Supplementary Material of
Noisy Sparse Subspace Clustering

Yu-Xiang Wang LUKE.YXWANG@GMAIL.COM
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

Appendices

A. Proof of Theorem 1

Our main deterministic result Theorem 1 is proved by duality. We first establish a set of conditions on the optimal dual variable of D_0 corresponding to *all* primal solutions satisfying self-expression property. Then we construct such a dual variable ν, hence certify that the optimal solution of P_0 satisfies the LASSO Subspace Detection Property.

A.1. Optimality Condition

Define general convex optimization:

$$\min_{c,e} \|c\|_1 + \frac{\lambda}{2}\|e\|^2 \qquad s.t. \quad x = Ac + e. \tag{A.1}$$

We may state an extension of the Lemma 7.1 in Soltanolkotabi & Candes's SSC Proof.

Lemma A.1. *Consider a vector $y \in \mathbb{R}^d$ and a matrix $A \in \mathbb{R}^{d\times N}$. If there exists triplet (c, e, ν) obeying $y = Ac + e$ and c has support $S \subseteq T$, furthermore the dual certificate vector ν satisfies*

$$\begin{aligned} A_s^T\nu = sgn(c_S), \qquad & \nu = \lambda e, \\ \|A_{T\cap S^c}^T\nu\|_\infty \le 1, \quad & \|A_{T^c}^T\nu\|_\infty < 1, \end{aligned}$$

then all optimal solution (c^*, e^*) to (A.1) obey $c^*_{T^c} = 0$.

Proof. For optimal solution (c^*, e^*), we have:

$$\begin{aligned} &\|c^*\|_1 + \frac{\lambda}{2}\|e^*\|^2 \\ =&\|c^*_S\|_1 + \|c^*_{T\cap S^c}\|_1 + \|c^*_{T^c}\|_1 + \frac{\lambda}{2}\|e^*\|^2 \\ \ge&\|c_S\|_1 + \langle sgn(c_S), c^*_S - c_S\rangle + \|c^*_{T\cap S^c}\|_1 + \|c^*_{T^c}\|_1 \\ &+ \frac{\lambda}{2}\|e\|^2 + \langle \lambda e, e^* - e\rangle \\ =&\|c_S\|_1 + \langle \nu, A_S(c^*_S - c_S)\rangle + \|c^*_{T\cap S^c}\|_1 + \|c^*_{T^c}\|_1 \\ &+ \frac{\lambda}{2}\|e\|^2 + \langle \nu, e^* - e\rangle \\ =&\|c_S\|_1 + \frac{\lambda}{2}\|e\|^2 + \|c^*_{T\cap S^c}\|_1 - \langle \nu, A_{T\cap S^c}(c^*_{T\cap S^c})\rangle \\ &+ \|c^*_{T^c}\|_1 - \langle \nu, A_{T^c}(c^*_{T^c})\rangle \end{aligned} \tag{A.2}$$

To see $\frac{\lambda}{2}\|e^*\|^2 \ge \frac{\lambda}{2}\|e\|^2 + \langle \lambda e, e^* - e\rangle$, note that right hand side equals to $\lambda\left(-\frac{1}{2}e^Te + (e^*)^Te\right)$, which takes a maximal value of $\frac{\lambda}{2}\|e^*\|^2$ when $e = e^*$. The last equation holds because both (c, e) and (c^*, e^*) are feasible solution, such that $\langle \nu, A(c^* - c)\rangle + \langle \nu, e^* - e\rangle = \langle \nu, Ac^* + e^* - (Ac + e)\rangle = 0$. Also, note that $\|c_S\|_1 + \frac{\lambda}{2}\|e\|^2 = \|c\|_1 + \frac{\lambda}{2}\|e\|^2$.

With the inequality constraints of ν given in the Lemma statement, we know

$$\begin{aligned} \langle \nu, A_{T\cap S^c}(c^*_{T\cap S^c})\rangle =&\langle A^T_{T\cap S^c}\nu, (c^*_{T\cap S^c})\rangle \\ \le&\|A^T_{T\cap S^c}\nu\|_\infty \|c^*_{T\cap S^c}\|_1 \le \|c^*_{T\cap S^c}\|_1. \end{aligned}$$

Substitute into (A.2), we get:

$$\|c^*\|_1 + \frac{\lambda}{2}\|e^*\|^2 \ge \|c\|_1 + \frac{\lambda}{2}\|e\|^2 + (1 - \|A^T_{T^c}\nu\|_\infty)\|c^*_{T^c}\|_1,$$

where $(1 - \|A^T_{T^c}\nu\|_\infty)$ is strictly greater than 0.

Using the fact that (c^*, e^*) is an optimal solution, $\|c^*\|_1 + \frac{\lambda}{2}\|e^*\|^2 \le \|c\|_1 + \frac{\lambda}{2}\|e\|^2$. Therefore, $\|c^*_{T^c}\|_1 = 0$ and (c, e) is also an optimal solution. This concludes the proof. □

Apply Lemma A.1 with $x = x_i^{(\ell)}$ and $A = X_{-i}$, we know that if we can construct a dual certificate ν such that all conditions are satisfied with respect to a feasible solution (c, e) and c satisfy SEP, then the all optimal solution of (4.1) satisfies SEP, in other word $c_i = \left[0, ..., 0, (c_i^{(\ell)})^T, 0, ..., 0\right]^T$.

By definition of LASSO detection property, we must further ensure $\|c_i^{(\ell)}\|_1 \neq 0$ to avoid the trivial solution that $x_i^{(\ell)} = e^*$. This is a non-convex constraint and hard to impose. To this matter, we note that given sufficiently large λ, $\|c_i^{(\ell)}\|_1 \neq 0$ never occurs.

