij}^2 \le C \right\}.$$

we have,

$$\mathbb{P}\left(\|\nabla \mathcal{L}_n(\theta^*)\|_{\bar{9},2,\infty} \ge \delta \right) \le \mathbb{P}(\mathcal{E}^c) + \mathbb{P}\left(\|\nabla \mathcal{L}_n(\theta^*)\|_{\bar{9},2,\infty} \ge \delta | \mathcal{E} \right) \mathbb{P}(\mathcal{E}).$$

First we show that $\mathbb{P}(\mathcal{E}^c)$ is small. Since each x_{ij} is a sub-Gaussian variable with sub-Gaussian parameter σ_x, defining $z_{ij} = x_{ij}^2 - E[x_{ij}^2]$,

$$\mathbb{P}(\mathcal{E}^c) \le p\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n x_{ij}^2 \ge 36\sigma_x^2 \right) \le p\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n z_{ij} \ge 32\sigma_x^2 \right)$$

where we use the fact that $E[x_{ij}^2] \le 4\sigma_x^2$. We note that $(z_{ij})_{i=1}^n$ are i.i.d. samples from mean-zero distribution with sub-Exponential tail with parameter $\nu = b = 16\sigma_x^2$ by applying Lemma S2.6 with $s = 1$. By Bernstein-type tail bound of the sub-exponential random variable,

$$\mathbb{P}(\mathcal{E}^c) \le p\mathbb{P}\left(\frac{1}{n}\sum_{i=1}^n z_{ij} \ge 32\sigma_x^2 \right) \le \exp(-\frac{n}{2}(2 - \frac{2\log p}{n})) \le \exp(-n/2), \tag{S12}$$

by the sample size condition $n \gtrsim \log J + m$, assuming sufficiently large n. Now we show that $\frac{1}{n}\sum_{i=1}^n V_{ij}$ is a sub-Gaussian variable on $\mathcal{E}$. In particular, we show that $E[\exp(t\frac{1}{n}\sum_{i=1}^n V_{ij})|\mathcal{E}] \le \exp(t^2 v^2/2)$ for some $v > 0$.

Defining $t_i := \dfrac{t}{n(1+e^{\theta^{*T}x_i})}$, by definition of V_{ij}, we have

$$\begin{aligned} E\left[\exp(\frac{t}{n}V_{ij})|x_i\right] &= E\left[\exp\left(-t_i z_i x_{ij}\right)\cdot\exp\left(t_i\mu(f(x_i^T\theta^*))x_{ij}\right)|x_i\right] \\ &= E\left[\exp\left(-t_i z_i x_{ij}\right)|x_i\right]\cdot\exp\left(t_i\mu(f(x_i^T\theta^*))x_{ij}\right). \end{aligned} \tag{S13}$$

By the property of exponential family, we obtain

$$\begin{aligned} E\left[\exp\left(-t_i z_i x_{ij}\right)|x_i\right] &= \int \exp\left(-t_i z x_{ij}\right)\cdot\exp(zf(x_i^T\theta^*) - A(f(x_i^T\theta^*))dz \\ &= \exp\left\{A(f(x_i^T\theta^*) - t_i x_{ij}) - A(f(x_i^T\theta^*))\right\}. \end{aligned} \tag{S14}$$

Therefore combining (S13) and (S14), we obtain

$$\begin{aligned} E\left[\exp(\frac{t}{n}V_{ij})|x_i\right] &= \exp\left\{A(f(x_i^T\theta^*) - t_i x_{ij}) - A(f(x_i^T\theta^*)) + t_i\mu(f(x_i^T\theta^*))x_{ij}\right\} \\ &\le \exp\left\{\frac{1}{8n^2}(tx_{ij})^2\right\} \end{aligned}$$

where the second inequality comes from the second order Taylor expansion, $\mu(\cdot) = A'(\cdot)$, $\sup_u A''(u) \le 1/4$, and $t_i \le t/n$. Therefore

$$\prod_{i=1}^n E\left[\exp(\frac{t}{n}V_{ij})|x_i\right] \le \exp\left(\frac{t^2}{8n^2}\sum_{i=1}^n x_{ij}^2\right),$$

and conditioned on $\mathcal{E}$, we have the bound

$$\exp\left(\frac{t^2}{8n^2}\sum_{i=1}^n x_{ij}^2\right) \le \exp\left(\frac{t^2 C}{8n}\right).$$

Therefore, $\frac{1}{n}\sum_{i=1}^n V_{ij} \sim \text{subG}(C/4n)$, i.e. $\triangledown\mathcal{L}_n(\theta^*)_j \sim \text{subG}(C/4n)$ for all j.

Now we discuss the distribution of $u^T\triangledown\mathcal{L}_n(\theta^*)_{g_j}$ and $u^T\triangledown\mathcal{L}_n(\theta^*)_{g_j}\circ\triangledown\mathcal{L}_n(\theta^*)_{g_j}$ on $\mathcal{E}$, for any $u \in \mathbb{R}^{|g_j|}$, to apply Lemma S2.9. By Assumption 1, $(\triangledown\mathcal{L}_n(\theta^*)_j)_{j\in g_j}$ are independent. With independence, it is easy to see for any j and any fixed $u \in \mathbb{R}^{|g_j|}$, $u^T\triangledown\mathcal{L}_n(\theta^*)_{g_j} \sim \text{subG}(\|u\|_2^2(C/4n))$ and $E[\triangledown\mathcal{L}_n(\theta^*)] = 0$. Then Lemma S2.6 gives

$$u^T(\triangledown\mathcal{L}_n(\theta^*)_{g_j}\circ\triangledown\mathcal{L}_n(\theta^*)_{g_j}) \sim \text{subExp}(\|u\|_2(4C/n), \|u\|_\infty(4C/n))$$

for any j and fixed $u \in \mathbb{R}^{|g_j|}$. Therefore the condition of Lemma S2.9 is satisfied with $\sigma^2 = C/4n$ and $\nu = 16\sigma^2 = 4C/n$.

We let $\delta^2 = 16C(\log J + m)/(\min_j w_j^2 n)$ and note that

$$C_\delta = \frac{(\min_j w_j^2)\delta^2}{m} - \frac{C}{n} = \frac{16C(\log J + m)}{mn} - \frac{C}{n} = \frac{4C}{n}\left(\frac{16\log J}{4m} + \frac{15}{4}\right)$$

By Lemma S2.9,

$$\mathbb{P}\left(\|\nabla \mathcal{L}_n(\theta^*)\|_{\bar{\mathcal{G}},2,\infty} \geq \delta | \mathcal{E}\right) \leq \exp\left(-\frac{m}{2}\min\left(\frac{C_\delta^2}{(4C/n)^2}, \frac{C_\delta}{4C/n}\right) + \log J\right),$$

and because $\log J/m \geq 0$, $C_\delta \geq 4C/n$, and $\min\left(\frac{C_\delta^2}{(4C/n)^2}, \frac{C_\delta}{4C/n}\right) = \frac{C_\delta}{4C/n}$ if $C_\delta \geq 4C/n$, we have,

$$\begin{aligned}\mathbb{P}\left(\|\nabla \mathcal{L}_n(\theta^*)\|_{\bar{\mathcal{G}},2,\infty} \geq \delta | \mathcal{E}\right) &\leq \exp\left(-\frac{m}{2}\left(\frac{4\log J}{m} + \frac{15}{4}\right) + \log J\right) \\ &\leq \exp\left(-\log J - m\right). \end{aligned} \tag{S15}$$

Putting (S12) and (S15) together, and noting $\delta = (24\sigma_x/\min_j w_j)\sqrt{\frac{\log J + m}{n}}$, we obtain

$$\mathbb{P}\left(\|\nabla \mathcal{L}_n(\theta^*)\|_{\bar{\mathcal{G}},2,\infty} \geq (24\sigma_x/\min_j w_j)\sqrt{\frac{\log J + m}{n}}\right) \leq \exp(-0.5n) + \exp(-\log J - m) \leq \epsilon$$

where the last inequality follows from the sample size condition $n \gtrsim (\log J + m) \vee (1/\epsilon)^{1/\beta}$.

