Appendices for "Semismooth Newton Coordinate Descent Algorithm for Elastic-Net Penalized Huber Loss Regression and Quantile Regression"

Congrui Yi and Jian Huang

Department of Statistics and Actuarial Science
University of Iowa
Iowa City, IA 52242

In these appendices, we provide some background information on the KKT conditions and the Newton derivative, derive the SNA for penalized Huber loss regression, and prove Theorems 3.3 and 3.4 in the main text. The proof of Theorem 3.1 is omitted since it is very similar to that of Theorem 3.3.

A Background on Convex Analysis and Properties of Newton Derivative

To derive the KKT conditions (2.3), we recall some background in convex analysis. We also describe some useful properties of Newton derivative.

For a convex function f, a vector w is called a *subgradient* of f at z if

$$f(x) - f(z) \geq w^\top (x - z), \quad \forall x. \tag{A.1}$$

The set of all subgradients of f at z is called the *subdifferential*, denoted as $\partial f(z)$. For example, the subdifferential of the absolute value function has the following form

$$\partial |z| = \begin{cases} \{\text{sign}(z)\} & \text{if } z \neq 0, \\ [-1, 1] & \text{if } z = 0. \end{cases} \tag{A.2}$$

For convex optimization problems, the necessary and sufficient optimality conditions are called the KKT conditions. In the case of unconstrained optimization, the KKT conditions can be stated in terms of Fermat's rule (Rockafellar, 1970): for a convex function f,

$$\mathbf{0} \in \partial f(z^*) \Leftrightarrow z^* = \arg\min_z f(z). \tag{A.3}$$

This holds because by definition $\mathbf{0} \in \partial f(z^*)$ if and only if for any z we have $f(z) - f(z^*) \geq \mathbf{0}^\top (z - z^*) = 0$, i.e. $z^* = \arg\min_z f(z)$.

A more general result (Combettes and Wajs, 2005) is

$$w \in \partial f(z) \Leftrightarrow z = \mathrm{Prox}_f(z + w), \tag{A.4}$$

where Prox_f is the *proximity operator* for f defined as

$$\mathrm{Prox}_f(z) \coloneqq \arg\min_x \frac{1}{2}\|x - z\|_2^2 + f(x).$$

The second statement can be shown as follows. Applying Fermat's rule,

$$z = \mathrm{Prox}_f(z + w) = \arg\min_x \frac{1}{2}\|x - z - w\|_2^2 + f(x),$$

if and only if there exists $s \in \partial f(x)$ such that

$$\mathbf{0} = (z - z - w) + s = -w + s,$$

that is,

$$w = s \in \partial f(x).$$

It can shown that the proximity operator of the absolute value $|\cdot|$ is given in closed form by the soft-thresholding operator with threshold 1, i.e.

$$\mathrm{Prox}_{|\cdot|}(z) = S(z) = \mathrm{sgn}(z)(|z| - 1)_+. \tag{A.5}$$

Then it follows from (A.4) that $s_j \in \partial|\beta_j|$ can be expressed as an equation

$$\beta_j - S(\beta_j + s_j) = 0. \tag{A.6}$$

According to the Fermat's rule (A.3), the KKT conditions for the penalized Huber loss regression (2.1) are

$$\begin{cases} -\frac{1}{n}\sum_i h'_\gamma(y_i - \widehat{\beta}_0 - x_i^\top \widehat{\beta}) = 0, \\ -\frac{1}{n}\sum_i h'_\gamma(y_i - \widehat{\beta}_0 - x_i^\top \widehat{\beta}) x_{ij} + \lambda\alpha\widehat{s}_j + \lambda(1-\alpha)\widehat{\beta}_j = 0, \\ \widehat{s}_j \in \partial|\widehat{\beta}_j|, \quad j = 1, \ldots, p, \end{cases} \tag{A.7}$$

where $(\widehat{\beta}_0, \widehat{\beta})$ is an optimizer. Rewriting the last row by (A.6), we obtain the KKT conditions as a system of equations (2.3).

The definition of "Newton derivative" is already given in the main text. Now we provide several properties useful for calculating Newton derivatives. The first one is the following chain rule for Newton derivatives (Ito and Kunisch, 2008).

Lemma A.1. *If $F : \mathbb{R}^m \to \mathbb{R}^n$ is continuously Fréchet differentiable at $z \in \mathbb{R}^m$ with Jacobian J_F and $G : \mathbb{R}^n \to \mathbb{R}^l$ is Newton differentiable at $F(z)$ with a Newton derivative H_G. Then $G \circ F$ is Newton differentiable at z with a Newton derivative $H_G(F(z+h))J_F(z+h)$ for h sufficiently small.*

We also derived two other results.