Our strategy of avoiding this trivial solution is hence showing the existence of a λ such that the dual optimal value is smaller than the trivial optimal value, namely:

$$OptVal(\mathbf{D}_0) = \langle x_i, \nu \rangle - \frac{1}{2\lambda}\|\nu\|^2 < \frac{\lambda}{2}\|x_i^{(\ell)}\|^2. \quad \text{(A.3)}$$

A.2. Constructing candidate dual vector ν

A natural candidate of the dual solution ν is the dual point corresponding to the optimal solution of the following fictitious optimization program.

$$\begin{aligned} \mathbf{P}_1 : \quad & \min_{c_i^{(\ell)}, e_i} \|c_i^{(\ell)}\|_1 + \frac{\lambda}{2}\|e_i\|^2 \\ & s.t. \quad y_i^{(\ell)} + z_i = (Y_{-i}^{(\ell)} + Z_{-i}^{(\ell)})c_i^{(\ell)} + e_i \end{aligned} \quad \text{(A.4)}$$

$$\begin{aligned} \mathbf{D}_1 : \quad & \max_{\nu} \langle x_i^{(\ell)}, \nu \rangle - \frac{1}{2\lambda}\nu^T\nu \\ & s.t. \quad \|(X_{-i}^{(\ell)})^T\nu\|_\infty \leq 1. \end{aligned} \quad \text{(A.5)}$$

This optimization is feasible because $y_i^{(\ell)} \in span(Y_{-i}^{(\ell)}) = \mathcal{S}_\ell$ so any $c_i^{(\ell)}$ obeying $y_i^{(\ell)} = Y_{-i}^{(\ell)}c_i^{(\ell)}$ and corresponding $e_i = z_i - Z_{-i}^{(\ell)}c_i^{(\ell)}$ is a pair of feasible solution. Then by strong duality, the dual program is also feasible, which implies that for every optimal solution (c, e) of (A.4) with c supported on S, there exist ν satisfying:

$$\left\{ \begin{aligned} &\|((Y_{-i}^{(\ell)})_{S^c}^T + (Z_{-i}^{(\ell)})_{S^c}^T)\nu\|_\infty \leq 1, \quad \nu = \lambda e, \\ &((Y_{-i}^{(\ell)})_S^T + (Z_{-i}^{(\ell)})_S^T)\nu = sgn(c_S). \end{aligned} \right\}$$

This construction of ν satisfies all conditions in Lemma A.1 with respect to

$$\begin{cases} c_i = [0, ..., 0, c_i^{(\ell)}, 0, ..., 0] \text{ with } c_i^{(\ell)} = c, \\ e_i = e, \end{cases} \quad \text{(A.6)}$$

except

$$\left\|[X_1, ..., X_{\ell-1}, X_{\ell+1}, ..., X_L]^T\nu\right\|_\infty < 1,$$

i.e., we must check for all data point $x \in \mathcal{X} \setminus \mathcal{X}^\ell$,

$$|\langle x, \nu \rangle| < 1. \quad \text{(A.7)}$$

Showing the solution of (A.5) ν also satisfies (A.7) gives precisely a dual certificate as required in Lemma A.1, hence implies that the candidate solution (A.6) associated with optimal (c, e) of (A.4) is indeed the optimal solution of (4.1).

A.3. Dual separation condition

In this section, we establish the conditions required for (A.7) to hold. The idea is to provide an upper bound of $|\langle x, \nu \rangle|$ then make it smaller than 1.

First, we find it appropriate to project ν to the subspace $\mathcal{S}_\ell$ and its complement subspace then analyze separately. For convenience, denote $\nu_1 := \mathbb{P}_{\mathcal{S}_\ell}(\nu)$, $\nu_2 := \mathbb{P}_{\mathcal{S}_\ell^c}(\nu)$. Then

$$\begin{aligned} |\langle x, \nu \rangle| =& |\langle y + z, \nu \rangle| \leq |\langle y, \nu_1 \rangle| + |\langle y, \nu_2 \rangle| + |\langle z, \nu \rangle| \\ \leq& \mu(\mathcal{X}_\ell)\|\nu_1\| + \|y\|\|\nu_2\|| \cos(\angle(y, \nu_2))| \\ &+ \|z\|\|\nu\|| \cos(\angle(z, \nu))|. \end{aligned} \quad \text{(A.8)}$$

To see the last inequality, check that by Definition 3, $|\langle y, \frac{\nu_1}{\|\nu_1\|} \rangle| \leq \mu(\mathcal{X}_\ell)$.

Since we are considering general (possibly adversarial) noise, we will use the relaxation $|\cos(\theta)| \leq 1$ for all cosine terms (a better bound under random noise will be given later). Now all we have to do is to bound $\|\nu_1\|$ and $\|\nu_2\|$ (note $\|\nu\| = \sqrt{\|\nu_1\|^2 + \|\nu_2\|^2} \leq \|\nu_1\| + \|\nu_2\|$).

A.3.1. Bounding $\|\nu_1\|$

We first bound $\|\nu_1\|$ by exploiting the feasible region of ν_1 in (A.5).

$$\|(X_{-i}^{(\ell)})^T\nu\|_\infty \leq 1$$

is equivalent to

$$x_i^T\nu \leq 1$$

for every x_i that is the column of $X_{-i}^{(\ell)}$. Decompose the condition into

$$y_i^T\nu_1 + (\mathbb{P}_{\mathcal{S}_\ell}z_i)^T\nu_1 + z_i^T\nu_2 \leq 1.$$

Now we relax each of the term into

$$y_i^T\nu_1 + (\mathbb{P}_{\mathcal{S}_\ell}z_i)^T\nu_1 \leq 1 - z_i^T\nu_2 \leq 1 + \delta\|\nu_2\|. \quad \text{(A.9)}$$

The relaxed condition contains the feasible region of ν_1 in (A.5).

It turns out that the geometric interpretation of the relaxed constraints gives a upper bound of $\|\nu_1\|$.

Definition A.1 (polar set)**.** *The polar set $\mathcal{K}^o$ of set $\mathcal{K} \in \mathbb{R}^d$ is defined as*

$$\mathcal{K}^o = \left\{ y \in \mathbb{R}^d : \langle x, y \rangle \le 1 \text{ for all } x \in \mathcal{K} \right\}$$

By the polytope geometry, we have

$$\begin{aligned} &\|(Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)}))^T \nu_1\|_\infty \le 1 + \delta\|\nu_2\| \\ \Leftrightarrow\ & \nu_1 \in \left[\mathcal{P}\left(\frac{Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)})}{1 + \delta\|\nu_2\|}\right)\right]^o := \mathcal{T}^o. \end{aligned} \tag{A.10}$$

Now we introduce the concept of circumradius.

