Supplementary Material

Wenzhuo Yang A0096049@NUS.EDU.SG

Department of Mechanical Engineering, National University of Singapore, Singapore 117576

Huan Xu MPEXUH@NUS.EDU.SG

Department of Mechanical Engineering, National University of Singapore, Singapore 117576

1. Notation Table

$[m]$	The set $\{1, \cdots, m\}$
g	A subset of $[m]$
g^c	The complement of g, $g^c = [m] \setminus g$
$\mathbf{I}$	The identity matrix
$\mathbf{X}$	The sample matrix $\mathbf{X} \in \mathcal{R}^{n \times m}$
$\mathbf{X}_i$	The ith column of $\mathbf{X}$
$\boldsymbol{\beta}$	Vector $\boldsymbol{\beta} \in \mathcal{R}^m$
β_i	The ith element of $\boldsymbol{\beta}$
$\boldsymbol{\beta}_g$	The vector whose ith element is β_i if $i \in g$ or 0 otherwise
$\boldsymbol{\Delta}^{(i)}$	The ith disturbance matrix
$\boldsymbol{\Delta}_j^{(i)}$	The jth column of $\boldsymbol{\Delta}^{(i)}$
$\boldsymbol{\Delta}_g$	The matrix whose ith column is $\boldsymbol{\Delta}_i$ if $i \in g$ or 0 otherwise
$\mathbf{W}_i$	Matrix $\mathbf{W}_i \in \mathcal{R}^{m \times m}$
$\text{vec}(\cdot)$	The operator vectorizing a matrix by stacking its columns
$\|\mathbf{X}\|_p$	The ℓ_p-norm of $\text{vec}(\mathbf{X})$, $\|\text{vec}(\mathbf{X})\|_p$

2. Proofs in Section 2

To prove the corollaries in Section 2, we give the following lemma.

Lemma 1. *If any two different groups g_p and g_q in G_i in the uncertainty set U (4) are non-overlapping for $i = 1, \cdots, t$, which means $g_p \cap g_q = \emptyset$, then the optimization problem (5) is equivalent to*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^{t} \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^*\} \tag{1}$$

Proof. Since any two different groups g_p and g_q in G_i are non-overlapping, we have

$$\sum_{i=1}^{t} \max_{\forall g \in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \le c_g} {\boldsymbol{\alpha}^{(i)}}^\top \mathbf{W}_i \boldsymbol{\beta} = \sum_{i=1}^{t} \sum_{g \in G_i} \max_{\|\boldsymbol{\alpha}_g^{(i)}\|_p \le c_g} \boldsymbol{\alpha}_g^{(i)} (\mathbf{W}_i \boldsymbol{\beta})_g = \sum_{i=1}^{t} \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* \tag{2}$$

Hence the lemma holds. □

By using Theorem 3 and Lemma 1, we have

1. *Proof of Corollary 1:* $G_1 = \{[m]\}$ satisfies the condition of Lemma 1, so we have

$$\sum_{i=1}^{t} \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* = \sum_{g \in G_1} c \|\boldsymbol{\beta}_g\|_2^* = c\|\beta\|_2. \tag{3}$$

2. *Proof of Corollary 2:* $G_1 = \{\{1\}, \cdots, \{m\}\}$ satisfies the condition of Lemma 1, then

$$\sum_{i=1}^{t} \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* = \sum_{g \in G_1} c_g \|\boldsymbol{\beta}_g\|_p^* = \sum_{i=1}^{m} c_i |\beta_i|. \tag{4}$$

3. *Proof of Corollary 3:* $G_1 = \{g_1, \cdots, g_k\}$ satisfies the condition of Lemma 1, so we have

$$\sum_{i=1}^{t} \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* = \sum_{i=1}^{k} c_{g_i} \|\boldsymbol{\beta}_{g_i}\|_p^*. \tag{5}$$

4. *Proof of Theorem 2:* $G_i = \{g_i, g_i^c\}$ satisfies the condition of Lemma 1 and $c_{g_i^c} = 0$, so that

$$\sum_{i=1}^{t} \sum_{g \in G_i} c_g \|(\mathbf{W}_i \boldsymbol{\beta})_g\|_p^* = \sum_{i=1}^{k} (c_{g_i} \|\boldsymbol{\beta}_{g_i}\|_p^* + c_{g_i^c} \|\boldsymbol{\beta}_{g_i^c}\|_p^*) = \sum_{i=1}^{k} c_{g_i} \|\boldsymbol{\beta}_{g_i}\|_p^*. \tag{6}$$