S2.5 Proof of Theorem 3.2

S2.5.1 Proof Outline

Defining $f(\theta^T x) = \log(n_l/\pi n_u) + \theta^T x - \log(1 + e^{\theta^T x})$, we recall that

$$\mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left(-z_i f(\theta^T x_i) + \log(1 + e^{f(\theta^T x_i)})\right).$$

Taking a derivative with respect to θ of $\mathcal{L}_n(\theta)$, we obtain

$$\nabla \mathcal{L}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \left(-z_i + \mu(f(\theta^T x_i))\right) f'(\theta^T x_i) x_i$$

and

$$\begin{aligned}
&\left(\nabla \mathcal{L}_n(\theta) - \nabla \mathcal{L}_n(\theta^*)\right)^T \Delta \\
&= \left(\frac{1}{n}\sum_{i=1}^{n} \left(\mu(f(\theta^T x_i)) - z_i\right) f'(\theta^T x_i) - \left(\mu(f(\theta^{*T} x_i)) - z_i\right) f'(\theta^{*T} x_i) \right) x_i^T \Delta \qquad \text{(S16)}
\end{aligned}$$

where Δ is defined as $\Delta := \theta - \theta^*$, and $A(\cdot), \mu(\cdot)$ defined as $A(\eta) := \log(1+e^\eta)$, $\mu(\eta) := A'(\eta) = e^\eta/(1+e^\eta)$. Also we let $e_i := \mu(f(\theta^{*T} x_i)) - z_i$.

To prove that (S16) is positive with high probability, we decompose (S16) into two terms, whose first term I has a positive expectation and the second term II has an expectation zero. To do so, we add and subtract $\frac{1}{n}\sum_{i=1}^n e_i f'(\theta^T x_i) x_i$ to (S16) to obtain

$$\text{(S16)} = \frac{1}{n}\sum_{i=1}^{n} \Big(\mu(f(\theta^T x_i)) - \mu(f(\theta^{*T} x_i)\Big)) f'(\theta^T x_i) x_i^T \Delta + e_i (f'(\theta^T x_i) - f'(\theta^{*T} x_i)) x_i^T \Delta.$$

Applying a Taylor expansion around $f(\theta^{*T} x_i)$, we obtain

$$\begin{aligned}
&\left(\nabla \mathcal{L}_n(\theta) - \nabla \mathcal{L}_n(\theta^*)\right)^T \Delta \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n} A''(f(\theta^{*T} x_i) + v_i(f(\theta^T x_i) - f(\theta^{*T} x_i)))(f(\theta^T x_i) - f(\theta^{*T} x_i)) f'(\theta^T x_i) x_i^T \Delta}_{\text{I}} \qquad \text{(S17)} \\
&\quad + \underbrace{\frac{1}{n}\sum_{i=1}^{n} e_i (f'(\theta^T x_i) - f'(\theta^{*T} x_i)) x_i^T \Delta}_{\text{II}} \text{ for } v_i \in [0,1] \qquad \text{(S18)}
\end{aligned}$$

where $A''(\eta) = e^\eta/(1+e^\eta)^2$. We will show that the expectation of I is positive. We immediately see $E[e_i(f'(\theta^T x_i) - f'(\theta^{*T} x_i)) x_i^T \Delta] = 0$ because $E[e_i | x_i] = 0$.

We aim to show each inequality

$$I \geq \kappa_0 \|\Delta\|_2^2 - \kappa_1 \|\Delta\|_{\mathcal{G},2,1} \|\Delta\|_2 \sqrt{\frac{\log J + m}{n}} \tag{S19}$$

$$|II| \leq \kappa_2 \|\Delta\|_{\mathcal{G},2,1} \sqrt{\frac{\log J + m}{n}} \tag{S20}$$

holds for all $\Delta \in \{\Delta; \|\Delta\|_2 \leq r\}$ with probability at least $1 - \epsilon/2$ for some $\kappa_0, \kappa_1, \kappa_2 > 0$.

Then

$$I + II \geq \kappa_0 \|\Delta\|_2^2 - \kappa_1 \|\Delta\|_{\mathcal{G},2,1} \|\Delta\|_2 \sqrt{\frac{\log J + m}{n}} - \kappa_2 \|\Delta\|_{\mathcal{G},2,1} \sqrt{\frac{\log J + m}{n}}$$

holds for all $\Delta \in \{\Delta; \|\Delta\|_2 \leq r\}$ with probability at least $1 - \epsilon$. Finally, by the inequality $a^2 + b^2 \geq 2ab$, we obtain,

$$I + II \geq (\kappa_0/2) \|\Delta\|_2^2 - (2\kappa_1^2/\kappa_0) \left(\frac{\log J + m}{n} \right) \|\Delta\|_{\mathcal{G},2,1}^2 - \kappa_2 \sqrt{\frac{\log J + m}{n}} \|\Delta\|_{\mathcal{G},2,1}$$

for all $\Delta \in \{\Delta; \|\Delta\|_2 \leq r\}$ with probability at least $1 - \epsilon$.

S2.5.2 Obtaining a lower bound of term I

We use a similar argument in Negahban et al. (2012) to obtain a lower bound of the first term. The main difference is that we get the dependence on θ for a curvature term, which is not the case for a canonical link $f(\theta^T x) = \theta^T x$. Since $f'(u) = \dfrac{1}{1 + e^u}$, the first term I becomes

$$I = \frac{1}{n} \sum_{i=1}^{n} A''(f(\theta^{*T} x_i) + v_i(f(\theta^T x_i) - f(\theta^{*T} x_i))) \frac{(x^T \Delta)^2}{(1 + e^{x_i^T \theta^* + v_i' x_i^T \Delta})(1 + e^{x_i^T \theta})}.$$

for some $v_i' \in [0, 1]$ by Taylor expansion. We note

$$I \geq \frac{1}{n} \sum_{i=1}^{n} \frac{A''(f(\theta^{*T} x_i) + v_i(f(\theta^T x_i) - f(\theta^{*T} x_i)))}{(1 + e^{x_i^T \theta^* + v_i' x_i^T \Delta})(1 + e^{x_i^T \theta})} (x_i^T \Delta)^2 \mathbb{1}\{|\Delta^T x_i| \leq \tau \|\Delta\|_2\}$$

for any $\tau \geq 0$, as $A''(u) = \frac{e^u}{(1+e^u)^2} \geq 0, \forall u$. A suitable τ will be chosen shortly. Since on the event

$$|\Delta^T x_i| \leq \tau \|\Delta\|_2, \tag{S21}$$

we have $\theta^T x_i \leq |\theta^{*T} x_i| + |\Delta^T x_i| \leq K_1^r + \tau r$ and

$$\begin{aligned} |f(\theta^{*T} x_i) + v_i(f(\theta^{*T} x_i) - f(\theta^T x_i))| &\leq |f(\theta^{*T} x_i)| + |f(\theta^{*T} x_i) - f(\theta^T x_i)| \\ &\leq \left|\log \frac{n_l}{\pi n_u}\right| + |\theta^{*T} x_i| + |\Delta^T x_i|, \end{aligned}$$

by Assumption 3 and the fact that $x^T\theta - \log(1 + e^{x^T\theta})$ is 1-Lipschitz in $x^T\theta$, I can be further lower-bounded by