Lemma A.2. *In the following, assume $F : \mathbb{R}^m \to \mathbb{R}^l$, $G : \mathbb{R}^m \to \mathbb{R}^l$, $z \in \mathbb{R}^m$, $F = (F_1, \ldots, F_l)^\top$ and $H = (H_1^\top, \ldots, H_l^\top)^\top$, where $H_i \in \mathbb{R}^{1\times m}$, $\quad i = 1, \ldots, l$.*

(i) If F is continuously Fréchet differentiable at z, then F is also Newton differentiable at z and $J_F \in \nabla_N F(z)$;

(ii) If F is Newton differentiable at z, then for any integer $k > 0$ and $A \in \mathbb{R}^{k\times l}$, AF is Newton differentiable at z; if $H \in \nabla_N F(z)$, then $AH \in \nabla_N AF(z)$;

(iii) If F and G are Newton differentiable at z, then $F + G$ is Newton differentiable at z; if $H_F \in \nabla_N F(z)$, $H_G \in \nabla_N G(z)$, then $H_F + H_G \in \nabla_N (F + G)(z)$;

(iv) F is Newton differentiable at z if and only if $F_1, \ldots, F_l$ are all Newton differentiable at z and $H \in \nabla_N F(z) \Leftrightarrow H_i \in \nabla_N F_i(z), \quad i = 1, \ldots, l$;

Lemma A.3. *A univariate piecewise-smooth real function f is everywhere Newton differentiable, with a Newton derivative H given by*

$$H(z) = \begin{cases} f'(z) & \text{if } f \text{ is differentiable at } z, \\ r_z \in \mathbb{R}^1 & \text{if } f \text{ is not differentiable at } z. \end{cases}$$

B Derivation of SNA for Penalized Huber Loss Regression

Following section 2.3.1, denote $\mathcal{S}(z) = (S(z_1), \ldots, S(z_p))^\top$ and $d(\beta_0, \beta) = (h'_\gamma(y_1 - \beta_0 - x_1^\top \beta), \ldots, h'_\gamma(y_n - \beta_0 - x_n^\top \beta))^\top$, then the KKT conditions (2.3) can be written as (2.11).

Since the soft-thresholding operator is piecewise linear as shown in (2.8), we define

$$\begin{aligned} A &= \{j : |\beta_j + s_j| > 1\}, \\ B &= \{j : |\beta_j + s_j| \le 1\}. \end{aligned} \tag{B.1}$$

The set A works as an estimate for the support of β. In fact, if $(\widehat{s}, \widehat{\beta}_0, \widehat{\beta})$ satisfies the KKT conditions, then the set A defined on $(\widehat{\beta}, \widehat{s})$ is exactly the support for $\widehat{\beta}$. This is easy

to see: since $\widehat{s}_j \in \partial|\widehat{\beta}_j|$, if $\widehat{\beta}_j \neq 0$ then $|\widehat{\beta}_j + \widehat{s}_j| = |\widehat{\beta}_j + \mathrm{sgn}(\widehat{\beta}_j)| = |\widehat{\beta}_j| + 1 > 1$; otherwise, if $\widehat{\beta}_j = 0$ then $|\widehat{\beta}_j + \widehat{s}_j| = |\widehat{s}_j| \leq 1$.

We decompose β into β_A, β_B and s into s_A, s_B, and denote $Z = (s_A^\top, \beta_B^\top, \beta_0, \beta_A^\top, s_B^\top)^\top$. Then KKT conditions (2.11) can be rewritten as

$$F(Z) = \begin{bmatrix} \beta_A - \mathcal{S}(\beta_A + s_A) \\ \beta_B - \mathcal{S}(\beta_B + s_B) \\ -\frac{1}{n} 1^\top d \\ -\frac{1}{n} X_A^\top d + \lambda\alpha s_A + \lambda(1-\alpha)\beta_A \\ -\frac{1}{n} X_B^\top d + \lambda\alpha s_B + \lambda(1-\alpha)\beta_B \end{bmatrix} = \mathbf{0}. \tag{B.2}$$

And from (2.8) we have

$$\begin{cases} \beta_A - \mathcal{S}(\beta_A + s_A) = -s_A + \mathrm{sgn}(\beta_A + s_A), \\ \beta_B - \mathcal{S}(\beta_B + s_B) = \beta_B. \end{cases} \tag{B.3}$$

Let ψ_γ be as in (2.7), and for brevity denote

$$\Psi = \Psi(\beta_0, \beta) = \frac{1}{n} \mathrm{diag}(\psi_\gamma(\, y_1 - \beta_0 - x_1^\top \beta), \ldots, \psi_\gamma(y_n - \beta_0 - x_n^\top \beta)). \tag{B.4}$$

Then the following result gives a Newton derivative of $F(Z)$.