5. *Proof of Corollary 4:* The dual problem of the optimization problem

$$\min_{\sum \mathbf{v}_{g_i} = \boldsymbol{\beta},\ \mathrm{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \sum_{i=1}^{k} c_{g_i} \|\mathbf{v}_{g_i}\|_p^*$$

can be formulated as

$$\begin{aligned}
&\max_{\boldsymbol{\alpha}} \min_{\forall i, \mathrm{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \{\sum_{i=1}^{k} c_{g_i} \|\mathbf{v}_{g_i}\|_p^* - \boldsymbol{\alpha}^\top \sum_{i=1}^{k} \mathbf{v}_{g_i} + \boldsymbol{\alpha}^\top \boldsymbol{\beta}\} \\
=& \max_{\boldsymbol{\alpha}} \{\boldsymbol{\alpha}^\top \boldsymbol{\beta} + \min_{\forall i, \mathrm{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \{\sum_{i=1}^{k} c_{g_i} \|\mathbf{v}_{g_i}\|_p^* - \boldsymbol{\alpha}_{g_i}^\top \mathbf{v}_{g_i}\}\} \\
=& \max_{\boldsymbol{\alpha}} \{\boldsymbol{\alpha}^\top \boldsymbol{\beta} - \max_{\forall i, \mathrm{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \{\sum_{i=1}^{k} \boldsymbol{\alpha}_{g_i}^\top \mathbf{v}_{g_i} - c_{g_i} \|\mathbf{v}_{g_i}\|_p^*\}\} \\
=& \max_{\forall i, \|\boldsymbol{\alpha}_{g_i}\| \leq c_{g_i}} \boldsymbol{\alpha}^\top \boldsymbol{\beta}
\end{aligned} \tag{7}$$

Since the constraints in the primal problem satisfy Slater's condition, the strong duality holds. From the duality and the condition in Corollary 4, we have

$$\begin{aligned}
&\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^{t} \max_{\forall g \in G_i, \|\boldsymbol{\alpha}_g^{(i)}\|_p \leq c_g} \boldsymbol{\alpha}^{(i)\top} \mathbf{W}_i \boldsymbol{\beta}\} \\
=& \min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \max_{\forall g \in G_1, \|\boldsymbol{\alpha}_g\|_p \leq c_g} \boldsymbol{\alpha}^\top \boldsymbol{\beta}\} \\
=& \min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \min_{\sum \mathbf{v}_{g_i} = \boldsymbol{\beta},\ \mathrm{supp}(\mathbf{v}_{g_i}) \subseteq g_i} \sum_{i=1}^{k} c_{g_i} \|\mathbf{v}_{g_i}\|_p^*\}.
\end{aligned} \tag{8}$$

6. *Proof of Corollary 5:* From Theorem 2 and Lemma 1, we have

$$\begin{aligned}&\sum_{i=1}^{t}\sum_{g\in G_i} c_g\|(\mathbf{W}_i\boldsymbol{\beta})_g\|_p^* \\ =&\sum_{g\in G_1} c_g\|\boldsymbol{\beta}_g\|_p^* + \sum_{g\in G_2} c_g'\|(\mathbf{W}_2\boldsymbol{\beta})_g\|_p^* \\ =&\sum_{i=1}^{m} c_i|\beta_i| + \sum_{i=1}^{m-1} c_i'|\beta_i-\beta_{i+1}|.\end{aligned} \tag{9}$$

7. *Proof of Corollary 6:* By using the proofs of Corollary 1 and Corollary 3, we can obtain Corollary 6.

8. *Proof of Corollary 7:* $G_1 = \{\{1\},\cdots,\{m\}\}$ satisfies the condition of Lemma 1. Since $t = 1$, $c_{\{i\}} = \lambda$ and $\mathbf{W}_1 = \mathbf{D}$, we have

$$\sum_{i=1}^{t}\sum_{g\in G_i} c_g\|(\mathbf{W}_i\boldsymbol{\beta})_g\|_p^* = \sum_{g\in G_1}\lambda\|(\mathbf{D}\boldsymbol{\beta})_g\|_p^* = \sum_{i=1}^{m}\lambda|(\mathbf{D}\boldsymbol{\beta})_i| = \lambda\|\mathbf{D}\boldsymbol{\beta}\|_1. \tag{10}$$