$$I \geq \frac{L_0(\tau)}{n} \sum_{i=1}^{n} (x_i^T \Delta)^2 \mathbb{1}\{|\Delta^T x_i| \leq \tau \|\Delta\|_2\},$$

where $L_0(\tau)$ is defined as $L_0(\tau) := \inf_{|u| \leq K_2 + K_1^\tau + \tau r} \frac{A''(u)}{(1 + e^{K_1^\tau + \tau r})^2}$. Finally, we truncate each term $(x_i^T \Delta)^2 \mathbb{1}\{|\Delta^T x_i| \leq \tau \|\Delta\|_2\}$ so that each term is Lipschitz in $(x_i^T \Delta)$. For a truncation level $\tau > 0$, we define the following function:

$$\varphi_\tau(u) = \begin{cases} u^2 & \text{if } |u| \leq \frac{\tau}{2} \\ (\tau - u)^2 & \text{if } \frac{\tau}{2} \leq |u| \leq \tau \\ 0 & \text{otherwise} \end{cases}$$

and note that $I \geq \frac{1}{n} \sum_{i=1}^n L_0(\tau) \varphi_{\tau\|\Delta\|_2}(\Delta^T x_i)$, since if the event (S21) holds, $(\Delta^T x_i)^2 \geq \varphi_{\tau\|\Delta\|_2}(\Delta^T x_i)$, and both left and right-hand sides are 0 if the event does not hold.

Defining I_ℓ as

$$I_\ell := \frac{L_0(\tau)}{n} \sum_{i=1}^{n} \varphi_{\tau\|\Delta\|_2}(\Delta^T x_i), \tag{S22}$$

we note that it is sufficient to show the inequality

$$I_\ell \geq \kappa_0 \|\Delta\|_2^2 - \kappa_1 \|\Delta\|_{\mathcal{G},2,1} \|\Delta\|_2 \sqrt{\frac{\log J + m}{n}} \tag{S23}$$

holds with high probability for all $\Delta \in \{\Delta; \|\Delta\|_2 \leq r\}$ to prove (S19). To do so, first we will show the inequality (S23) is true for $\Delta \in \mathbb{S}(\delta, t)$, where we define

$$\mathbb{S}(\delta, t) := \{\Delta \in \mathbb{R}^p; \|\Delta\|_2 = \delta, \|\Delta\|_{\mathcal{G},2,1}/\|\Delta\|_2 \leq t\}. \tag{S24}$$

If $\Delta = 0$, the inequality (S23) is trivially true. Otherwise, we show that

$$\frac{L_0(\tau)}{n\delta^2} \sum_{i=1}^{n} \varphi_{\tau\|\Delta\|_2}(\Delta^T x_i) \geq \kappa_0 - \kappa_1 t \sqrt{\frac{\log J + m}{n}}, \tag{S25}$$

is true for all $\Delta \in \mathbb{S}(\delta, t)$ with high probability. Then we will use a homogeneity property of φ and peeling argument to obtain a uniform result over (δ, t).

S2.5.3 Bounding Expectation of Term I

We note that I_ℓ is lower bounded by,

$$I_\ell = E[I_\ell] + (I_\ell - E[I_\ell]) \geq E[I_\ell] - \sup_{\Delta \in \mathbb{S}(\delta,t)} |I_\ell - E[I_\ell]|.$$

In this sub-section, we obtain the lower bound of $E[I_\ell]$, which is strictly positive with a suitably chosen τ. In the next sub-section, we will control the deviation term $\sup_{\Delta \in \mathbb{S}(\delta,t)} |I_\ell - E[I_\ell]|$. First we have $E[I_\ell] = L_0(\tau) E\left[\varphi_{\tau\|\Delta\|_2}(\Delta^T x)\right]$ where $x \overset{d}{=} x_i$, and

$$E\left[\varphi_{\tau\|\Delta\|_2}(\Delta^T x)\right] = E[(\Delta^T x)^2] - E[(\Delta^T x)^2 - \varphi_{\tau\|\Delta\|_2}(\Delta^T x)].$$

We lower and upper bound each two terms on the right-hand side by

$$E[(\Delta^T x)^2] \geq K_0 \|\Delta\|_2^2$$

and

$$E[(\Delta^T x)^2 - \varphi_{\tau\|\Delta\|_2}(\Delta^T x)] \leq E\left[(\Delta^T x)^2 \mathbb{1}\left\{|\Delta^T x| \geq \frac{\tau\|\Delta\|_2}{2}\right\}\right]$$

Applying the Cauchy-Schwarz inequality, we obtain

$$E\left[(\Delta^T x)^2 \mathbb{1}\left\{|\Delta^T x| \geq \frac{\tau\|\Delta\|_2}{2}\right\}\right] \leq \sqrt{E(\Delta^T x)^4}\sqrt{\mathbb{P}\left(|\Delta^T x| \geq \frac{\tau\|\Delta\|_2}{2}\right)}$$
$$\leq 4\sqrt{2}\sigma_x^2 \exp\left(-\frac{\tau^2}{16\sigma_x^2}\right)\|\Delta\|_2^2$$

by using expectation and tail-bound of sub-Gaussians, since $\Delta^T x \sim \text{subG}(\|\Delta\|_2^2\sigma_x^2)$. As $4\sqrt{2}\sigma_x^2\left(\exp\left(-\frac{\tau^2}{16\sigma_x^2}\right)\right) \leq \frac{K_0}{4}$ for $\tau^2 \geq 16\sigma_x^2 \log\frac{16\sqrt{2}\sigma_x^2}{K_0}$, we take $\tau = K_3 := 4\sigma_x\left(\log\frac{16\sqrt{2}\sigma_x^2}{K_0}\right)^{1/2}$ to have

$$\begin{aligned} E[I_\ell] &= L_0(K_3)E\left[\varphi_{K_3\|\Delta\|_2}(\Delta^T x)\right] \\ &\geq L_0(K_3)\|\Delta\|_2^2\left(K_0 - 4\sqrt{2}\sigma_x^2\exp\left(-\frac{\tau^2}{16\sigma_x^2}\right)\right) \\ &\geq \|\Delta\|_2^2\frac{3L_0(K_3)K_0}{4}. \end{aligned} \tag{S26}$$

For simplicity, we write $L_0 := L_0(K_3)$ for future references.

S2.5.4 Controlling the difference of Term I from its expectation

We now bound the term $\sup_{\Delta\in\mathbb{S}(\delta,t)}|I_\ell - E[I_\ell]|$ using the concentration property of an empirical process. We have $\sup_{\Delta\in\mathbb{S}(\delta,t)}|I_\ell - E[I_\ell]| = \delta^2 L_0 U_1(t)$, where we define $U_1(t)$ as

$$U_1(t) := \sup_{\Delta\in\mathbb{S}(\delta,t)}\left|\frac{1}{n\|\Delta\|_2^2}\sum_{i=1}^n \varphi_{K_3\|\Delta\|_2}(\Delta^T x_i) - E\left[\varphi_{K_3\|\Delta\|_2}(\Delta^T x)\right]\right|,$$

since $\|\Delta\|_2 = \delta$ for all $\Delta \in \mathbb{S}(\delta, t)$. Since we have $\|\varphi_{K_3\|\Delta\|_2}\|_\infty \leq \frac{K_3^2\|\Delta\|_2^2}{4}$ by definition of $\varphi_\tau(\cdot)$, we apply bounded difference inequality with $c_i = K_3^2/2n$ (Theorem S2.3) to obtain

$$\mathbb{P}(U_1(t) \geq EU_1(t) + u_1) \leq 2\exp\left(-\frac{8nu_1^2}{K_3^4}\right).$$