Theorem B.1. *$F(Z)$ is Newton differentiable for any $Z \in \mathbb{R}^{2p+1}$ and*

$$\begin{aligned} H(Z) \;\coloneqq\; & \begin{bmatrix} -I_{|A|} & \mathbf{0} & 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & I_{|B|} & 0 & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_n^\top \Psi X_B & \mathbf{1}_n^\top \Psi \mathbf{1}_n & \mathbf{1}_n^\top \Psi X_A & \mathbf{0} \\ \lambda\alpha I_{|A|} & X_A^\top \Psi X_B & X_A^\top \Psi \mathbf{1}_n & X_A^\top \Psi X_A + \lambda(1-\alpha) I_{|A|} & \mathbf{0} \\ \mathbf{0} & X_B^\top \Psi X_B + \lambda(1-\alpha) I_{|B|} & X_B^\top \Psi \mathbf{1}_n & X_B^\top \Psi X_A & \lambda\alpha I_{|B|} \end{bmatrix} \\ \in \;& \nabla_N F(Z). \end{aligned} \tag{B.5}$$

Furthermore, for any $\gamma > 0$ and $\alpha \in (0,1)$, on the set $\{Z = (s, \beta_0, \beta) : \;$ there exists $i \in \{1, \ldots, n\}$ such that $|y_i - \beta_0 - x_i^\top \beta| \leq \gamma\}$, $H(Z)$ is invertible and $H(Z)^{-1}$ is uniformly bounded in spectral norm.

From Theorems 2.1 and B.1, we immediately obtain the following result.

Theorem B.2. *Given $\lambda, \gamma, \alpha \in (0,1)$, define Z and $F(Z)$ as (B.2). Suppose $\widehat{Z}$ solves $F(Z) = 0$ and there exists a neighborhood $\mathcal{N}(\widehat{Z})$ such that for any $Z \in \mathcal{N}(\widehat{Z})$ there is an $i \in \{1, \ldots, n\}$ that satisfies $|y_i - \beta_0 - x_i^\top \beta| \leq \gamma$, then the Newton-type iteration*

$$Z^{k+1} = Z^k - H(Z^k)^{-1} F(Z^k)$$

converges superlinearly to $\widehat{Z}$ provided that $\|Z^0 - \widehat{Z}\|_2$ is sufficiently small.

Now we describe the algorithm in details. The $(k+1)$-th iteration can be split into two steps:

1. Solve D_k from $H(Z^k)D_k = -F(Z^k)$;
2. Update $Z^{k+1} = Z^k + D^k$.

At the first glance, step 1 seems to involve inverting a $(2p+1)\times(2p+1)$ matrix, which is intractable in high dimensional settings. However, the definitions of sets A, B in (B.1) motivate an "active set strategy" for dimension reduction. Given the estimates from the kth iteration, define the active set A_k and its complement B_k by (2.12), $d_k = d(\beta_0^k, \beta^k)$, and $D_k = (D_{A_k}^{s\top}, D_{B_k}^{\beta\top}, D_0^{\beta_0}, D_{A_k}^{\beta\top}, D_{B_k}^{s\top})^\top$ corresponding to Z_k.

Now substituting these identities into step 1 and combining (B.3) we have

$$\begin{aligned}
D_{A_k}^s &= -s_{A_k}^k + \mathrm{sgn}(\beta_{A_k}^k + s_{A_k}^k),\\
D_{B_k}^\beta &= -\beta_{B_k}^k,\\
\begin{bmatrix} D_0^{\beta_0} \\ D_{A_k}^\beta \end{bmatrix} &= \begin{bmatrix} \mathbf{1}_n^\top \Psi_k \mathbf{1}_n & \mathbf{1}_n^\top \Psi_k X_{A_k} \\ X_{A_k}^\top \Psi_k \mathbf{1}_n & X_{A_k}^\top \Psi_k X_{A_k} + \lambda(1-\alpha) I_{|A_k|} \end{bmatrix}^{-1} \\
&\quad \begin{bmatrix} \frac{1}{n} 1^\top d_k + \mathbf{1}_n^\top \Psi_k X_{B_k} \beta_{B_k}^k \\ \frac{1}{n} X_{A_k}^\top d_k - \lambda(1-\alpha)\beta_{A_k}^k - \lambda\alpha \mathrm{sgn}(\beta_{A_k}^k + s_{A_k}^k) + X_{A_k}^\top \Psi_k X_{B_k} \beta_{B_k}^k \end{bmatrix},\\
D_{B_k}^s &= -s_{B_k}^k + \frac{1}{\lambda\alpha} X_{B_k}^\top \left(\frac{1}{n} d_k + \Psi_k X_{B_k} \beta_{B_k}^k - \Psi_k \mathbf{1}_n D_0^{\beta_0} + \Psi_k X_{A_k} D_{A_k}^\beta \right).
\end{aligned}$$

Combining steps 1 and 2, the $(k+1)$th iteration of SNA is carried out as follows:

(i) Update $s_{A_k}^{k+1}$ and $\beta_{B_k}^{k+1}$:

$$\begin{aligned}
s_{A_k}^{k+1} &= \mathrm{sgn}(\beta_{A_k}^k + s_{A_k}^k),\\
\beta_{B_k}^{k+1} &= \mathbf{0}.
\end{aligned}$$