3. Proofs in Section 3

3.1. Proof of Theorem 4:

From the definition of $\hat{U}$, we have

$$\begin{aligned}&\max_{\boldsymbol{\Delta}\in\hat{U}}\|\mathbf{y}-(\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_p \\ =&\max_{\mathbf{c}\in Z}\max_{\forall i,\forall g\in G_i,\|\boldsymbol{\Delta}_g^{(i)}\|_p\le c_g}\|\mathbf{y}-(\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_p \\ =&\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_p + \max_{\mathbf{c}\in Z}\sum_{i=1}^{t}\max_{\forall g\in G_i,\|\boldsymbol{\alpha}_g^{(i)}\|_p\le c_g}{\boldsymbol{\alpha}^{(i)}}^\top\mathbf{W}_i\boldsymbol{\beta} \\ =&\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_p + \max_{\mathbf{c}|\mathbf{c}\ge0;f_i(\mathbf{c})\le0}\sum_{i=1}^{t}\max_{\forall g\in G_i,\|\boldsymbol{\alpha}_g^{(i)}\|_p\le c_g}{\boldsymbol{\alpha}^{(i)}}^\top\mathbf{W}_i\boldsymbol{\beta} \\ =&\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_p + \min_{\boldsymbol{\lambda}\in\mathcal{R}_+^q,\boldsymbol{\kappa}\in\mathcal{R}_+^k}\max_{\mathbf{c}\in\mathcal{R}^k}\{\sum_{i=1}^{t}\max_{\forall g\in G_i,\|\boldsymbol{\alpha}_g^{(i)}\|_p\le c_g}{\boldsymbol{\alpha}^{(i)}}^\top\mathbf{W}_i\boldsymbol{\beta}+\boldsymbol{\kappa}^\top\mathbf{c}-\sum_{i=1}^{q}\lambda_i f_i(\mathbf{c})\} \\ =&\|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_p + \min_{\boldsymbol{\lambda}\in\mathcal{R}_+^q,\boldsymbol{\kappa}\in\mathcal{R}_+^k}\upsilon(\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{\beta})\end{aligned} \tag{11}$$

Hence we establish the theorem by taking minimum over $\boldsymbol{\beta}$ on both sides. Now we show the optimization problem is convex and tractable. we first prove that $\upsilon(\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{\beta})$ is a convex function of $\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{\beta}$. Since

$$\upsilon(\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{\beta}) = \max_{\substack{\mathbf{c}\in\mathcal{R}^k,\\ \forall i,\, g\in G_i,\, \|\boldsymbol{\alpha}_g^{(i)}\|_p\le c_g}}\{\sum_{i=1}^{t}{\boldsymbol{\alpha}^{(i)}}^\top\mathbf{W}_i\boldsymbol{\beta}+\boldsymbol{\kappa}^\top\mathbf{c}-\sum_{i=1}^{q}\lambda_i f_i(\mathbf{c})\} = \max_{\substack{\mathbf{c}\in\mathcal{R}^k,\\ \forall i,\, g\in G_i,\, \|\boldsymbol{\alpha}_g^{(i)}\|_p\le c_g}}\mu(\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{\beta}). \tag{12}$$

For fixed $\mathbf{c}$ and $\boldsymbol{\alpha}_g^{(i)}$, $\mu(\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{\beta})$ is a linear function of $\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{\beta}$. Thus $\upsilon(\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{\beta})$ is convex, which implies the optimization problem is convex. By choosing parameter γ, the optimization problem can be reformulated as

$$\begin{aligned}\min\quad & \|\mathbf{y}-\mathbf{X}\boldsymbol{\beta}\|_p \\ \text{s.t.}\quad & \upsilon(\boldsymbol{\lambda},\boldsymbol{\kappa},\boldsymbol{\beta})\le\gamma \\ & \boldsymbol{\lambda}\in\mathcal{R}_+^p,\boldsymbol{\kappa}\in\mathcal{R}_+^k,\boldsymbol{\beta}\in\mathcal{R}^m\end{aligned}$$

To show the problem is tractable, it suffices to construct a polynomial-time *separation oracle* for the feasible set S (Grötschel et al. (Grötschel et al., 1988)). A separation oracle is a routine such that for a solution $(\boldsymbol{\lambda}_0,\boldsymbol{\kappa}_0,\boldsymbol{\beta}_0)$, it can find, in polynomial time, that (a) whether $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0)$ belongs to S or not; and (b) if $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0) \notin S$, a hyperplane that separates $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0)$ with S.

To verify the feasibility of $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0)$, notice that $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0) \in S$ if and only if the optimal value of the optimization problem (12) is smaller than or equal to γ, which can be verified in polynomial time. If $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0) \notin S$, then by solving (12), we can find in polynomial time $\mathbf{c}_0, \boldsymbol{\alpha}_0^{(i)}$ such that

$$\sum_{i=1}^{t} {\boldsymbol{\alpha}_0^{(i)}}^\top \mathbf{W}_i \boldsymbol{\beta} + \boldsymbol{\kappa}^\top \mathbf{c}_0 - \sum_{i=1}^{q} \lambda_i f_i(\mathbf{c}_0) > \gamma.$$

which is the hyperplane separates $(\boldsymbol{\lambda}_0, \boldsymbol{\kappa}_0, \boldsymbol{\beta}_0)$ with S.