Setting $u_1 = K_0/4$,

$$\mathbb{P}(U_1(t) \geq \mathbb{E}[U_1(t)] + \frac{K_0}{4}) \leq 2\exp(-c_1 n) \tag{S27}$$

where $c_1 = K_0^2/2K_3^4$ is a constant depending on K_0 and K_3. Now we calculate $EU_1(t)$. By symmetrization and contraction inequalities (Theorems S2.1, S2.2), we have

$$\begin{aligned} E[U_1(t)] &\leq 2E\left[\sup_{\Delta\in\mathbb{S}(\delta,t)}\left|\frac{1}{n\|\Delta\|_2^2}\sum_{i=1}^n \epsilon_i\varphi_{K_3\|\Delta\|_2}(\Delta^T x_i)\right|\right] \\ &\leq \frac{8K_3\delta}{\delta^2}E\left[\sup_{\Delta\in\mathbb{S}(\delta,t)}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i\Delta^T x_i\right|\right] \\ &\leq 8K_3\delta^{-1}\left(\sup_{\Delta\in\mathbb{S}(\delta,t)}\|\Delta\|_{\mathcal{G},2,1}\right)E\left[\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\|_{\bar{\mathcal{G}},2,\infty}\right] \\ &\leq 8K_3K_4 t\sqrt{\frac{\log J+m}{n}} \end{aligned} \tag{S28}$$

where $(\epsilon_i)_{i=1}^n$ are i.i.d Rademacher variables and $K_4 := 20\sigma_x(\min_j w_j)^{-1}$. Note that $\varphi_{K_3\|\Delta\|_2}$ is a Lipschitz function with the Lipschitz constant $= 2K_3\|\Delta\|_2 = 2K_3\delta$ for $\Delta \in \mathbb{S}(\delta,t)$ which allows us to apply the Ledoux-Talagrand contraction theorem. The second last inequality is from Lemma S2.4 and the last inequality follows from $E\left[\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\|_{\bar{\mathcal{G}},2,\infty}\right] \leq K_4\sqrt{\frac{\log J+m}{n}}$, which will be proven shortly in Lemma S2.10.

Therefore, combining (S26), (S27) and (S28), we have

$$\inf_{\Delta\in\mathbb{S}(\delta,t)}\frac{L_0}{n\|\Delta\|_2^2}\sum_{i=1}^n \varphi_{K_3\|\Delta\|_2}(\Delta^T x_i) \geq \kappa_0 - \kappa_1' t\sqrt{\frac{\log J+m}{n}} \tag{S29}$$

with probability at least $1-\exp(-c_1 n)$ where $\kappa_0 = K_0L_0/2$ and $\kappa_1' = 8L_0K_3K_4$. It remains to prove Lemma S2.10.

Lemma S2.10.

$$E\left[\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\|_{\bar{\mathcal{G}},2,\infty}\right] \leq c\sqrt{\frac{\log J+m}{n}} \tag{S30}$$

for $n \geq \log p$, where $c := 20\sigma_x(\min_j w_j)^{-1}$ is a constant depending on $\sigma_x, (w_j)_1^J$.

Proof. Conditioned on x_1^n, $\frac{1}{n}\sum_{i=1}^n \epsilon_i x_{ij}$ is a sub-Gaussian with a parameter $\frac{1}{n^2}\sum_i x_{ij}^2$, since $\epsilon_i \sim \text{subG}(1)$. Then $\frac{1}{n}\sum_{i=1}^n \epsilon_i x_{ij} \sim \text{subG}(C(x)/n)$, where we define $C(x) = \max_{1\leq j\leq p}\frac{1}{n}\sum_i x_{ij}^2$

conditioned on x_1^n. Defining $u := [u_1, \ldots, u_p]^T \in \mathbb{R}^p$ as $u_j = \frac{1}{n}\sum_{i=1}^n \epsilon_i x_{ij}$, we have independence of $(u_j)_{j\in g_j}$ by Assumption 1. Following similar arguments as in the proof of Lemma 3.1, we obtain for any j and $v \in \mathbb{R}^{|g_j|}$, $v^T u_{g_j} \sim \mathrm{subG}((C(x)/n)\|v\|_2^2)$ and $v^T(u_{g_j} \circ u_{g_j}) \sim \mathrm{subExp}(\nu\|v\|_2, \nu\|v\|_\infty)$ with $\nu = 16C(x)/n$. Then Lemma S2.8 gives,

$$E\left[\|\frac{1}{n}\sum_{i=1}^n \epsilon_i u_i\|_{\bar{\mathcal{G}},2,\infty}|x_1^n\right] \le 4(\min_j w_j)^{-1}\sqrt{C(x)}\sqrt{\frac{\log J + m}{n}}.$$

Therefore,

$$E\left[\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\|_{\bar{\mathcal{G}},2,\infty}\right] \le 4(\min_j w_j)^{-1}\sqrt{\frac{\log J + m}{n}}E[\sqrt{C(x)}]$$

Now we upper-bound $E[\sqrt{C(x)}]$. By Holder's inequality,

$$E\left[\sqrt{\max_{1\le j\le p}\frac{1}{n}\sum_{i=1}^n x_{ij}^2}\right] \le E\left[\max_{1\le j\le p}\frac{1}{n}\sum_{i=1}^n x_{ij}^2\right]^{1/2}$$

Now we define $z_{ij} := x_{ij}^2 - E[x_{ij}^2]$ for each $1 \le i \le n$ and $1 \le j \le p$ and $z_j = [z_{1j}, \ldots, z_{nj}]^T$. Using Lemma S2.5, we have,

$$E\left[\max_{1\le j\le p}\frac{1}{n}\sum_{i=1}^n x_{ij}^2\right] \le E\left[\max_{1\le j\le p}\frac{1}{n}\sum_{i=1}^n z_{ij}\right] + 4\sigma_x^2.$$

Since $\mathbb{1}^T z_j \sim \mathrm{subExp}(16\sigma_x^2\sqrt{n}, 16\sigma_x^2)$ by Lemma S2.6, we apply Lemma S2.7 with $\nu_* = 16\sigma_x^2\sqrt{n}$, $c = 1/\sqrt{n}$ (taking $m_j = 1, \forall j$) to obtain

$$n^{-1}E[\max_{1\le j\le p}\mathbb{1}^T z_j] \le n^{-1}16\sigma_x^2(\log p + n/2) = 16\sigma_x^2\frac{\log p}{n} + 8\sigma_x^2,$$

Hence,

$$E[\sqrt{C(x)}] \le 4\sigma_x\sqrt{\log p/n + 1/2} \le 5\sigma_x$$

by the condition of $\log p/n \le 1$, and thus,

$$E\left[\|\frac{1}{n}\sum_{i=1}^n \epsilon_i x_i\|_{\bar{\mathcal{G}},2,\infty}\right] \le 20\sigma_x(\min_j w_j)^{-1}\sqrt{\frac{\log J + m}{n}}$$