Definition A.2 (circumradius)**.** *The circumradius of a convex body $\mathcal{P}$, denoted by $R(\mathcal{P})$, is defined as the radius of the smallest Euclidean ball containing $\mathcal{P}$.*

The magnitude $\|\nu_1\|$ is bounded by $R(\mathcal{T}^o)$. Moreover, by the the following lemma we may find the circumradius by analyzing the polar set of $\mathcal{T}^o$ instead. By the property of polar operator, polar of a polar set gives the tightest convex envelope of original set, i.e., $(\mathcal{K}^o)^o = conv(\mathcal{K})$. Since $\mathcal{T} = \text{conv}\left(\pm \frac{Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)})}{1+\delta\|\nu_2\|}\right)$ is convex in the first place, the polar set of $\mathcal{T}^o$ is essentially $\mathcal{T}$.

Lemma A.2. *For a symmetric convex body $\mathcal{P}$, i.e. $\mathcal{P} = -\mathcal{P}$, inradius of $\mathcal{P}$ and circumradius of polar set of $\mathcal{P}$ satisfy:*

$$r(\mathcal{P})R(\mathcal{P}^o) = 1.$$

Lemma A.3. *Given $X = Y + Z$, denote $\rho := \max_i \|\mathbb{P}_{\mathcal{S}} z_i\|$, furthermore $Y \in \mathcal{S}$ where $\mathcal{S}$ is a linear subspace, then we have:*

$$r(\text{Proj}_{\mathcal{S}}(\mathcal{P}(X))) \ge r(\mathcal{P}(Y)) - \rho$$

Proof. First note that projection to subspace is a linear operator, hence $\text{Proj}_{\mathcal{S}}(\mathcal{P}(X)) = \mathcal{P}(\mathbb{P}_{\mathcal{S}} X)$. Then by definition, the boundary set of $\mathcal{P}(\mathbb{P}_{\mathcal{S}} X)$ is $\mathcal{B} := \{y \mid y = \mathbb{P}_{\mathcal{S}} Xc; \|c\|_1 = 1\}$. Inradius by definition is the largest ball containing in the convex body, hence $r(\mathcal{P}(\mathbb{P}_{\mathcal{S}} X)) = \min_{y \in \mathcal{B}} \|y\|$. Now we provide a lower bound of it:

$$\begin{aligned} \|y\| \ge& \|Yc\| - \|\mathbb{P}_{\mathcal{S}} Zc\| \ge r(\mathcal{P}(Y)) - \sum\nolimits_j \|\mathbb{P}_{\mathcal{S}} z_j\| |c_j| \\ \ge& r(\mathcal{P}(Y)) - \rho\|c\|_1. \end{aligned}$$

This concludes the proof. □

A bound of $\|\nu_1\|$ follows directly from Lemma A.2 and Lemma A.3:

$$\begin{aligned} \|\nu_1\| \le& (1 + \delta\|\nu_2\|) R(\mathcal{P}(Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)}))) \\ =& \frac{1 + \delta\|\nu_2\|}{r(\mathcal{P}(Y_{-i}^{(\ell)} + \mathbb{P}_{\mathcal{S}_\ell}(Z_{-i}^{(\ell)}))} = \frac{1 + \delta\|\nu_2\|}{r(\text{Proj}_{\mathcal{S}_\ell}(\mathcal{P}(X_{-i}^{(\ell)})))} \\ \le& \frac{1 + \delta\|\nu_2\|}{r(\mathcal{Q}_{-i}^{\ell}) - \delta_1}. \end{aligned} \tag{A.11}$$

This bound unfortunately depends $\|\nu_2\|$. This can be extremely loose as in general, ν_2 is not well-constrained (see the illustration in Figure C.2 and C.3). That is why we need to further exploit the fact ν is the optimal solution of (A.5), which provides a reasonable bound of $\|\nu_2\|$.

A.3.2. Bounding $\|\nu_2\|$

By optimality condition:

$$\nu = \lambda e_i = \lambda(x_i - X_{-i} c)$$

and

$$\nu_2 = \lambda \mathbb{P}_{\mathcal{S}_\ell^\perp}(x_i - X_{-i} c) = \lambda \mathbb{P}_{\mathcal{S}_\ell^\perp}(z_i - Z_{-i} c)$$

so

$$\begin{aligned} \|\nu_2\| &\le \lambda\left(\|\mathbb{P}_{\mathcal{S}_\ell^\perp} z_i\| + \|\mathbb{P}_{\mathcal{S}_\ell^\perp} Z_{-i} c\|\right) \\ &\le \lambda(\|\mathbb{P}_{\mathcal{S}_\ell^\perp} z_i\| + \sum_{j \in S} |c_j| \|\mathbb{P}_{\mathcal{S}_\ell^\perp} z_j\|) \\ &\le \lambda(\|c\|_1 + 1)\delta_2 \le \lambda(\|c\|_1 + 1)\delta \end{aligned} \tag{A.12}$$

Now we will bound $\|c\|_1$. As c is the optimal solution, $\|c\|_1 \le \|c\|_1 + \frac{\lambda}{2}\|e\|^2 \le \|\tilde{c}\|_1 + \frac{\lambda}{2}\|\tilde{e}\|^2$ for any feasible solution $(\tilde{c}, \tilde{e})$. Let $\tilde{c}$ be the solution of

$$\begin{aligned} &\min_c \|c\|_1 \\ &s.t. \quad y_i^{(\ell)} = Y_{-i}^{(\ell)} c, \end{aligned} \tag{A.13}$$

then by strong duality,

$$\|\tilde{c}\|_1 = \max_\nu \left\{ \langle \nu, y_i^{(\ell)} \rangle \mid \|[Y_{-i}^{(\ell)}]^T \nu\|_\infty \le 1 \right\}.$$

By Lemma A.2, optimal dual solution $\tilde{\nu}$ satisfies $\|\tilde{\nu}\| \le \frac{1}{r(\mathcal{Q}_{-i}^\ell)}$. It follows that

$$\|\tilde{c}\|_1 = \langle \tilde{\nu}, y_i^{(\ell)} \rangle = \|\tilde{\nu}\| \|y_i^{(\ell)}\| \le \frac{1}{r(\mathcal{Q}_{-i}^\ell)}.$$