3.2. Extension of Corollary 8:

Theorem 1. *Let $g_1, \cdots, g_t$ be t groups such that $\bigcup_{i=1}^t g_i = [m]$, and $\bar{\boldsymbol{\Delta}}_i$ be a $n \times m$ matrix whose columns except the ith one are all zero. Suppose that $\mathbf{c}_{g_i}$ is a $|g_i|$ dimension vector whose elements give the norm bound of $\bar{\boldsymbol{\Delta}}_j$ for $j \in g_i$, e.g. $\|\bar{\boldsymbol{\Delta}}_j\|_2 \le c_{g_i}^j$, and $\mathbf{c} = (\mathbf{c}_{g_1}, \cdots, \mathbf{c}_{g_t})$. We define the uncertainty set as $\hat{U} = \{\sum_{i=1}^t \sum_{j \in g_i} \bar{\boldsymbol{\Delta}}_j | \exists \mathbf{c}$ such that $\mathbf{c} \ge 0$ and $\|\mathbf{c}_{g_i}\|_q^* \le s_i, \forall i \in [t]; \|\bar{\boldsymbol{\Delta}}_j\|_2 \le c_{g_i}^j, \forall i \in [t], \forall j \in g_i\}$, then the equivalent linear regularized regression problem is*

$$\min_{\boldsymbol{\beta} \in \mathcal{R}^m} \{\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_p + \sum_{i=1}^t s_i \|\boldsymbol{\beta}_{g_i}\|_q\},$$

where $\|\cdot\|_q^*$ is the dual norm of $\|\cdot\|_q$.

Proof. From Theorem 3 and Theorem 4, we have

$$\begin{aligned} &\min_{\lambda \in \mathcal{R}_+, \boldsymbol{\kappa} \in \mathcal{R}_+^m} \upsilon(\lambda, \boldsymbol{\kappa}, \boldsymbol{\beta}) \\ =& \min_{\lambda \in \mathcal{R}_+, \boldsymbol{\kappa} \in \mathcal{R}_+^m} \max_{\mathbf{c} \in \mathcal{R}^m} \{\sum_{j=1}^t \sum_{i \in g_j} (\kappa_i + |\beta_i|) c_i - \\ &\sum_{i=1}^t \lambda_i (\|\mathbf{c}_{g_i}\|_q^* + s_i)\}. \end{aligned}$$

Define $\mathbf{r}_{g_i}$ as the vector whose elements are $\kappa_j + |\beta_j|$ for $j \in g_i$, then the equation above is equivalent to

$$\min_{\lambda \in \mathcal{R}_+, \boldsymbol{\kappa} \in \mathcal{R}_+^m | \|\mathbf{r}_{g_i}\|_q \le \lambda_i, \forall i in [t]} \boldsymbol{\lambda}^\top \mathbf{s} = \sum_{i=1}^t s_i \|\boldsymbol{\beta}_{g_i}\|_q,$$

which establishes the theorem. □

4. Proofs in Section 5

Recall that the uncertainty set considered in this paper is

$$U = \{\boldsymbol{\Delta}^{(1)} \mathbf{W}_1 + \cdots + \boldsymbol{\Delta}^{(t)} \mathbf{W}_t | \forall i, \forall g \in G_i, \|\boldsymbol{\Delta}_g^{(i)}\|_2 \le c_g\} \tag{13}$$

where G_i is the set of the groups of $\boldsymbol{\Delta}^{(i)}$ and c_g gives the bound of $\boldsymbol{\Delta}_g^{(i)}$ for group g. We denote $\bar{G}_i$ and $\bar{G}_i^c$ as the set $\{g \in G_i | c_g \ne 0\}$ and $G_i - \bar{G}_i$, respectively. In this theorem, we restrict our discussion to the case that $\mathbf{W}_i = \mathbf{I}$ for $i = 1, \cdots, t$ and the bound c_g of $\boldsymbol{\Delta}_g^{(i)}$ for each group g equals $\sqrt{n} c_n$ or 0, so the uncertainty set can be rewritten as

$$U = \{\boldsymbol{\Delta}^{(1)} + \cdots + \boldsymbol{\Delta}^{(t)} | \forall i, \forall g \in \bar{G}_i, \|\boldsymbol{\Delta}_g^{(i)}\|_2 \le \sqrt{n} c_n\} \tag{14}$$

Note that the constraint $\|\mathbf{\Delta}\|_2 \le c$ can be reformulated as the union of several element-wise constraints. Denote $\mathcal{D} = \{\mathbf{D} | \sum_i \sum_j D_{ij}^2 = c^2, D_{ij} \ge 0\}$ (we call an element $\mathbf{D} \in \mathcal{D}$ *decomposition*), then we have

$$\{\mathbf{\Delta} \mid \|\mathbf{\Delta}\|_2 \le c\} = \bigcup_{\mathbf{D} \in \mathcal{D}} \{\mathbf{\Delta} \mid \forall i, j, |\Delta_{ij}| \le D_{ij}\}.$$