□

S2.5.5 Extending the inequality (S29) for all $\Delta \in \mathbb{B}_2(r)$

In this section, we show

$$\frac{L_0}{n\|\Delta\|_2^2}\sum_{i=1}^{n}\varphi_{K_3\|\Delta\|_2}(\Delta^T x_i) \geq \kappa_0 - \kappa_1\left(\frac{\|\Delta\|_{\mathcal{G},2,1}}{\|\Delta\|_2}\right)\sqrt{\frac{\log J + m}{n}} \tag{S31}$$

holds for all $\|\Delta\|_2 = \delta$ with probability at least $1-\epsilon/2$ where $\kappa_1 = 2\kappa_1'$. Note if (S31) holds, for any Δ' such that $\|\Delta'\|_2 = \delta' \neq \delta$, we can apply (S31) to $\Delta = \Delta'(\delta/\delta')$ to obtain

$$\frac{L_0}{n\|\Delta'\|_2^2}\sum_{i=1}^{n}\varphi_{K_3\|\Delta'\|_2}(\Delta'^T x_i) \geq \kappa_0 - \kappa_1\left(\frac{\|\Delta'\|_{\mathcal{G},2,1}}{\|\Delta'\|_2}\right)\sqrt{\frac{\log J + m}{n}}$$

by using homogeneity property of φ,i.e. $\varphi_\tau(x) = c^{-2}\varphi_{c\tau}(cx)$ for any $c > 0$. Thus proving that (S31) holds for all $\|\Delta\|_2 = \delta$ with probability at least $1-\epsilon/2$ is enough to prove that the same inequality holds for all $\|\Delta\|_2 \leq r$ with the same high probability. We let $\mathbb{S}_2(\delta) := \{\Delta \in \mathbb{R}^p; \|\Delta\|_2 = \delta\}$ and $K_w > 0$ be a constant such that $\min_j w_j \geq K_w$, where the existence of K_w is guaranteed by Assumption 4.

$$\begin{aligned}
&\mathbb{P}\left(\exists \Delta \in \mathbb{S}_2(\delta) \text{ such that inequality (S31) fails }\right)\\
&\leq \sum_{l=1}^{N_L}\mathbb{P}\left(\exists \Delta \in \mathbb{S}_2(\delta); K_w 2^{l-1} \leq \frac{\|\Delta\|_{\mathcal{G},2,1}}{\|\Delta\|_2} \leq K_w 2^l \text{ s.t inequality (S31) fails }\right)
\end{aligned} \tag{S32}$$

where $2^{N_L} \leq (\max_j w_j/K_w)\sqrt{J}$, i.e. $N_L := \left\lceil \log_2\left(\max_j w_j\sqrt{J}/K_w\right)\right\rceil$, by the inequality $K_w\|\Delta\|_2 \leq (\min_j w_j)\|\Delta\|_2 \leq \|\Delta\|_{\mathcal{G},2,1} \leq (\max_j w_j)\sqrt{J}\|\Delta\|_2$.

$$\begin{aligned}
&\sum_{l=1}^{N_L}\mathbb{P}\left(\exists \Delta \in \mathbb{S}_2(\delta); K_w 2^{l-1} \leq \frac{\|\Delta\|_{\mathcal{G},2,1}}{\|\Delta\|_2} \leq K_w 2^l \text{ such that inequality (S31) fails }\right)\\
&\leq \sum_{l=1}^{N_L}\mathbb{P}\left(\inf_{\Delta\in\mathbb{S}_2(\delta); \frac{\|\Delta\|_{\mathcal{G},2,1}}{\|\Delta\|_2}\leq(K_w 2^l)} \frac{L_0}{n\|\Delta\|_2^2}\sum_{i=1}^{n}\varphi_{K_3\|\Delta\|_2}(\Delta^T x_i) < \kappa_0 - \kappa_1(K_w 2^{l-1})\sqrt{\frac{\log J + m}{n}}\right)\\
&= \sum_{l=1}^{N_L}\mathbb{P}\left(\inf_{\Delta\in\mathbb{S}(\delta,(K_w 2^l))} \frac{L_0}{n\|\Delta\|_2^2}\sum_{i=1}^{n}\varphi_{K_3\|\Delta\|_2}(\Delta^T x_i) < \kappa_0 - \kappa_1'(K_w 2^l)\sqrt{\frac{\log J + m}{n}}\right)\\
&\leq \exp(-c_1 n + \log N_L)
\end{aligned}$$

by $\kappa_1 = 2\kappa_1'$ and the inequality (S29). Finally,

$$\exp(-c_1 n + \log N_L) \leq \exp\left(-c_1 n + \log\log_2(J^{3/2}/K_w)\right) \lesssim \exp\left(-c_1 n + \log\log J\right) \leq \epsilon/2$$

by the sample size condition $n \gtrsim (\log J + m) \vee (1/\epsilon)^{1/\beta}$ and $\max_j w_j/J \leq 1$.

S2.5.6 Controlling the difference of Term II from its expectation

For the second term, we recall the definition :

$$II = \frac{1}{n}\sum_{i=1}^{n} e_i(f'(\theta^T x_i) - f'(\theta^{*T} x_i))x_i^T \Delta,$$

and note that $E[II] = 0$ by $E[e_i|x_i] = 0$. Similar to $U_1(t)$, we define a following quantity,

$$U_2(t) := \sup_{(1/2)t \leq \|\Delta\|_{\mathcal{G},2,1} \leq t} \left| \frac{1}{n\|\Delta\|_{\mathcal{G},2,1}} \sum_{i=1}^{n} e_i(f'(\theta^T x_i) - f'(\theta^{*T} x_i))x_i^T \Delta. \right|$$

, and bound $E(U_2(t))$ using symmetrization and contraction theorem. First we define

$$g_i(\Delta^T x_i) := e_i\left(f'(\theta^{*T} x_i + \Delta^T x_i) - f'(\theta^{*T} x_i)\right)\Delta^T x_i.$$

and prove that g_i/L_g is a contraction map where $L_g := 3 + (K_1^r/4)$.

Lemma S2.11. *$g_i(s)/L_g$ is a contraction map with $g_i(0) = 0$.*

Proof. We consider the first derivative of g_i. For ease of notation, we let $u_i^* := \theta^{*T} x_i$. We note $f'(u) = 1/(1+e^u), f'(u) = -e^u/(1+e^u)^2$. Thus $\sup_u |f'(u)| \leq 1, \sup_u |f''(u)| \leq 1/4$. Also, elementary calculation shows that $\sup_u |uf'(u)|, \sup_u |uf''(u)| \leq 1/2$. Since,

$$g_i(u) = e_i(f'(u_i^* + u) - f'(u_i^*))u$$

we have,

$$\begin{aligned}
|g_i'(u)| &= |e_i(f''(u_i^* + u)u + f'(u_i^* + u) - f'(u_i^*))| \\
&\leq |f''(u_i^* + u)(u_i^* + u) - f''(u_i^* + u)u_i^* + f'(u_i^* + u) - f'(u_i^*))| \\
&\leq 3 + (1/4)|u_i^*|
\end{aligned}$$

where $|e_i| \leq 1$ was used in the first inequality. By Assumption 2, $u_i^* := |\theta^{*T} x_i| \leq K_1^r$, thus we can take $L_g := 3 + (1/4)K_1^r$. □

Back to $E(U_2(t))$, by symmetrization and contraction theorem (Theorems S2.1,S2.2),

$$\begin{aligned} E(U_2(t)) &\leq 4L_g E\left[\sup_{(1/2)t \leq \|\Delta\|_{\mathcal{G},2,1} \leq t} \left| \frac{1}{n\|\Delta\|_{\mathcal{G},2,1}} \sum_{i=1}^n \epsilon_i \Delta^T x_i \right| \right] \\ &\leq 4L_g E\left[\sup_{(1/2)t \leq \|\Delta\|_{\mathcal{G},2,1} \leq t} \frac{1}{(1/2)t} \|\Delta\|_{\mathcal{G},2,1} \| \frac{1}{n} \sum_{i=1}^n \epsilon_i x_i \|_{\bar{\mathcal{G}},2,\infty} \right] \\ &\leq 8K_4 L_g \sqrt{\frac{\log J + m}{n}}. \end{aligned} \tag{S33}$$

where the second inequality uses the fact that $(1/2)t \leq \|\Delta\|_{\mathcal{G},2,1} \leq t$ and Lemma S2.4, and the last inequality comes from Lemma S2.10.