(ii) Find the direction $D_0^{\beta_0}$ for the intercept β_0, and $D_{A_k}^\beta$ for the active coefficients β_{A_k}:

$$\begin{aligned}
\begin{bmatrix} D_0^{\beta_0} \\ D_{A_k}^\beta \end{bmatrix} &= \begin{bmatrix} \mathbf{1}_n^\top \Psi_k \mathbf{1}_n & \mathbf{1}_n^\top \Psi_k X_{A_k} \\ X_{A_k}^\top \Psi_k \mathbf{1}_n & X_{A_k}^\top \Psi_k X_{A_k} + \lambda(1-\alpha) I_{|A_k|} \end{bmatrix}^{-1} \\
&\quad \begin{bmatrix} \frac{1}{n} 1^\top d_k + \mathbf{1}_n^\top \Psi_k X_{B_k} \beta_{B_k}^k \\ \frac{1}{n} X_{A_k}^\top d_k - \lambda(1-\alpha)\beta_{A_k}^k - \lambda\alpha s_{A_k}^{k+1} + X_{A_k}^\top \Psi_k X_{B_k} \beta_{B_k}^k \end{bmatrix}.
\end{aligned}$$

(iii) Update the intercept, the active coefficients, and the inactive subgradients:

$$\begin{aligned}
\beta_0^{k+1} &= \beta_0^k + D_0^{\beta_0},\\
\beta_{A_k}^{k+1} &= \beta_{A_k}^k + D_{A_k}^\beta,\\
s_{B_k}^{k+1} &= \frac{1}{\lambda\alpha} X_{B_k}^\top \left(\frac{1}{n} d_k + \Psi_k X_{B_k} \beta_{B_k}^k - \Psi_k \mathbf{1}_n D_0^{\beta_0} + \Psi_k X_{A_k} D_{A_k}^\beta \right).
\end{aligned}$$

C Proofs

Here we give proofs of Theorems 3.3, 3.4 in the main text and Lemmas A.2, A.3 and Theorem B.1 in the appendices.

Proof of Theorem 3.3.

Proof. Without loss of generality, assume θ_k has exactly one cluster point $\theta^\star$, i.e. $\theta_k \to \theta^\star$. Notice that

$$|t| - \frac{\gamma}{2} \le h_\gamma(t) \le |t|,$$

hence

$$f_A(\theta;\lambda) - \frac{\gamma}{2} \le f_H(\theta;\lambda,\gamma) \le f_A(\theta;\lambda).$$

Let $\widehat{\theta}_A$ be a minimizer of $f_A(\theta;\lambda)$, and $f_A^0 = \min_\theta f_A(\theta;\lambda) = f_A(\widehat{\theta}_A;\lambda)$, then

$$f_H(\theta_k;\lambda,\gamma_k) \le f_H(\widehat{\theta}_A;\lambda,\gamma_k) \le f_A(\widehat{\theta}_A;\lambda) = f_A^0.$$

For any $\epsilon > 0$, there exists K such that for $k \ge K$, $\gamma_k < 2\epsilon$, then

$$f_H(\theta_k;\lambda,\gamma_k) \ge f_A(\theta_k;\lambda) - \epsilon \ge f_A^0 - \epsilon.$$

Hence for $k \ge K$,

$$f_A^0 - \epsilon \le f_A(\theta_k;\lambda) - \epsilon \le f_A^0.$$

Let $k \to \infty$, we obtain $f_A^0 \le f_A(\theta^\star) \le f_A^0 + \epsilon$. Since ϵ is arbitrary, we have $f_A(\theta^\star) = f_A^0$. □

Proof of Theorem 3.4.

Proof. Without loss of generality, assume θ_k has exactly one cluster point $\theta^\star$, i.e. $\theta_k \to \theta^\star$. Notice that

$$\gamma h_\gamma(t) \le \frac{1}{2}t^2,$$

which implies

$$\gamma f_H(\theta;\lambda/\gamma,\gamma) \le f_S(\theta;\lambda).$$

Let $\widehat{\theta}_S$ be a minimizer of $f_S(\theta;\lambda)$, and $f_S^0 = \min_\theta f_S(\theta;\lambda) = f_S(\widehat{\theta}_S;\lambda)$, then

$$\gamma_k f_H(\theta_k;\lambda/\gamma_k,\gamma_k) \le \gamma_k f_H(\widehat{\theta}_S;\lambda/\gamma_k,\gamma_k) \le f_S(\widehat{\theta}_S;\lambda) = f_S^0.$$

Since $\theta_k = (\beta_0^k, \beta^k)$ is convergent, $r^k = y - \beta_0^k \mathbf{1} - X\beta^k$ is convergent too. Then there exists $M > 0$ such that $\|r^k\|_\infty \le M$. There exists K such that for $k \ge K$, $\gamma_k > M$, then $h_\gamma(r_i k) = \frac{1}{2} r_i k^2$, and

$$\gamma_k f_H(\theta_k;\lambda/\gamma_k,\gamma_k) = f_S(\theta_k;\lambda).$$

Hence for $k \geq K$,

$$f_S(\theta_k; \lambda) \leq f_S^0.$$

Let $k \to \infty$, we obtain $f_S(\theta^\star; \lambda) \leq f_S^0$. Since $f_S(\theta^\star; \lambda) \geq \min_\theta f_S(\theta; \lambda) = f_S^0$, we have $f_S(\theta^\star) = f_S^0$. □

Proof of Lemma A.2.