On the other hand, $\tilde{e} = z_i - Z_{-i}^{(\ell)} \tilde{c}$, so $\|\tilde{e}\|^2 \le (\|z_i\| + \sum_j \|z_j\| |\tilde{c}_j|)^2 \le (\delta + \|\tilde{c}\|_1 \delta)^2$, thus:

$$\|c\|_1 \le \|\tilde{c}\|_1 + \frac{\lambda}{2}\|\tilde{e}\|^2 \le \frac{1}{r(\mathcal{Q}_{-i}^\ell)} + \frac{\lambda}{2}\delta^2\left[1 + \frac{1}{r(\mathcal{Q}_{-i}^\ell)}\right]^2.$$

This gives the bound we desired:

$$\|\nu_2\| \leq \lambda\left(\frac{1}{r(\mathcal{Q}^\ell_{-i})} + \frac{\lambda}{2}\delta^2\left[1+\frac{1}{r(\mathcal{Q}^\ell_{-i})}\right]^2 + 1\right)\delta$$
$$= \lambda\delta\left(\frac{1}{r(\mathcal{Q}^\ell_{-i})}+1\right) + \frac{\delta}{2}\left\{\lambda\delta\left(\frac{1}{r(\mathcal{Q}^\ell_{-i})}+1\right)\right\}^2.$$

By choosing λ satisfying

$$\lambda\delta^2 \leq \frac{2}{1+1/r(\mathcal{Q}^\ell_{-i})}, \tag{A.14}$$

the bound can be simplified to:

$$\|\nu_2\| \leq 2\lambda\delta\left(\frac{1}{r(\mathcal{Q}^\ell_{-i})}+1\right) \tag{A.15}$$

A.3.3. Conditions for $|\langle x, \nu\rangle| < 1$

Putting together (A.8), (A.11) and (A.15), we have the upper bound of $|\langle x, \nu\rangle|$:

$$|\langle x,\nu\rangle| \leq (\mu(\mathcal{X}_\ell) + \|\mathbb{P}_{\mathcal{S}_\ell} z\|)\|\nu_1\| + (\|y\| + \|\mathbb{P}_{\mathcal{S}_\ell^\perp} z\|)\|\nu_2\|$$
$$\leq \frac{\mu(\mathcal{X}_\ell)+\delta_1}{r(\mathcal{Q}^\ell_{-i})-\delta_1} + \left(\frac{(\mu(\mathcal{X}_\ell)+\delta_1)\delta}{r(\mathcal{Q}^\ell_{-i})-\delta_1} + 1 + \delta\right)\|\nu_2\|$$
$$\leq \frac{\mu(\mathcal{X}_\ell)+\delta_1}{r(\mathcal{Q}^\ell_{-i})-\delta_1} + 2\lambda\delta(1+\delta)\left(\frac{1}{r(\mathcal{Q}^\ell_{-i})}+1\right)$$
$$+ \frac{2\lambda\delta^2(\mu(\mathcal{X}_\ell)+\delta_1)}{r(\mathcal{Q}^\ell_{-i})-\delta_1}\left(\frac{1}{r(\mathcal{Q}^\ell_{-i})}+1\right)$$

For convenience, we further relax the second $r(\mathcal{Q}^\ell_{-i})$ into $r(\mathcal{Q}^\ell_{-i}) - \delta_1$. The dual separation condition is thus guaranteed with

$$\frac{\mu(\mathcal{X}_\ell)+\delta_1+2\lambda\delta(1+\delta)+2\lambda\delta^2(\mu(\mathcal{X}_\ell)+\delta_1)}{r(\mathcal{Q}^\ell_{-i})-\delta_1}$$
$$+2\lambda\delta(1+\delta) + \frac{2\lambda\delta^2(\mu(\mathcal{X}_\ell)+\delta_1)}{r(\mathcal{Q}^\ell_{-i})(r(\mathcal{Q}^\ell_{-i})-\delta_1)} < 1.$$

Denote $\rho := \lambda\delta(1+\delta)$, assume $\delta < r(\mathcal{Q}^\ell_{-i})$, $(\mu(\mathcal{X}_\ell)+\delta_1) < 1$ and simplify the form with

$$\frac{2\lambda\delta^2(\mu(\mathcal{X}_\ell)+\delta_1)}{r(\mathcal{Q}^\ell_{-i})-\delta_1} + \frac{2\lambda\delta^2(\mu(\mathcal{X}_\ell)+\delta_1)}{r(\mathcal{Q}^\ell_{-i})(r(\mathcal{Q}^\ell_{-i})-\delta_1)}$$
$$< \frac{2\rho}{r(\mathcal{Q}^\ell_{-i})-\delta_1},$$

we get a sufficient condition

$$\mu(\mathcal{X}_\ell) + 3\rho + \delta_1 < (1-2\rho)\left(r(\mathcal{Q}^\ell_{-i})-\delta_1\right). \tag{A.16}$$

To generalize (A.16) to all data of all subspaces, the following must hold for each $\ell = 1, ..., k$:

$$\mu(\mathcal{X}_\ell) + 3\rho + \delta_1 < (1-2\rho)\left(\min_{\{i: x_i \in X^{(\ell)}\}} r(\mathcal{Q}^{(\ell)}_{-i}) - \delta_1\right). \tag{A.17}$$

This gives a first condition on δ and λ, which we call it "**dual separation condition**" under noise. Note that this reduces to exactly the geometric condition in Soltanolkotabi & Candes's Theorem 2.5 when $\delta = 0$.