Now, we apply bounded difference inequality to show that $U_2(t)$ is close to $E(U_2(t))$ with probability at least $1 - \exp(-c'n)$. We have,

$$\begin{aligned} \sup_{i,\theta} \frac{1}{n\|\Delta\|_{\mathcal{G},2,1}} |g_i(\Delta^T x_i)| &= \sup_{i,\theta} \frac{1}{n\|\Delta\|_{\mathcal{G},2,1}} \left| e_i \left(f'(\theta^{*T} x_i + \Delta^T x_i) - f'(\theta^{*T} x_i) \right) \Delta^T x_i \right| \\ &\leq \sup_{i,\theta} \frac{2}{n\|\Delta\|_{\mathcal{G},2,1}} |\Delta^T x_i| \leq \frac{2}{n} \max_{i,j} w_j^{-1} \|(x_i)_{g_j}\|_2 \end{aligned}$$

by Lemma S2.4. We note for any $w \in \mathbb{R}^p$ such that $w_{g_j^c} = 0$ and $\|w\|_2 = 1$, $u \in \mathbb{R}^p$, defined as $u := \theta^* + rw$, satisfies $\|u - \theta^*\|_2 \leq r$ and $supp(u - \theta^*) \subseteq g_j$. By Assumption 2, $|x_i^T u| \leq K_1^r$ a.s. for all i. Then $|x_i^T w| = |x_i^T (u - \theta^*)|/r \leq 2K_1^r/r$ a.s., which implies $\|(x_i)_{g_j}\|_2 \leq 2K_1^r/r$ since $\|(x_i)_{g_j}\|_2 = \sup_{v \in \mathbb{R}^{|g_j|}; \|v\|_2 = 1} |(x_i)_{g_j}^T v| = \sup_{w \in \mathbb{R}^p; \|w\|_2 = 1, w_{g_j^c} = 0} |x_i^T w|$. As the bound holds for any i, j, we have $\max_{i,j} \|(x_i)_{g_j}\|_2 \leq 2K_1^r/r$.

Hence by applying Theorem S2.3 with $c_i = (8K_1^r/K_w r)n^{-1}$, we obtain

$$\mathbb{P}(U_2(t) \geq EU_2(t) + u_2) \leq \exp(-2u_2^2 / \sum_{i=1}^n c_i^2)$$

Taking $u_2 = K_4 L_g \sqrt{\frac{\log J + m}{n}}$, we get

$$\mathbb{P}\left(U_2(t) \geq 9K_4 L_g \sqrt{\frac{\log J + m}{n}} \right) \leq \exp(-c_2(\log J + m))$$

where $c_2 := (K_w r K_4 L_g)^2 / 32 (K_1^r)^2$. In other words, we have shown, for any $t > 0$,

$$\mathbb{P}\left(\left| \frac{1}{n} \sum_{i=1}^{n} e_i (f'(\theta^{*T} x_i + \Delta^T x_i) - f'(\theta^{*T} x_i)) x_i^T \Delta \right| \leq \kappa_2' \|\Delta\|_{\mathcal{G},2,1} \sqrt{\frac{\log J + m}{n}}, \forall (1/2)t \leq \|\Delta\|_{\mathcal{G},2,1} \leq t \right)$$
$$\geq 1 - \exp(-c_2(\log J + m)) \tag{S34}$$

where we define $\kappa_2' := 9K_4 L_g$.

S2.5.7 Extending the inequality (S34) for all $\Delta \in \mathbb{B}_2(r)$

In this section, we obtain a uniform result for term II. More concretely, we consider the following inequality:

$$\left| \frac{1}{n} \sum_{i=1}^{n} e_i (f'(\theta^{*T} x_i + \Delta^T x_i) - f'(\theta^{*T} x_i)) x_i^T \Delta \right| \leq \kappa_2 \|\Delta\|_{\mathcal{G},2,1} \sqrt{\frac{\log J + m}{n}} \tag{S35}$$

where $\kappa_2 := 10 K_4 L_g$. Equivalently, defining

$$\phi(\Delta; x_1^n, z_1^n) := \frac{1}{n \|\Delta\|_{\mathcal{G},2,1}} \sum_{i=1}^{n} e_i (f'(\theta^{*T} x_i + \Delta^T x_i) - f'(\theta^{*T} x_i)) x_i^T \Delta$$

for $\Delta \neq 0$, we aim to establish the result,

$$\mathbb{P}\left(|\phi(\Delta; x_1^n, z_1^n)| \leq \kappa_2 \sqrt{\frac{\log J + m}{n}}, \forall \Delta \in \mathbb{B}_2(r) \right) \geq 1 - \epsilon/2.$$

We first define

$$\mathbb{A}(r_1, r_2) := \{\Delta \in \mathbb{R}^p; r_1 < \|\Delta\|_{\mathcal{G},2,1} \leq r_2\}$$

and decompose $\mathbb{B}_2(r)$ into different regions. We have,

$$\begin{aligned}
&\mathbb{P}\left(\exists \Delta \in \mathbb{B}_2(r) \text{ such that inequality (S35) fails }\right) \\
&\leq \mathbb{P}\left(\exists \Delta \in \mathbb{A}(0, C_n) \text{ such that inequality (S35) fails }\right) \qquad \text{(S36)} \\
&+ \sum_{k=1}^{N_K} \mathbb{P}\left(\exists \Delta \in \mathbb{A}(r_{k-1}, r_k) \text{ such that inequality (S35) fails }\right) \qquad \text{(S37)}
\end{aligned}$$

where we define

$$\begin{aligned}
C_n &:= K_4 L_g \left(\frac{(\min_j w_j) r}{K_1^r}\right)^2 \sqrt{\frac{\log J + m}{n}} \\
r_k &:= C_n 2^k.
\end{aligned}$$

Here C_n is chosen to ensure the probability (S36) to be small enough, which will be shown shortly. We take N_K such that $r_{N_K} = C_n 2^{N_K} \geq r \max_j w_j \sqrt{J}$ since $\|\Delta\|_{\mathcal{G},2,1} \leq (\max_j w_j)\sqrt{J}\|\Delta\|_2 \leq r(\max_j w_j)\sqrt{J}$. Then we can let,

$$N_K := \left\lceil \log_2 \left(c \max_j w_j \sqrt{\frac{nJ}{\log J + m}} \right) \right\rceil$$

for $c := (K_1^r)^2/(rK_w^2 K_4 L_g) \vee 1$. By the sample size assumption, $\max_j w_j/n \leq 1$ and $J \gtrsim n^\beta$, thus

$$N_K \leq \log_2 \left(c \max_j w_j \sqrt{\frac{nJ}{\log J + m}} \right) \leq 2 \log \left(c' n^{(3+\beta)/2} \right).$$

for some $c' > 1$. Since $\mathbb{P}\left(\exists \Delta \in \mathbb{A}(r_{k-1}, r_k) \text{ such that inequality (S35) fails }\right) \leq \exp(-c_2(m + \log J))$ for any k by (S34), we have for (S37),

$$\begin{aligned}
\text{(S37)} &\leq \exp(-c_2(m + \log J) + \log N_K) \\
&\leq 2 \exp\left(-c_2(m + \log J) + \log\log c' n^{(3+\beta)/2}\right) \\
&\leq c_3 \exp(-c_2(m + \log J) + \log\log n)
\end{aligned}$$

for $c_3 = 2((3+\beta)/2 + \log c') > 1$, as $\log\log c' n^{(3+\beta)/2} \leq \log\log n + \log((3+\beta)/2 + \log c')$.