Proof. (i) By assumption, the Jacobian J_F is continuous at z. Since

$$\begin{aligned}
&\frac{\|F(z+h) - F(z) - J_F(z+h)h\|_2}{\|h\|_2} \\
\leq\ & \frac{\|F(z+h) - F(z) - J_F(z)h\|_2 + \|(J_F(z) - J_F(z+h))h\|_2}{\|h\|_2} \\
\leq\ & \frac{\|F(z+h) - F(z) - J_F(z)h\|_2}{\|h\|_2} + \|J_F(z) - J_F(z+h)\| \\
\to\ & 0
\end{aligned}$$

as $h \to \mathbf{0}$, by definition $J_F \in \nabla_N F(z)$.

(ii)

$$\|AF(z+h) - AF(z) - AH(z+h)h\|_2 \leq \|A\|\|F(z+h) - F(z) - H(z+h)h\|_2 = o(\|h\|_2),$$

hence $AH \in \nabla_N AF(z)$.

(iii)

$$\begin{aligned}
& \|(F(z+h) + G(z+h)) - (F(z) + G(z)) - (H_F(z+h) + H_G(z+h))h\|_2 \\
\leq\ & \|F(z+h) - F(z) - H_F(z+h)h\|_2 + \|G(z+h) - G(z) - H_G(z+h)h\|_2 \\
=\ & o(\|h\|_2),
\end{aligned}$$

hence $H_F + H_G \in \nabla_N (F+G)(z)$.

(iv) It can be seen by observing that

$$\|F(z+h) - F(z) - H(z+h)h\|_2^2 = \sum_{i=1}^{l} (F_i(z+h) - F_i(z) - H_i(z+h)h)^2.$$

□

Proof of Lemma A.3.

Proof. If f is differentiable at z with derivative f' defined in its neighborhood, by smoothness assumption and Lemma A.2(i), $f' \in \nabla_N f(z)$.

If f is not differentiable at z, by assumption there exists $s > 0$ such that f is smooth on both $(z-s, z)$ and $(z, z+s)$ implying that $f'(z-) = \lim_{h\to 0^-} \frac{f(z+h)-f(z)}{h}$ and $f'(z+) = \lim_{h\to 0^+} \frac{f(z+h)-f(z)}{h}$ exist and

$$\begin{aligned} f'(z+h) &\to f'(z-) \qquad \text{as } h \to 0^-, \\ f'(z+h) &\to f'(z+) \qquad \text{as } h \to 0^+. \end{aligned}$$

Hence for any $\varepsilon > 0$, there exists a sufficiently small $\delta > 0$ such that

$$\begin{aligned} &\forall\, x \in (z-\delta, z), \quad \tfrac{|f(x)-f(z)-f'(z-)(x-z)|}{|x-z|} < \varepsilon/2, \quad |f'(x) - f'(z-)| < \varepsilon/2; \\ &\forall\, x \in (z, z+\delta), \quad \tfrac{|f(x)-f(z)-f'(z+)(x-z)|}{|x-z|} < \varepsilon/2, \quad |f'(x) - f'(z+)| < \varepsilon/2. \end{aligned}$$

Thus for $x \in (z-\delta, z)$,

$$\frac{|f(x)-f(z)-f'(x)(x-z)|}{|x-z|} \le \frac{|f(x)-f(z)-f'(z-)(x-z)|}{|x-z|} + |f'(z-) - f'(x)| < \varepsilon,$$

and similarly for $x \in (z, z+\delta)$. Define $H(z)$ as in the lemma, then the above implies

$$\forall \varepsilon > 0, \exists \delta > 0 \text{ s.t. } \forall |x-\delta| < z, \ \frac{|f(x)-f(z)-H(x)(x-z)|}{|x-z|} < \epsilon.$$

In other word, f is Newton differentiable at z with $H \in \nabla_N f(z)$.

□

Proof of Theorem B.1.

Proof. Notice that the derivative of Huber loss h'_γ is piecewise-smooth, which implies by Lemma A.3 that $\psi_\gamma \in \nabla_N h'_\gamma(t), \forall t \in \mathbb{R}$. As shown in (2.8), the soft-thresholding operator is also piecewise-smooth. Hence by Lemma A.1, A.2 and A.3, it is easy to show $F(Z)$ is Newton differentiable with a Newton derivative $H(Z)$ taking the form of (B.5).