A.4. Avoid trivial solution

In this section we provide sufficient conditions on λ such that trivial solution $c = 0$, $e = x_i^{(\ell)}$ is not the optimal solution. For any optimal triplet (c, e, ν) we have $\nu = \lambda e$, a condition: $\|\nu\| < \lambda\|x_i^{(\ell)}\|$ implies that optimal $\|e\| < \|x_i^{(\ell)}\|$, so $e \neq x_i^{(\ell)}$. By the equality constraint, $X_{-i}^{(\ell)} c = x_i^{(\ell)} - e \neq 0$, therefore $\|c\|_1 \neq 0$. Now we will establish the condition on λ such that:

$$\|\nu\| < \lambda\|x_i^{(\ell)}\|.$$

An upper bound of $\|\nu\|$ and a lower bound of $\lambda\|x_i^{(\ell)}\|$ are readily available:

$$\|\nu\| \leq \|\nu_1\| + \|\nu_2\| \leq \frac{1}{r(\mathcal{Q}^\ell_{-i})-\delta_1}$$
$$+ 2\lambda\delta\left(\frac{1}{r(\mathcal{Q}^\ell_{-i})}+1\right)\left(1+\frac{\delta}{r(\mathcal{Q}^\ell_{-i})-\delta_1}\right)$$
$$\leq \frac{1+3\lambda\delta+2\lambda\delta^2}{r(\mathcal{Q}^\ell_{-i})-\delta_1} + 2\lambda\delta,$$
$$\lambda\|x_i^{(\ell)}\| \geq \lambda(\|y_i^{(\ell)}\| - \|z_i^{(\ell)}\|) \geq \lambda(1-\delta).$$

So the sufficient condition on λ such that solution is non-trivial is

$$\frac{1+3\lambda\delta+2\lambda\delta^2}{r(\mathcal{Q}^\ell_{-i})-\delta_1} + 2\lambda\delta < \lambda(1-\delta).$$

Reorganize the condition, we reach

$$\lambda > \frac{1}{(r(\mathcal{Q}^\ell_{-i})-\delta_1)(1-3\delta)-3\delta-2\delta^2}. \tag{A.18}$$

For the inequality operations above to be valid, we need:

$$\begin{cases} r(\mathcal{Q}^\ell_{-i}) - \delta_1 > 0 \\ (r(\mathcal{Q}^\ell_{-i})-\delta_1)(1-3\delta)-3\delta-2\delta^2 > 0 \end{cases}$$

Relax δ_1 to δ and solve the system of inequalities, we get:

$$\delta < \frac{3r+4-\sqrt{9r^2+20r+16}}{2} = \frac{2r}{3r+4+\sqrt{9r^2+20r+16}}.$$

Use $\sqrt{9r^2+20r+16} \le 3r+4$ and impose the constraint for all $x_i^{(\ell)}$, we choose to impose a stronger condition for every $\ell = 1, ..., L$:

$$\delta < \frac{\min_i r\left(\mathcal{Q}^{\ell}_{-i}\right)}{3\min_i r\left(\mathcal{Q}^{\ell}_{-i}\right)+4}. \tag{A.19}$$

A.5. Existence of a proper λ

Basically, (A.17), (A.18) and (A.14) must be satisfied simultaneously for all $\ell = 1, ..., L$. Essentially (A.18) gives condition of λ from below, the other two each gives a condition from above. Denote $r_\ell := \min_{\{i:x_i\in X^{(\ell)}\}} r(\mathcal{Q}^{(\ell)}_{-i})$, $\mu_\ell := \mu(\mathcal{X}_\ell)$, the condition on λ is:

$$\begin{cases}\lambda & > \max_\ell \frac{1}{(r_\ell-\delta_1)(1-3\delta)-3\delta-2\delta^2}\\ \lambda & < \min_\ell\left(\frac{r_\ell-\mu_\ell-2\delta_1}{\delta(1+\delta)(3+2r_\ell-2\delta_1)} \vee \frac{2r_\ell}{\delta^2(r_\ell+1)}\right)\end{cases}$$

Note that on the left

$$\max_\ell\left\{\frac{1}{(r_\ell-\delta_1)(1-3\delta)-3\delta-2\delta^2}\right\}$$
$$=\frac{1}{(\max_\ell r_\ell-\delta_1)(1-3\delta)-3\delta-2\delta^2}.$$

On the right

$$\min_\ell\left\{\frac{2r_\ell}{\delta^2(r_\ell+1)}\right\} = \frac{2\min_\ell r_\ell}{\delta^2(\min_\ell r_\ell+1)}.$$

Denote $r = \min_\ell r_\ell$, it suffices to guarantee for each ℓ:

$$\begin{cases}\lambda & > \frac{1}{(r-\delta_1)(1-3\delta)-3\delta-2\delta^2}\\ \lambda & < \frac{r_\ell-\mu_\ell-2\delta_1}{\delta(1+\delta)(3+2r_\ell-2\delta_1)} \vee \frac{2r}{\delta^2(r+1)}\end{cases} \tag{A.20}$$

To understand this, when δ and μ is small then any λ values satisfying $\Theta(r) < \lambda < \Theta(r/\delta)$ will satisfy separation condition. We will now derive the condition on δ such that (A.20) is not an empty set.

A.6. Lower bound of break-down point

(A.19) gives one requirement on δ and the range of (A.20) being non-empty gives another. Combining these two leads to lower bound of the breakdown point. In other word, the algorithm will be robust to arbitrary corruptions with magnitude less than this point for some λ.

Again, we relax δ_1 to δ in (A.20) to get:

$$\begin{cases}\frac{1}{(r-\delta)(1-3\delta)-3\delta-2\delta^2} < \frac{r_\ell-\mu_\ell-2\delta}{\delta(1+\delta)(3+2r_\ell-2\delta)}\\ \frac{1}{(r-\delta)(1-3\delta)-3\delta-2\delta^2} < \frac{2r}{\delta^2(r+1)}.\end{cases}$$

The first inequality in standard form is:

$$A\delta^3+B\delta^2+C\delta+D<0$$

with

$$\begin{cases}A=0\\ B=-(6r-r_\ell+7-\mu_\ell)\\ C=3r_\ell r+6r_\ell+2r-3\mu_\ell r+3-4\mu_\ell\\ D=-r(r_\ell-\mu_\ell)\end{cases}$$