Similarly, the uncertainty set $\{\mathbf{\Delta} \mid \|\mathbf{\Delta}_g\|_2 \le c\}$ is equivalent to

$$\bigcup_{\mathbf{D} \in \mathcal{D}_g} \{\mathbf{\Delta} \mid \forall i, \forall j \in g, |\Delta_{ij}| \le D_{ij}\},$$

where $\mathcal{D}_g = \{\mathbf{D} | \sum_i \sum_{j \in g} D_{ij}^2 = c^2, D_{ij} \ge 0\}$. After the constraints of the uncertainty sets are decomposed into element-wise constraints, the set $\{\mathbf{X} + \mathbf{\Delta}^{(1)} + \cdots + \mathbf{\Delta}^{(t)}\}$ can also be represented by an element-wise way. The notation is a little complicated so we first consider three simple cases:

- One uncertainty set $\mathbf{\Delta}$ such that $\|\mathbf{\Delta}\|_2 \le c$: for fixed $\mathbf{D} \in \mathcal{D}$, we have $\{X_{ij} + \Delta_{ij}\} = [X_{ij} - D_{ij}, X_{ij} + D_{ij}]$.
- Two uncertainty sets $\mathbf{\Delta}^{(1)}$ and $\mathbf{\Delta}^{(2)}$ such that $\|\mathbf{\Delta}^{(1)}\|_2 \le c$ and $\|\mathbf{\Delta}^{(2)}\|_2 \le c$: for fixed $\mathbf{D}^{(1)} \in \mathcal{D}$ and $\mathbf{D}^{(2)} \in \mathcal{D}$, we have $\{X_{ij} + \Delta_{ij}^{(1)} + \Delta_{ij}^{(2)}\} = [X_{ij} - D_{ij}^{(1)} - D_{ij}^{(2)}, X_{ij} + D_{ij}^{(1)} + D_{ij}^{(2)}]$.
- One uncertainty set $\mathbf{\Delta}$ and two overlapping groups p and q such that $\|\mathbf{\Delta}_p\|_2 \le c$ and $\|\mathbf{\Delta}_q\|_2 \le c$: for fixed $\mathbf{P} \in \mathcal{D}_p$ and $\mathbf{Q} \in \mathcal{D}_q$, we have

$$\{X_{ij} + \Delta_{ij}\} = \begin{cases} [X_{ij} - P_{ij}, X_{ij} + P_{ij}] & j \in p,\ j \notin q \\ [X_{ij} - Q_{ij}, X_{ij} + Q_{ij}] & j \notin p,\ j \in q \\ [X_{ij} - \min\{P_{ij}, Q_{ij}\}, X_{ij} + \min\{P_{ij}, Q_{ij}\}] & j \in p,\ j \in q \end{cases}$$

Thus, if the decomposition $\mathbf{D} \in \mathcal{D}_g$ for each $\mathbf{\Delta}_g^{(i)}$ is fixed, we have $\{X_{ij} + \Delta_{ij}^{(1)} + \cdots + \Delta_{ij}^{(t)}\} = [X_{ij} - \gamma_{ij}, X_{ij} + \gamma_{ij}]$ where γ_{ij} is determined by the decomposition $\mathbf{D}$s. Since the number of the elements of $\mathbf{\Delta}_g^{(i)}$ is less than or equal to mn (m is the feature dimension and n is the number of samples), there exists a decomposition $\mathbf{D}$ for each $\mathbf{\Delta}_g^{(i)}$ such that $[X_{ij} - \frac{c_n}{\sqrt{m}}, X_{ij} + \frac{c_n}{\sqrt{m}}] \subseteq [X_{ij} - \gamma_{ij}, X_{ij} + \gamma_{ij}]$. We now prove the theorem.

Proposition 1. *(Xu et al., 2010) Given a function $h : \mathcal{R}^{m+1} \mapsto R$ and Borel sets $Z_1, \cdots, Z_n \subseteq \mathcal{R}^{m+1}$, let*

$$P_n = \{\mu \in P | \forall S \subseteq \{1, \cdots, n\} : \mu(\bigcup_{i \in S} Z_i) \ge |S|/n\}.$$

The following holds

$$\frac{1}{n} \sum_{i=1}^{n} \sup_{(b_i, \mathbf{r}_i) \in Z_i} h(b_i, \mathbf{r}_i) = \sup_{\mu \in P_n} \int_{\mathcal{R}^{m+1}} h(b_i, \mathbf{r}_i) d\mu(b_i, \mathbf{r}_i).$$

Step 1: Using the notation above, we first give the following corollary:

Corollary 1. *Given $\mathbf{y} \in \mathcal{R}^n$, $\mathbf{X} \in \mathcal{R}^{n \times m}$, the following equation holds for any $\boldsymbol{\beta} \in \mathcal{R}^m$,*

$$\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|_2 + \sqrt{\frac{n}{m}} c_n + \sum_{i=1}^{t} \max_{\forall g \in \bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \le \sqrt{n} c_n} {\boldsymbol{\alpha}^{(i)}}^\top \boldsymbol{\beta} = \sup_{\mu \in \hat{P}(n)} \sqrt{n \int_{\mathcal{R}^{m+1}} (b' - \mathbf{r}'^\top \boldsymbol{\beta})^2 d\mu(b', \mathbf{r}')} \tag{15}$$