Next we show $H = H(Z)$ is invertible with its inverse H^{-1} bounded in spectral norm. Denote

$$H = \begin{bmatrix} H_1 & \mathbf{0} \\ H_2 & H_3 \end{bmatrix},$$

where

$$H_1 = \begin{bmatrix} -I_{|A|} & \mathbf{0} \\ \mathbf{0} & I_{|B|} \end{bmatrix} \in \mathbb{R}^{p\times p}, \quad H_2 = \begin{bmatrix} \mathbf{0} & \mathbf{1}_n^\top \Psi X_B \\ \lambda\alpha I_{|A|} & X_A^\top \Psi X_B \\ \mathbf{0} & X_B^\top \Psi X_B + \lambda(1-\alpha) I_{|B|} \end{bmatrix} \in \mathbb{R}^{(p+1)\times p},$$

$$H_3 = \begin{bmatrix} \mathbf{1}_n^\top \Psi \mathbf{1}_n & \mathbf{1}_n^\top \Psi X_A & \mathbf{0} \\ X_A^\top \Psi \mathbf{1}_n & X_A^\top \Psi X_A + \lambda(1-\alpha) I_{|A|} & \mathbf{0} \\ X_B^\top \Psi \mathbf{1}_n & X_B^\top \Psi X_A & \lambda\alpha I_{|B|} \end{bmatrix} \in \mathbb{R}^{(p+1)\times(p+1)}. \tag{C.1}$$

Clearly, H_1 is invertible. From Lemma C.1 to be stated later, H_3 is also invertible. Then via some algebra we have

$$H^{-1} = \begin{bmatrix} H_1^{-1} & \mathbf{0} \\ -H_3^{-1} H_2 H_1^{-1} & H_3^{-1} \end{bmatrix}. \tag{C.2}$$

Let $g = (g_1^\top, g_2^\top)^\top \in \mathbb{R}^p \times \mathbb{R}^{p+1}$, then

$$\begin{aligned} \|H^{-1}g\|_2^2 &= \|H_1^{-1}g_1\|_2^2 + \| - H_3^{-1}H_2H_1^{-1}g_1 + H_3^{-1}g_2\|_2^2 \\ &\le \|H_1^{-1}\|^2\|g_1\|_2^2 + (\|H_3^{-1}\|\|H_2\|\|H_1^{-1}\|\|g_1\|_2 + \|H_3^{-1}\|\|g_2\|_2)^2 \\ &\le (\|H_1^{-1}\|\|g_1\|_2 + \|H_3^{-1}\|\|H_2\|\|H_1^{-1}\|\|g_1\|_2 + \|H_3^{-1}\|\|g_2\|_2)^2 \\ &\le (\|H_1^{-1}\| + \|H_3^{-1}\| + \|H_3^{-1}\|\|H_2\|\|H_1^{-1}\|)^2\|g\|_2^2 \end{aligned} \tag{C.3}$$

which implies

$$\|H^{-1}\| \le \|H_1^{-1}\| + \|H_3^{-1}\| + \|H_3^{-1}\|\|H_2\|\|H_1^{-1}\|. \tag{C.4}$$

Notice $\|X_A\| \vee \|X_B\| \le \|X\|$. Taking X_A for example, without loss of generality shuffle columns of X such that $X = (X_A|X_B)$, then for any $g \in \mathbb{R}^{|A|}$ such that $\|g\|_2 = 1$, we have

$$\|X_A g\|_2 = \|X \begin{bmatrix} g \\ \mathbf{0} \end{bmatrix}\|_2 \le \sup\{\|Xv\|_2 : \ \|v\|_2 = 1\} = \|X\|,$$

implying that $\|X_A\| \le \|X\|$. Similarly, we can show $\|X_B\| \le \|X\|$.

In addition, with a similar argument as (C.3) for H_2, we have

$$\|H_2\| \le 1 + \alpha + 2\|X\|^2. \tag{C.5}$$

Note that $\|H_1^{-1}\| = 1$. Combining (C.4), (C.5) with Lemma C.1, we obtain the uniform boundedness of H in spectral norm, i.e.,

$$\begin{aligned} \|H^{-1}\| \ \le \ & 1 + \left[\frac{1}{\lambda\alpha} + \left(\frac{1}{\lambda(1-\alpha)} + \frac{\lambda_{\max}(X^\top X)^2 + n\gamma\lambda(1-\alpha)}{\lambda(1-\alpha)}\left(1 + \frac{\|X\|}{\sqrt{n}\gamma\lambda(1-\alpha)}\right)^2\right)\right. \\ & \left. \times \left(1 + \frac{2\|X\|}{\sqrt{n}\gamma\lambda\alpha}\right)\right](2 + \alpha + 2\|X\|^2). \end{aligned}$$

□

In order to complete the proof of Theorem B.1, we need the following lemma.