This is an extremely complicated 3^{rd} order polynomial. We will try to simplify it imposing a stronger condition. First extract and regroup μ_ℓ in first three terms, we get $(\delta^2-4\delta-3r\delta)\mu_\ell$ which is negative, so we drop it. Second we express the remaining expression using:

$$f(r,\delta)\delta < r(r-\mu),$$

where

$$f(r,\delta) = -(6r-r_\ell+7)\delta+3r_\ell r+6r_\ell+2r+2.$$

Note that since $\delta<1$, we can write

$$f(r,\delta)\le f(r,0) = 3r_\ell r+6r_\ell+2r+2 \le 3r_\ell^2+8r_\ell+2.$$

Thus, a stronger condition on δ is established:

$$\delta < \frac{r(r_\ell-\mu_\ell)}{3r_\ell^2+8r_\ell+2} \tag{A.21}$$

The second inequality in standard form is:

$$(1-r)\delta^2+(6r^2+8r)\delta-2r^2<0$$

By definition $r<1$, we solve the inequality and get:

$$\begin{cases}\delta > \frac{-3r^2-4r-r\sqrt{9r^2+22r+18}}{1-r}\\ \delta < \frac{-3r^2-4r+r\sqrt{9r^2+22r+18}}{1-r}\end{cases}$$

The lower constraint is always satisfied. Rationalized the expression of the upper constraint, $1-r$ gets cancelled out:

$$\delta < \frac{2r^2}{3r^2+4r+r\sqrt{9r^2+22r+18}}.$$

It turns out that (A.19) is sufficient for the inequality to hold. This is by $\sqrt{9r^2+22r+18} < \sqrt{9r^2+24r+16} = 3r+4$. Combine with (A.21) we reach the overall condition:

$$\delta < \left\{\frac{r(r_\ell-\mu_\ell)}{3r_\ell^2+8r_\ell+2}\right\} \vee \frac{r}{3r+4} = \frac{r(r_\ell-\mu_\ell)}{3r_\ell^2+8r_\ell+2}. \tag{A.22}$$

The first expression is always smaller because:

$$\frac{r}{3r+4} \ge \frac{rr_\ell}{3rr_\ell+4r_\ell} \ge \frac{rr_\ell}{3rr_\ell+4r_\ell+3r_\ell+2}$$
$$\ge \frac{r(r_\ell-\mu_\ell)}{3r_\ell^2+8r_\ell+2}.$$

Verify that when (A.22) is true for all ℓ, there exists a single λ for solution of (2.2) to satisfy subspace detection property for all x_i. The proof of Theorem 1 is now complete.

B. Proof of Randomized Results

In this section, we provide proof to the Theorems about the three randomized models:

- **Determinitic data+random noise**
- **Semi-random data+random noise**
- **Fully random**

To do this, we need to bound δ_1, $\cos(\angle(z,\nu))$ and $\cos(\angle(y,\nu_2))$ when the Z follows *Random Noise Model*, such that a better dual separation condition can be obtained. Moreover, for *Semi-random* and *Random data model*, we need to bound $r(\mathcal{Q}_{-i}^{(\ell)})$ when data samples from each subspace are drawn uniformly and bound $\mu(\mathcal{X}_\ell)$ when subspaces are randomly generated.

These requires the following Lemmas.

Lemma B.1 (Upper bound on the area of spherical cap)**.** *Let $a\in\mathbb{R}^n$ be a random vector sampled from a unit sphere and z is a fixed vector. Then we have:*

$$Pr\left(|a^Tz|>\epsilon\|z\|\right)\leq 2e^{\frac{-n\epsilon^2}{2}}$$

This Lemma is extracted from an equation in page 29 of Soltanolkotabi & Candes (2012), which is in turn adapted from the upper bound on the area of spherical cap in Ball (1997). By definition of Random Noise Model, z_i has spherical symmetric, which implies that the direction of z_i distributes uniformly on an n-sphere. Hence Lemma B.1 applies whenever an inner product involves z.

As an example, , we write the following lemma

Lemma B.2 (Properties of Gaussian noise)**.** *For Gaussian random matrix $Z\in\mathbb{R}^{n\times N}$, if each entry $Z_{i,j}\sim N(0,\frac{\sigma}{\sqrt{n}})$, then each column z_i satisfies:*

1. $Pr(\|z_i\|^2>(1+t)\sigma^2)\leq e^{\frac{n}{2}(\log(t+1)-t)}$
2. $Pr(|\langle z_i,z\rangle|>\epsilon\|z_i\|\|z\|)\leq 2e^{\frac{-n\epsilon^2}{2}}$

where z is any fixed vector(or random generated but independent to z_i).

Proof. The second property follows directly from Lemma B.1 as Gaussian vector has uniformly random direction.

To show the first property, we observe that the sum of n independent square Gaussian random variables follows χ^2 distribution with d.o.f n, in other word, we have

$$\|z_i\|^2=|Z_{1i}|^2+...+|Z_{ni}|^2\sim\frac{\sigma^2}{n}\chi^2(n).$$

By Hoeffding's inequality, we have an approximation of its CDF (Dasgupta & Gupta, 2002), which gives us

$$Pr(\|z_i\|^2>\alpha\sigma^2)=1-\mathrm{CDF}_{\chi_n^2}(\alpha)\leq(\alpha e^{1-\alpha})^{\frac{n}{2}}.$$

Substitute $\alpha=1+t$, we get exactly the concentration statement. □

By Lemma B.2, $\delta=\max_i\|z_i\|$ is bounded with high probability. δ_1 has an even tighter bound because each $\mathcal{S}_\ell$ is low-rank. Likewise, $\cos(\angle(z,\nu))$ is bounded to a small value with high probability. Moreover, since $\nu=\lambda e=\lambda(x_i-X_{-i}c)$, $\nu_2=\lambda\mathbb{P}_{\mathcal{S}_\ell^\perp}(z_i-Z_{-i}c)$, thus ν_2 is merely a weighted sum of random noise in a $(n-d_\ell)$-dimensional subspace. Consider y a fixed vector, $\cos(\angle(y,\nu_2))$ is also bounded with high probability.

Replace these observations into (A.7) and the corresponding bound of $\|\nu_1\|$ and $\|\nu_2\|$. We obtained the dual separation condition for under Random noise model.