Here,

$$\hat{P}(n) = \bigcup_{\mathcal{S} = \{\mathbf{D}_g^{(i)}\} | \mathbf{D}_g^{(i)} \in \mathcal{D}_g, \forall i, g \in \bar{G}_i} P_n(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n)$$

$$P_n(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n) = \{\mu \in P | Z_i = [y_i - \frac{c_n}{\sqrt{m}}, y_i + \frac{c_n}{\sqrt{m}}] \times \prod_{j=1}^{m} [X_{ij} - \gamma_{ij}, X_{ij} + \gamma_{ij}];$$

$$\forall S \subseteq \{1, \cdots, n\} : \mu(\bigcup_{i \in S} Z_i) \ge |S|/n\},$$

where γ_{ij} depends on the "decomposition" set $\mathcal{S}$.

Proof. The right hand side of Equation (15) is equal to

$$\sup_{\mathcal{S}=\{\mathbf{D}_g^{(i)}\}|\forall i,g\in\bar{G}_i,\mathbf{D}_g^{(i)}\in\mathcal{D}_g}\{\sup_{\mu\in P_n(\mathbf{X},\mathcal{S},\mathbf{y},c_n)}\sqrt{n\int_{\mathcal{R}^{m+1}}(b'-\mathbf{r}'^\top\boldsymbol{\beta})^2 d\mu(b',\mathbf{r}')}\}.$$

From Theorem 2, we know that the left hand side is equal to

$$\begin{aligned}&\sup_{\forall i,g\in G_i,\|\boldsymbol{\delta}_y\|_2\le\sqrt{\frac{n}{m}}c_n,\|\boldsymbol{\Delta}_g^{(i)}\|_2\le\sqrt{n}c_n}\|\mathbf{y}+\boldsymbol{\delta}_y-(\mathbf{X}+\boldsymbol{\Delta})\boldsymbol{\beta}\|_2\\ =&\sup_{\forall i,g\in G_i,\mathbf{D}_g^{(i)}\in\mathcal{D}_g}\{\sup_{\|\boldsymbol{\delta}_y\|_2^2\le\frac{n}{m}c_n^2,|\boldsymbol{\Delta}_g^{(i)}|\le\mathbf{D}_g^{(i)}}\|\mathbf{y}+\boldsymbol{\delta}_y-(\mathbf{X}+\boldsymbol{\Delta})\beta\|_2\}\\ =&\sup_{\forall i,g\in G_i,\mathbf{D}_g^{(i)}\in\mathcal{D}_g}\sqrt{\sum_{i=1}^n\sup_{(b_i,\mathbf{r}_i)\in[y_i-c_n/\sqrt{m},y_i+c_n/\sqrt{m}]\times\prod_{j=1}^m[X_{ij}-\gamma_{ij},X_{ij}+\gamma_{ij}]}(b_i-\mathbf{r}_i^\top\boldsymbol{\beta})}.\end{aligned}$$

Furthermore, applying Proposition 1 yields

$$\begin{aligned}&\sqrt{\sum_{i=1}^n\sup_{(b_i,\mathbf{r}_i)\in[y_i-c_n/\sqrt{m},y_i+c_n/\sqrt{m}]\times\prod_{j=1}^m[X_{ij}-\gamma_{ij},X_{ij}+\gamma_{ij}]}(b_i-\mathbf{r}_i^\top\boldsymbol{\beta})}\\ =&\sqrt{\sup_{\mu\in P(\mathbf{X},\mathcal{S},\mathbf{y},c_n)}n\int_{\mathcal{R}^{m+1}}(b'-\mathbf{r}'^\top\boldsymbol{\beta})^2 d\mu(b',\mathbf{r}')}\\ =&\sup_{\mu\in P(\mathbf{X},\mathcal{S},\mathbf{y},c_n)}\sqrt{n\int_{\mathcal{R}^{m+1}}(b'-\mathbf{r}'^\top\boldsymbol{\beta})^2 d\mu(b',\mathbf{r}')}\end{aligned}$$

which proves the corollary. □

Step 2: As (Xu et al., 2010), we consider the following kernel estimator given samples $(b_i,\mathbf{r}_i)_{i=1}^n$,

$$\begin{aligned}&h_n(b,\mathbf{r})=(nc^{m+1})^{-1}\sum_{i=1}^n K(\frac{b-b_i,\mathbf{r}-\mathbf{r}_i}{c})\\ &\text{where } K(\mathbf{x})=I_{[-1,1]^{m+1}}(\mathbf{x})/2^{m+1},\ \text{and } c=\frac{c_n}{\sqrt{m}}.\end{aligned}\tag{16}$$