Now we address (S36):

$$(\text{S36}) = \mathbb{P}\left(\exists \Delta \in \mathbb{A}(0, C_n); |\phi(\Delta; x_1^n, z_1^n)| > \kappa_2 \sqrt{\frac{\log J + m}{n}}\right)$$

For $s \in (0, C_n]$, we define a function $\widetilde{\phi} : \mathbb{R}_+ \times \mathbb{R}^p \to \mathbb{R}$, whose first argument takes the size (measured in $\|\cdot\|_{\mathcal{G},2,1}$ norm), second argument takes normalized direction (i.e. $\|d\|_{\mathcal{G},2,1} = 1$) such that

$$\widetilde{\phi}(s, d; x_1^n, z_1^n) := \frac{1}{n}\sum_{i=1}^n e_i(f'(\theta^{*T} x_i + s x_i^T d) - f'(\theta^{*T} x_i)) x_i^T d = \phi(sd; x_1^n, z_1^n)$$

In particular, for any $\Delta \in \mathbb{A}(0, C_n)$, we have $\widetilde{\phi}(\|\Delta\|_{\mathcal{G},2,1}, \Delta/\|\Delta\|_{\mathcal{G},2,1}; x_1^n, z_1^n) = \phi(\Delta; x_1^n, z_1^n)$.

Now we calculate how much ϕ changes when the size of the input vector varies while fixing the direction. In other words, we calculate the rate of change of $\widetilde{\phi}$ with respect to its first argument. To ease the notation, we suppress the dependence of $\phi, \widetilde{\phi}$ on (x_1^n, z_1^n).

$$\begin{aligned}
|\frac{d}{ds}\widetilde{\phi}(s, d)| &= \left|\frac{d}{ds}\left(\frac{1}{n}\sum_{i=1}^n e_i f'(\theta^{*T} x_i + s x_i^T d) - f'(\theta^{*T} x_i)) x_i^T d\right)\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n \left|e_i f''(\theta^{*T} x_i + s x_i^T d)\right| (x_i^T d)^2 \\
&\leq \frac{1}{4}\|x_i\|_{\mathcal{G},2,\infty}^2 \|d\|_{\mathcal{G},2,1}^2 \leq \left(\frac{K_1^r}{(\min_j w_j) r}\right)^2
\end{aligned}$$

by $|e_i| \leq 1$ and $\|f''\|_\infty \leq (1/4)$. Then for any normalized direction $d \in \mathbb{R}^p$ such that $\|d\|_{\mathcal{G},2,1} = 1$, we have,

$$|\widetilde{\phi}(s, d) - \widetilde{\phi}(u, d)| \leq \left(\frac{K_1^r}{(\min_j w_j) r}\right)^2 |s - u|$$

In particular, for any $0 < s \leq C_n$,

$$|\widetilde{\phi}(s, \Delta/\|\Delta\|_{\mathcal{G},2,1})| \leq |\widetilde{\phi}(C_n, \Delta/\|\Delta\|_{\mathcal{G},2,1})| + \left(\frac{K_1^r}{(\min_j w_j) r}\right)^2 C_n$$

Therefore,

$$\begin{aligned}
\text{(S36)} &= \mathbb{P}\left(\exists \Delta \in \mathbb{A}(0, C_n); |\widetilde{\phi}(\|\Delta\|_{\mathcal{G},2,1}, \Delta/\|\Delta\|_{\mathcal{G},2,1})| > \kappa_2 \sqrt{\frac{\log J + m}{n}}\right) \\
&\leq \mathbb{P}\left(\exists \Delta \in \mathbb{A}(0, C_n); |\widetilde{\phi}(C_n, \Delta/\|\Delta\|_{\mathcal{G},2,1})| > \kappa_2 \sqrt{\frac{\log J + m}{n}} - \left(\frac{K_1^r}{(\min_j w_j) r}\right)^2 C_n\right) \\
&= \mathbb{P}\left(\exists \Delta \in \mathbb{A}(0, C_n); |\widetilde{\phi}(C_n, \Delta/\|\Delta\|_{\mathcal{G},2,1})| > 9 K_4 L_g \sqrt{\frac{\log J + m}{n}}\right)
\end{aligned}$$

where the last line uses the fact $\left(\frac{K_1^r}{(\min_j w_j) r}\right)^2 C_n = K_4 L_g \sqrt{\frac{\log J + m}{n}}$. Since $\widetilde{\phi}(C_n, \Delta/\|\Delta\|_{\mathcal{G},2,1}) = \phi(C_n \Delta/\|\Delta\|_{\mathcal{G},2,1})$ and $C_n \Delta/\|\Delta\|_{\mathcal{G},2,1} \in \{\Delta' \in \mathbb{R}^p; \|\Delta'\|_{\mathcal{G},2,1} = C_n\}$, we have,

$$\begin{aligned}
\text{(S36)} &\leq \mathbb{P}\left(\sup_{\|\Delta\|_{\mathcal{G},2,1} = C_n} |\phi(\Delta)| > 9 K_4 L_g \sqrt{\frac{\log J + m}{n}}\right) \\
&\leq \mathbb{P}\left(\sup_{(1/2) C_n \leq \|\Delta\|_{\mathcal{G},2,1} \leq C_n} |\phi(\Delta)| > 9 K_4 L_g \sqrt{\frac{\log J + m}{n}}\right) \\
&\leq \exp(-c_2(\log J + m))
\end{aligned}$$

by (S34). Therefore,

$$\begin{aligned}
\text{(S36)} + \text{(S37)} &\leq \exp(-c_2(\log J + m)) + c_3 \exp(-c_2(m + \log J) + \log\log n) \\
&\leq 2 c_3 \exp(-c_2(m + \log J) + \log\log n) \leq \epsilon/2
\end{aligned}$$

by the sample size condition $n \gtrsim (\log J + m) \vee (1/\epsilon)^{1/\beta}$, noting $\log\log n = o(\log J)$.

S3 Supplementary simulation results in Section 4

In this section, we display additional classification performance results. We recall the simulation setting: dimension of features $p \in (10, 5000)$, auto-correlation level among features $\rho \in (0, 0.2, 0.4, 0.6, 0.8)$, separation distance $d \in (1.5, 2.5, 3.5)$, and the model specification scheme (logistic, misspecified). The sample size is $n_\ell = n_u = 500$ in all setting and experiments are repeated 50 times.

S3.1 The logistic model scheme

S3.1.1 F_1 scores under the logistic model scheme

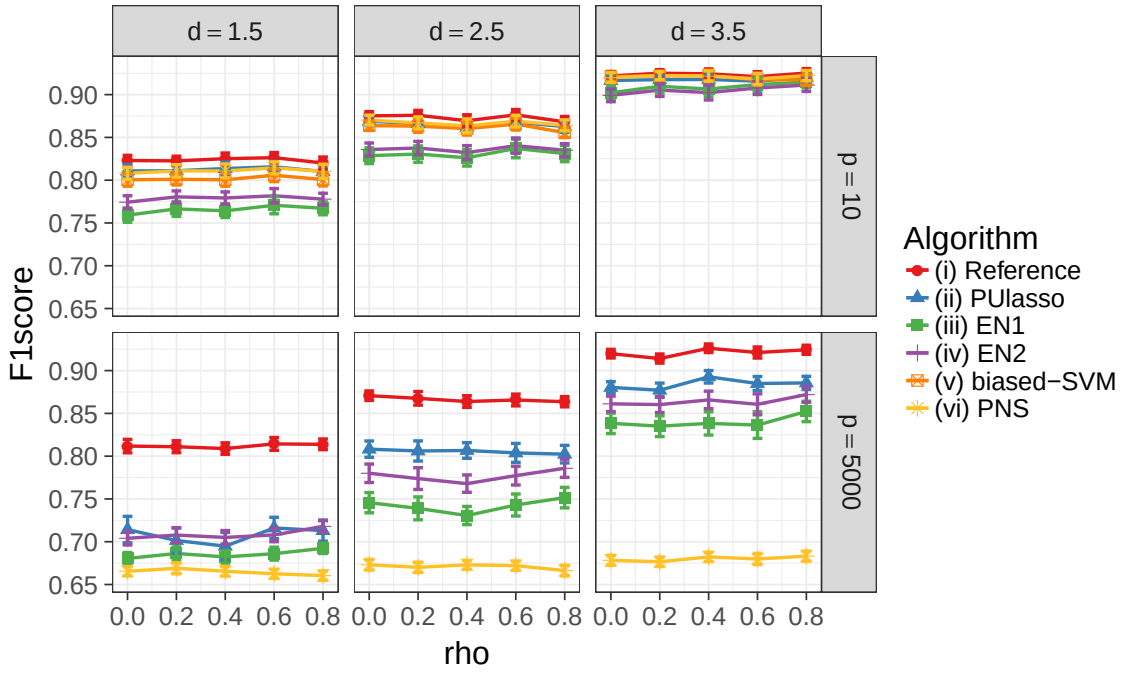

Figure S1: F_1 scores of algorithms (i)-(vi) under correct (logistic) model specification