Lemma B.3 (Dual separation condition under random noise)**.** *Let $\rho:=\lambda\delta(1+\delta)$ and*

$$\epsilon:=\sqrt{\frac{6\log N+2\log\max_\ell d_\ell}{n-\max_\ell d_\ell}}\leq\frac{C\log(N)}{\sqrt{n}}$$

for some constant C. Under random noise model, if for each $\ell=1,...,L$

$$\mu(\mathcal{X}_\ell)+3\rho\epsilon+\delta\epsilon\leq(1-2\rho\epsilon)(\max_i r(\mathcal{Q}_{-i}^{(\ell)})-\delta\epsilon),$$

then dual separation condition (A.7) holds for all data points with probability at least $1-7/N$.

Proof. Recall that we want to find an upper bound of $|\langle x,\nu\rangle|$.

$$\begin{aligned}|\langle x,\nu\rangle|\leq&\mu\|\nu_1\|+\|y\|\|\nu_2\||\cos(\angle(y,\nu_2))|\\&+\|z\|\|\nu\||\cos(\angle(z,\nu))|\end{aligned}\tag{B.1}$$

Here we will bound the two cosine terms and δ_1 under random noise model.

As discussed above, directions of z and ν_2 are independently and uniformly distributed on the n-sphere. Then by Lemma B.1,

$$\begin{cases}Pr\left(\cos(\angle(z,\nu))>\sqrt{\frac{6\log N}{n}}\right)\leq\frac{2}{N^3}\\Pr\left(\cos(\angle(y,\nu_2))>\sqrt{\frac{6\log N}{n-d_\ell}}\right)\leq\frac{2}{N^3}\\Pr\left(\cos(\angle(z,\nu_2))>\sqrt{\frac{6\log N}{n}}\right)\leq\frac{2}{N^3}\end{cases}$$

Using the same technique, we provide a bound for δ_1. Given orthonormal basis U of S_ℓ, $\mathbb{P}_{S_\ell}z=UU^Tz$, then

$$\|UU^Tz\|=\|U^Tz\|\leq\sum_{i=1,\ldots,d_\ell}|U_{:,i}^Tz|.$$

Apply Lemma B.1 for each i , then apply union bound, we get:

$$Pr\left(\|\mathbb{P}_{S_\ell} z\| > \sqrt{\frac{2\log d_\ell + 6\log N}{n}}\delta\right) \le \frac{2}{N^3}$$

Since δ_1 is the worse case bound for all L subspace and all N noise vector, then a union bound gives:

$$Pr\left(\delta_1 > \sqrt{\frac{2\log d_\ell + 6\log N}{n}}\delta\right) \le \frac{2L}{N^2}$$

Moreover, we can find a probabilistic bound for $\|\nu_1\|$ too by a random variation of (A.9) which is now

$$y_i^T\nu_1 + (\mathbb{P}_{\mathcal{S}_\ell} z_i)^T\nu_1 \le 1 - z_i^T\nu_2 \le 1 + \delta_2\|\nu_2\|| \cos\angle(z_i,\nu_2)|. \tag{B.2}$$

Substituting the upper bound of the cosines, we get:

$$|\langle x, \nu\rangle| \le \mu\|\nu_1\| + \|y\|\|\nu_2\|\sqrt{\frac{6\log N}{n - d_\ell}} + \|z\|\|\nu\|\sqrt{\frac{6\log N}{n}}$$

$$\|\nu_1\| \le \frac{1 + \delta\|\nu_2\|\sqrt{\frac{6\log N}{n}}}{r(\mathcal{Q}^\ell_{-i}) - \delta_1}, \quad \|\nu_2\| \le 2\lambda\delta\left(\frac{1}{r(\mathcal{Q}^\ell_{-i})} + 1\right)$$

Denote $r := r(\mathcal{Q}^\ell_{-i})$, $\epsilon := \sqrt{\frac{6\log N + 2\log\max_\ell d_\ell}{n - \max_\ell d_\ell}}$ and $\mu := \mu(\mathcal{X}_\ell)$ we can further relax the bound into

$$\begin{aligned}|\langle x, \nu\rangle| \le& \frac{\mu + \delta\epsilon}{r - \epsilon\delta} + \frac{(\mu + \delta\epsilon)2\delta^2\epsilon}{r - \epsilon\delta}\left(\frac{1}{r} + 1\right)\\ &+ 2\lambda\delta\epsilon\left(\frac{1}{r} + 1\right) + 2\lambda\delta^2\epsilon\left(\frac{1}{r} + 1\right)\\ \le& \frac{\mu + \delta\epsilon + 3\lambda\delta(1+\delta)\epsilon}{r - \epsilon\delta} + 2\lambda\delta(1+\delta)\epsilon.\end{aligned}$$

Note that here in order to get rid of the higher order term $\frac{1}{r(r-\epsilon\delta)}$, we used $\delta < r$ and $\mu + \delta\epsilon < 1$ to construct $\frac{(\mu+\delta\epsilon)\delta^2\epsilon}{r(r-\delta\epsilon)} < \frac{\delta\epsilon}{r-\delta\epsilon}$ as in the proof of Theorem 1. Now impose the dual detection constraint on the upper bound, we get:

$$2\lambda\delta(1+\delta)\epsilon + \frac{\mu + \delta\epsilon + 3\lambda\delta(1+\delta)\epsilon}{r - \delta\epsilon} < 1.$$

Replace $\rho := \lambda\delta(1+\delta)$ and reorganize the inequality, we reach the desired condition:

$$\mu + 3\rho\epsilon + \delta\epsilon \le (1 - 2\rho\epsilon)(r - \delta\epsilon).$$

There are N^2 instances for each of the three events related to the consine value, apply union bound we get the failure probability $\frac{6}{N} + \frac{2L}{N^2} \le \frac{7}{N}$. This concludes the proof. □

B.1. Proof of Theorem 2

Lemma B.3 has already provided the separation condition. The things left are to find the range of λ and update the condition of δ.