Observe that the estimated distribution above belongs to the set of distributions

$$\begin{aligned}P_n(\mathbf{X},\mathcal{S},\mathbf{y},c_n)=\{\mu\in P|Z_i=[y_i-\frac{c_n}{\sqrt{m}},y_i+\frac{c_n}{\sqrt{m}}]\times\prod_{j=1}^m[X_{ij}-\gamma_{ij},X_{ij}+\gamma_{ij}];\\ \forall S\subseteq\{1,\cdots,n\}:\mu(\bigcup_{i\in S}Z_i)\ge|S|/n\}\end{aligned}$$

and hence belongs to $\hat{P}(n)=\bigcup_{\mathcal{S}=\{\mathbf{D}_g^{(i)}\}|\mathbf{D}_g^{(i)}\in\mathcal{D}_g,\forall i,g\in\bar{G}_i}P_n(\mathbf{X},\mathcal{S},\mathbf{y},c_n)$.

Step 3: Combining the last two steps, and using the fact that $\int_{b,\mathbf{r}}|h_n(b,\mathbf{r})-h(b,\mathbf{r})|d(b,\mathbf{r})$ goes to zero almost surely when $c\downarrow 0$ and $nc^{m+1}\uparrow\infty$ or equivalently $c_n\downarrow 0$ and $nc_n^{m+1}\uparrow\infty$. Now we prove consistency of robust regression.

Proof. Let $f(\cdot)$ be the true probability density function of the samples, and $\hat{\mu}_n$ be the estimated distribution using Equation (16) given S_n and c_n, and denote its density function as $f_n(\cdot)$. The condition that $\|\boldsymbol{\beta}(c_n,S_n)\|_2\le H$ almost surely and P has a bounded support implies that there exists a universal constant C such that

$$\max_{b,\mathbf{r}}(b-\mathbf{r}^\top\boldsymbol{\beta}(c_n,S_n))^2\le C$$

almost surely.

By Corollary 1 and $\hat{\mu}_n \in \hat{P}(n)$, we have

$$\begin{aligned}
&\sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\hat{\mu}_n(b, \mathbf{r})} \\
\leq &\sup_{\mu \in \hat{P}(n)} \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\mu_n(b, \mathbf{r})} \\
= &\frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^{n} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(c_n, S_n))^2} + \sum_{i=1}^{t} \max_{\forall g \in \bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \leq c_n} {\boldsymbol{\alpha}^{(i)}}^\top \boldsymbol{\beta} + \frac{1}{\sqrt{m}} c_n \\
\leq &\frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^{n} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(P))^2} + \sum_{i=1}^{t} \max_{\forall g \in \bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \leq c_n} {\boldsymbol{\alpha}^{(i)}}^\top \boldsymbol{\beta} + \frac{1}{\sqrt{m}} c_n
\end{aligned}$$

Notice that, $\sum_{i=1}^{t} \max_{\forall g \in \bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \leq c_n} {\boldsymbol{\alpha}^{(i)}}^\top \boldsymbol{\beta} + \frac{1}{\sqrt{m}} c_n$ converges to 0 as $c_n \downarrow 0$ almost surely, so the right-hand side converges to $\sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(P))^2 dP(b, \mathbf{r})}$ as $n \uparrow \infty$ and $c_n \downarrow 0$ almost surely. Furthermore, we have

$$\begin{aligned}
&\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 dP(b, \mathbf{r}) \\
\leq &\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\hat{\mu}_n(b, \mathbf{r}) + \max_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 \cdot \int_{b,\mathbf{r}} |f_n(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}) \\
\leq &\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\hat{\mu}_n(b, \mathbf{r}) + C \int_{b,\mathbf{r}} |f_n(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}),
\end{aligned}$$

where the last inequality follows from the definition of C. Notice that $\int_{b,\mathbf{r}} |f_n(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r})$ goes to zero almost surely when $c_n \downarrow 0$ and $n c_n^{m+1} \uparrow \infty$. Hence the theorem follows. □

As mentioned in the paper, the assumption that $\|\boldsymbol{\beta}(c_n, S_n)\|_2 \leq H$ in Theorem 7 can be removed, then we have

Theorem 2. *Let $\{c_n\}$ converge to zero sufficiently slowly. Then*

$$\lim_{n \to \infty} \sqrt{\int_{b,\mathbf{r}} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(c_n, S_n))^2 dP(b, \mathbf{r})} = \sqrt{\int_{b,\mathbf{r}} (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(P))^2 dP(b, \mathbf{r})}$$

almost surely.

We now prove this heorem. We establish the following lemma first.

Lemma 2. *Partition the support of P as $V_1, \cdots, V_T$ such that the l_∞ radius of each set is less than $\frac{c_n}{\sqrt{m}}$. If a distribution μ satisfies*

$$\mu(V_t) = \#((b_i, \mathbf{r}_i^\top) \in V_t)/n; \; t = 1, \cdots, T, \tag{17}$$

then $\mu \in \hat{P}(n)$.