S3.2 The misspecified model scheme

Heavy-tailed distribution tends to generate more separated samples, leading to better classification performance. The scaling of Σ_ρ, which sets $Var(x_i^T\theta^*)$ the same across ρ, indirectly changes the separation between the two classes. As a result, we observe improved classification performance with higher ρ in the misspecified setting. PUlasso algorithm continues to out-perform other algorithms in most cases, but performance difference among algorithms decreases under the model misspecification scheme.

S3.2.1 Mis-classification rates under the misspecified model

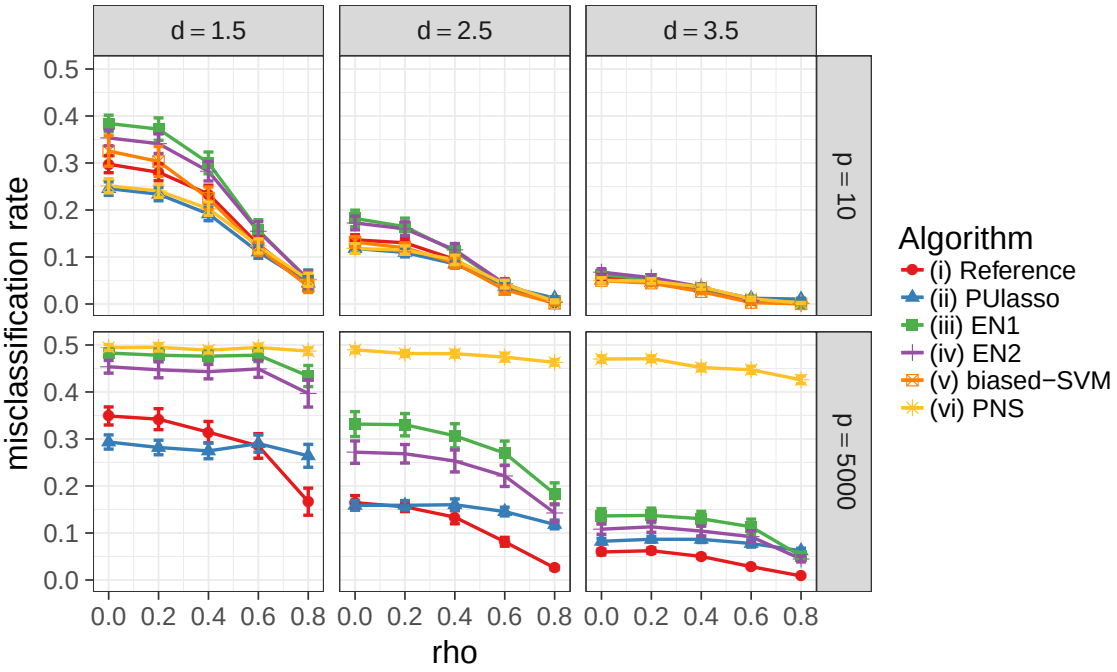

Figure S2: Mis-classification rates of algorithms (i)-(vi) under model misspecification.

S3.2.2 F_1 scores under the misspecified model

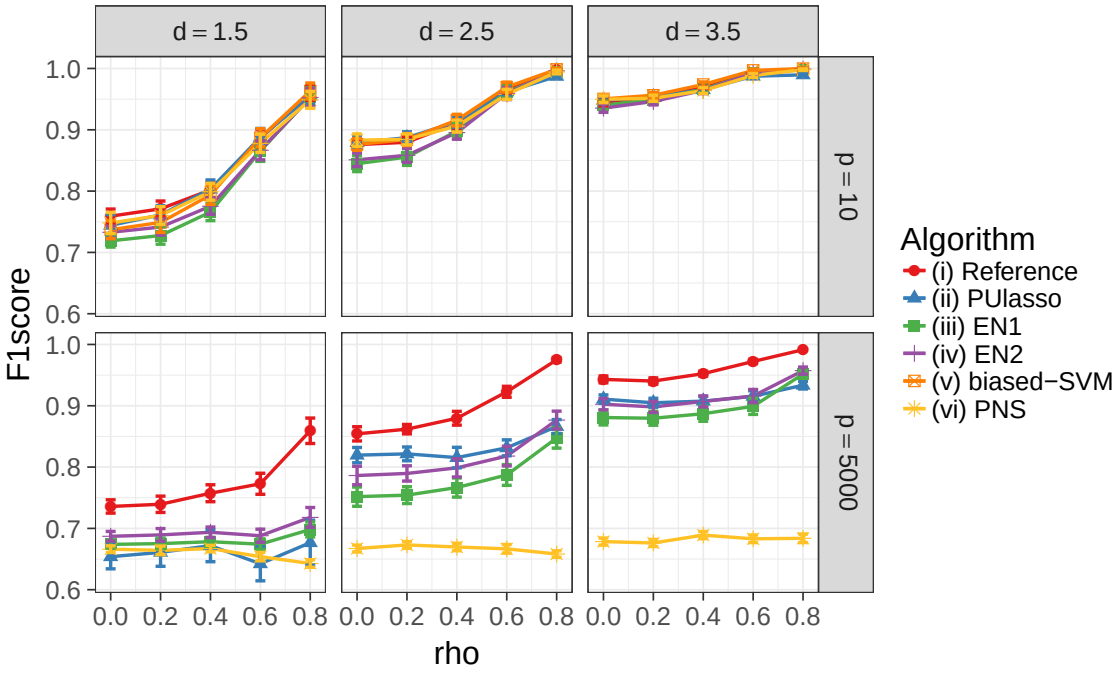

Figure S3: F_1 scores of algorithms (i)-(vi) under model misspecification

References

Vladimir Koltchinskii. *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: École D'Été de Probabilités de Saint-Flour XXXVIII-2008*. Springer Science & Business Media, July 2011.

M. Ledoux and M. Talagrand. *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, New York, NY, 1991.

P-L. Loh and M. J. Wainwright. Regularized M-estimators with nonconvexity: Statistical and algorithmic theory for local optima. *Journal of Machine Learning Research*, 1:1–9, 2013.

Colin McDiarmid. On the method of bounded differences. *Surveys in combinatorics*, 141 (1):148–188, 1989.

S. N. Negahban, R. Pradeep, Bin Yu, and M. J. Wainwright. A Unified Framework for High-Dimensional Analysis of M-Estimators with Decomposable Regularizers. *Statistica Sinica*, 27(4):538–557, 2012.

A W van der Vaart and J Wellner. *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics. Springer, 1996.

C. F. Jeff Wu. On the Convergence Properties of the EM Algorithm. *The Annals of Statistics*, 11(1):95–103, 1983.

W I Zangwill. *Nonlinear programming: a unified approach*. Prentice-Hall international series in management. Prentice-Hall, 1969.