Lemma C.1. *Given $\alpha \in (0,1)$ and β_0, β satisfy $|y_i - \beta_0 - x_i^\top \beta| \le \gamma$ for some i, then H_3 in (C.1) is invertible with its inverse uniformly bounded in spectral norm, i.e.*

$$\|H_3^{-1}\| \le \frac{1}{\lambda\alpha} + \left[\frac{1}{\lambda(1-\alpha)} + \frac{\lambda_{\max}(X^\top X)^2 + n\gamma\lambda(1-\alpha)}{\lambda(1-\alpha)}\left(1 + \frac{\|X\|}{\sqrt{n}\gamma\lambda(1-\alpha)}\right)^2\right]\left(1 + \frac{2\|X\|}{\sqrt{n}\gamma\lambda\alpha}\right).$$

Proof. Denote $J = n\gamma\Psi$, then J is diagonal and idempotent. We have

$$\mathbf{1}_n^\top \Psi \mathbf{1}_n = \frac{1}{n\gamma}\mathbf{1}_n^\top J \mathbf{1}_n = \frac{1}{n\gamma}(J\mathbf{1}_n)^\top (J\mathbf{1}_n),$$

and

$$\begin{aligned}
&\mathbf{1}_n^\top \Psi X_A \left(X_A^\top \Psi X_A + \lambda(1-\alpha) I_{|A|}\right)^{-1} X_A^\top \Psi \mathbf{1}_n \\
&= \frac{1}{n\gamma}(J\mathbf{1}_n)^\top (JX_A)\left((JX_A)^\top (JX_A) + n\gamma\lambda(1-\alpha) I_{|A|}\right)^{-1} (JX_A)^\top (J\mathbf{1}_n).
\end{aligned}$$

Denote $a = J\mathbf{1}_n$, $Z = JX_A$, $t = n\gamma\lambda(1-\alpha)$, and $m = |A|$. Then the LHS becomes

$$\frac{1}{n\gamma}\left(a^\top a - a^\top Z(Z^\top Z + tI_m)^{-1} Z^\top a\right).$$

Since $|y_i - \beta_0 - x_i^\top \beta| \le \gamma$ for some i, we have $\psi_i = \frac{1}{n\gamma} > 0$, implying that $J_{ii} = 1$ and $a^\top a \ge J_{ii}^2 = 1$. Thus we are guaranteed that $a = J\mathbf{1}_n$ is not a zero vector.

Now apply SVD to Z such that $Z = UDV^\top$, where $U_{n\times n}$ and $V_{m\times m}$ are both orthogonal matrices, and $D_{n\times m}$ is a rectangular diagonal matrix with non-negative diagonal elements $d_1, \ldots, d_{m\wedge n}$. Hence

$$\begin{aligned}
Z(Z^\top Z + tI_m)^{-1} Z^\top &= UDV^\top (VD^\top U^\top UDV^\top + tI_m)^{-1} VD^\top U^\top \\
&= UDV^\top \left(V(D^\top D + tI_m)V^\top\right)^{-1} VD^\top U^\top \\
&= UDV^\top V(D^\top D + tI_m)^{-1} V^\top VD^\top U^\top \\
&= UD(D^\top D + tI_m)^{-1} D^\top U^\top.
\end{aligned}$$

When $n > m$,

$$D(D^\top D + tI_m)^{-1} D^\top = \mathrm{diag}\left(\frac{d_1^2}{d_1^2 + t}, \ldots, \frac{d_m^2}{d_m^2 + t}, 0, \ldots, 0\right),$$

and when $n \le m$,

$$D(D^\top D + tI_m)^{-1} D^\top = \mathrm{diag}\left(\frac{d_1^2}{d_1^2 + t}, \ldots, \frac{d_n^2}{d_n^2 + t}\right).$$

In either case $D(D^\top D + tI_m)^{-1} D^\top$ is p.s.d. with $\lambda_{\max}(D(D^\top D + tI_m)^{-1} D^\top) < 1$.

Next we will derive the upper bound of eigenvalues of the above matrix. First, for any eigenvalue d and corresponding nonzero eigenvector u of $Z^\top Z = X_A J X_A$, we have

$$du^\top u = u^\top X_A^\top J X_A u \leq \lambda_{\max}(X^\top J X) u^\top u,$$

hence $d \leq \lambda_{\max}(X^\top J X)$. Then again, for any eigenvalue c and corresponding nonzero eigenvector v of $X^\top J X$, we have

$$cv^\top v = v^\top X^\top J X v = \sum_i J_{ii} v^\top x_i x_i^\top v \leq \sum_i v^\top x_i x_i^\top v = v^\top X^\top X v \leq \lambda_{\max}(X^\top X) v^\top v,$$

implying that $c \leq \lambda_{\max}(X^\top X)$.