Proof. Let $Z_i = [y_i - \frac{c_n}{\sqrt{m}}, y_i + \frac{c_n}{\sqrt{m}}] \times \prod_{j=1}^{m} [X_{ij} - \frac{c_n}{\sqrt{m}}, X_{ij} + \frac{c_n}{\sqrt{m}}]$, recall that X_{ij} is the jth element of $\mathbf{r}_i$. Notice that the l_∞ radius of V_t is less than $\frac{c_n}{\sqrt{m}}$, we have

$$(b_i, \mathbf{r}_i^\top) \in V_t \Rightarrow V_t \subseteq Z_i.$$

Therefore, for any $S \subseteq \{1, \cdots, n\}$, the following holds

$$\mu(\bigcup_{i\in S} Z_i) \geq \mu(\bigcup V_t | \exists i \in S : (b_i, \mathbf{r}_i^\top) \in V_t)$$
$$= \sum_{t|\exists i \in S: (b_i, \mathbf{r}_i^\top) \in V_t} \mu(V_t) = \sum_{t|\exists i \in S: (b_i, \mathbf{r}_i^\top) \in V_t} \#((b_i, \mathbf{r}_i^\top) \in V_t)/n \geq |S|/n.$$

Hence $\mu \in P_n(\mathbf{X}, \mathcal{S}, \mathbf{y}, c_n)$ which implies $\mu \in \hat{P}(n)$. □

Partition the support of P into T subsets such that the l_∞ radius of each set is less than $\frac{c_n}{\sqrt{m}}$. Denote $\tilde{P}(n)$ as the set of probability measures satisfying Equation (17). Hence $\tilde{P}(n) \subseteq \hat{P}(n)$ by Lemma 1. Further notice that there exists a universal constant K such that $\|\boldsymbol{\beta}(c_n, S_n)\|_2 \leq K/c_n$ due to the fact that the square loss of the solution $\boldsymbol{\beta} = 0$ is bounded by a constant only depends on the support of P. Thus, there exists a constant C such that $\max_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 \leq C/c_n^2$. Follow a similar argument as the proof of Theorem 6, we have

$$\begin{aligned} &\sup_{\mu \in \tilde{P}(n)} \sqrt{\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\mu_n(b, \mathbf{r})} \\ \leq &\frac{\sqrt{n}}{n} \sqrt{\sum_{i=1}^n (b_i - \mathbf{r}_i^\top \boldsymbol{\beta}(P))^2 + \sum_{i=1}^t \max_{\forall g \in \bar{G}_i, \|\boldsymbol{\alpha}_g^{(i)}\|_2 \leq c_n} \boldsymbol{\alpha}^{(i)\top} \boldsymbol{\beta} + \frac{1}{\sqrt{m}} c_n} \end{aligned} \tag{18}$$

and

$$\begin{aligned} &\int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 dP(b, \mathbf{r}) \\ \leq &\inf_{\mu_n \in \tilde{P}(n)} \{ \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\mu_n(b, \mathbf{r}) + \max_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 \cdot \int_{b,\mathbf{r}} |f_{\mu_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}) \} \\ \leq &\sup_{\mu_n \in \tilde{P}(n)} \int_{b,\mathbf{r}} (b - \mathbf{r}^\top \boldsymbol{\beta}(c_n, S_n))^2 d\mu_n(b, \mathbf{r}) + 2C/c_n^2 \inf_{\mu_n \in \tilde{P}(n)} \int_{b,\mathbf{r}} |f_{\mu_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}), \end{aligned}$$

here f_μ stands for the density function of a measure μ. Notice that $\tilde{P}(n)$ is the set of distributions satisfying Equation (17), hence $\inf_{\mu_n \in \tilde{P}(n)} \int_{b,\mathbf{r}} |f_{\mu_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r})$ is upper-bounded by $\sum_{t=1}^T |P(V_t) - \#((b_i, \mathbf{r}_i^\top) \in V_t)|/n$, which goes to zero as n increases for any fixed c_n. Therefore,

$$2C/c_n^2 \inf_{\mu_n \in \tilde{P}(n)} \int_{b,\mathbf{r}} |f_{\mu_n}(b, \mathbf{r}) - f(b, \mathbf{r})| d(b, \mathbf{r}) \to 0,$$

if $c_n \downarrow 0$ sufficiently slow. Combining this with Inequality (18) proves the theorem.

References

Grötschel, Martin, Lovász, Lászlo, and Schrijver, Alexander. *Geometric Algorithms and Combinatorial Optimization*, volume 2. Springer, 1988.

Xu, H., Caramanis, C., and Mannor, S. Robust regression and lasso. *IEEE Transactions on Information Theory*, 56(7):3561–3574, 2010.