Therefore, we have $d \leq \lambda_{\max}(X^\top J X) \leq \lambda_{\max}(X^\top X)$. Then since the eigenvalues of $Z^\top Z$ are the diagonal elements of D, the eigenvalues of $D(D^\top D + tI_m)^{-1} D^\top$ are bounded by $\frac{\lambda_{\max}(X^\top X)^2}{\lambda_{\max}(X^\top X)^2 + t}$.

Then recall $t = n\gamma\lambda(1-\alpha)$ and $a^\top a \geq 1$, we have

$$\begin{aligned}
&\mathbf{1}_n^\top \Psi \mathbf{1}_n - \mathbf{1}_n^\top \Psi X_A \left(X_A^\top \Psi X_A + \lambda(1-\alpha) I_{|A|} \right)^{-1} X_A^\top \Psi \mathbf{1}_n \\
&= \frac{1}{n\gamma} \left(a^\top a - (U^\top a)^\top D (D^\top D + tI_m)^{-1} D^\top (U^\top a) \right) \\
&\geq \frac{1}{n\gamma} (a^\top a - \frac{\lambda_{\max}(X^\top X)^2}{\lambda_{\max}(X^\top X)^2 + t} (U^\top a)^\top U^\top a) \\
&= \frac{1}{n\gamma} \times \frac{n\gamma\lambda(1-\alpha)}{\lambda_{\max}(X^\top X)^2 + n\gamma\lambda(1-\alpha)} a^\top a \\
&\geq \frac{\lambda(1-\alpha)}{\lambda_{\max}(X^\top X)^2 + n\gamma\lambda(1-\alpha)} \\
&> 0.
\end{aligned}$$

Let

$$H_{31} = \begin{bmatrix} \mathbf{1}_n^\top \Psi \mathbf{1}_n & \mathbf{1}_n^\top \Psi X_A \\ X_A^\top \Psi \mathbf{1}_n & X_A^\top \Psi X_A + \lambda(1-\alpha) I_{|A|} \end{bmatrix}, \; H_{32} = \begin{bmatrix} X_B^\top \Psi \mathbf{1}_n & X_B^\top \Psi X_A \end{bmatrix}, \; H_{33} = \lambda\alpha I_{|B|}.$$

Observe that $H_{33}^{-1} = \frac{1}{\lambda\alpha} I_{|B|}$. Then if H_{31} is invertible, we have

$$H_3^{-1} = \begin{bmatrix} H_{31}^{-1} & \mathbf{0} \\ -\frac{1}{\lambda\alpha} H_{32} H_{31}^{-1} & \frac{1}{\lambda\alpha} I_{|B|} \end{bmatrix}.$$

Hence to show H_3 is invertible, it suffices to show H_{31} is invertible. Let

$$M = X_A^\top \Psi X_A + \lambda(1-\alpha) I_{|A|}, \; b = X_A^\top \Psi \mathbf{1}_n,$$

and

$$\kappa = \mathbf{1}_n^\top \Psi \mathbf{1}_n - \mathbf{1}_n^\top \Psi X_A \left(X_A^\top \Psi X_A + \lambda(1-\alpha) I_{|A|} \right)^{-1} X_A^\top \Psi \mathbf{1}_n.$$

Since $\kappa > 0$, we have

$$H_{31}^{-1} = \begin{bmatrix} \frac{1}{\kappa} & -\frac{1}{\kappa} b^\top M^{-1} \\ -\frac{1}{\kappa} M^{-1} b & M^{-1} + \frac{1}{\kappa} M^{-1} b b^\top M^{-1} \end{bmatrix},$$

and it follows that H_3 is invertible.

It can be easily shown that $\|b\| = \|b^\top\| \leq \frac{1}{\sqrt{n\gamma}} \|X\|$, $\|M^{-1}\| \leq \frac{1}{\lambda(1-\alpha)}$. Combine this with $\frac{1}{\kappa} \leq \frac{\lambda_{\max}(X^\top X)^2 + n\gamma\lambda(1-\alpha)}{\lambda(1-\alpha)}$, then similar to (C.3), we have

$$\|H_{31}^{-1}\| \leq \frac{1}{\lambda(1-\alpha)} + \frac{\lambda_{\max}(X^\top X)^2 + n\gamma\lambda(1-\alpha)}{\lambda(1-\alpha)} \left(1 + \frac{\|X\|}{\sqrt{n}\gamma\lambda(1-\alpha)}\right)^2$$

and then

$$\|H_3^{-1}\| \leq \frac{1}{\lambda\alpha} + \left[\frac{1}{\lambda(1-\alpha)} + \frac{\lambda_{\max}(X^\top X)^2 + n\gamma\lambda(1-\alpha)}{\lambda(1-\alpha)} \left(1 + \frac{\|X\|}{\sqrt{n}\gamma\lambda(1-\alpha)}\right)^2\right] \left(1 + \frac{2\|X\|}{\sqrt{n}\gamma\lambda\alpha}\right).$$

□