The range of λ: Follow the same arguments in Section A.4 and Section A.5, re-derive the upper bound from the relationship in Lemma B.3 and substitute the tighter bound of δ_1 where applicable. Again let $r_\ell = \min_i r(Q^\ell_{-i})$, $\mu_\ell = \mu(\mathcal{X}_\ell)$ and $r = \min_\ell r_\ell$. We get the range of λ under random noise model:

$$\begin{cases}\lambda > \dfrac{1}{(r - \delta\epsilon)(1 - 3\delta) - 3\delta - 2\delta^2}\\ \lambda < \displaystyle\min_{\ell=1,\dots,L}\left\{\frac{r_\ell - \mu_\ell - 2\delta\epsilon}{\epsilon\delta(1+\delta)(3 + 2r_\ell - 2\delta\epsilon)}\right\} \vee \frac{2r}{\delta^2(r+1)}\end{cases} \tag{B.3}$$

Remark B.1. *A critical difference from the deterministic noise model is that now under the paradigm of small μ and δ, if $\delta > \epsilon$, the second term in the upper bound is actually tight. Then the valid range of λ is expanded an order to $\Theta(1/r) \le \lambda < \Theta(r/\delta^2)$.*

The condition of δ: Re-derive (A.19) using $\delta_1 \le \epsilon\delta$, we get:

$$\delta < \frac{r}{3r + 3 + \epsilon} \tag{B.4}$$

Likewise, we re-derive (A.21) from the new range of λ in (B.3). The first inequality in standard form is,

$$A\delta^3 + B\delta^2 + C\delta + D < 0$$

with

$$\begin{cases}A = 6\epsilon^2 - 6\epsilon,\\ B = -(3\epsilon + 4\epsilon^2 + \epsilon r_\ell - 2r_\ell + 6\epsilon r + 2\mu_\ell - 3\mu_\ell\epsilon),\\ C = 3r_\ell r + 3r_\ell + 3\epsilon r_\ell + 3\epsilon + 2\epsilon r - 3\mu_\ell r - 3\mu_\ell - \epsilon\mu_\ell,\\ D = -r(r_\ell - \mu_\ell),\end{cases}$$

apply the same trick of removing the negative μ term and define

$$f(r,\delta) := A\delta^2 + B\delta + C$$

such that the 3^{rd}-order polynomial inequality becomes $f(r,\delta)\delta < r(r_\ell - \mu_\ell)$.Rearrange the expressions and drop negative terms, we get

$$\begin{aligned}&f(r,\delta) < B\delta + C\\ =&-\left[3\epsilon + 4\epsilon^2 + 2\epsilon(r_\ell - \mu_\ell) + 6\epsilon r\right]\delta + 2(r_\ell - \mu_\ell)\delta\\ &+ [3(r_\ell - \mu_\ell)r + 3(r_\ell - \mu_\ell) + 3\epsilon(r_\ell - \mu_\ell) + 2\epsilon r + 3\epsilon]\\ &+ (r_\ell - \mu_\ell)\epsilon\delta + 2\mu_\ell\epsilon\delta - \mu_\ell\epsilon\\ <&3(r_\ell - \mu_\ell)r + 5(r_\ell - \mu_\ell) + 4\epsilon(r_\ell - \mu_\ell) + 2\epsilon r + 3\epsilon.\end{aligned}$$

Therefore, a sufficient condition of δ is

$$\delta < \frac{r(r_\ell - \mu_\ell)}{3(r_\ell - \mu_\ell)r + 5(r_\ell - \mu_\ell) + 4\epsilon(r_\ell - \mu_\ell) + 2\epsilon r + 3\epsilon}. \tag{B.5}$$

When $r > r_\ell - \mu_\ell$, we have $(r_\ell - \mu_\ell)/r < 1$. Then

$$\begin{aligned} (B.5) \Leftarrow \quad & \delta < \frac{r_\ell - \mu_\ell}{3(r_\ell - \mu_\ell) + 5 + \epsilon(4 + 2 + 3/r)} \\ \Leftarrow \quad & \delta < \frac{r_\ell - \mu_\ell}{3r + 5 + \epsilon(6 + 3/r)}. \end{aligned}$$

When $r < r_\ell - \mu_\ell$, we have $r/(r_\ell - \mu_\ell) < 1$. Since $r < r_\ell$,

$$\begin{aligned} (B.5) \Leftarrow \quad & \delta < \frac{r}{3r + 5 + \epsilon(4 + 2 + 3/(r_\ell - \mu_\ell))} \\ \Leftarrow \quad & \delta < \frac{r}{3r + 5 + \epsilon(6 + 3/r)} \end{aligned}$$

Combining the two cases, we have:

$$\delta < \frac{\min\{r, r_\ell - \mu_\ell\}}{3r + 5 + \epsilon(6 + 3/r)} \tag{B.6}$$

For the second inequality, the quadratic polynomial is now

$$(1 + 5r - 6r\epsilon)\delta^2 + (6r^2 + 2\epsilon r + 6r)\delta - 2r^2 < 0.$$

Check that $1 + 5r - 6r\epsilon > 0$. We solve the quadratic inequality and get a slightly stronger condition than (B.4), which is

$$\delta < \frac{r}{3r + 4 + \epsilon}. \tag{B.7}$$

Note that (B.6) $\Rightarrow$ (B.7), so (B.6) alone is sufficient. In fact, when $\epsilon(6r + 3)/r < 1$ or equivalently $r > 3\epsilon/(1 - 6\epsilon)$, which are almost always true, a neater expression is:

$$\delta < \frac{\min\{r, r_\ell - \mu_\ell\}}{3r + 6}.$$

Finally, as the condition needs to be satisfied for all ℓ, the output of the min function at the smallest bound is always $r_\ell - \mu_\ell$. This observation allows us to replace $\min\{r, r_\ell - \mu_\ell\}$ with simple $(r_\ell - \mu_\ell)$, which concludes the proof for Theorem 2.

B.2. Proof of Theorem 3

To prove Theorem 3, we only need to bound inradii r and incoherence parameter μ under the new assumptions, then plug into Theorem 2.

Lemma B.4 (Inradius bound of random samples)**.** *In random sampling setting, when each subspace is sampled $N_\ell = \kappa_\ell d_\ell$ data points randomly, we have:*

$$Pr\left\{ c(\kappa_\ell)\sqrt{\frac{\beta \log(\kappa_\ell)}{d_\ell}} \leq r(\mathcal{Q}_{-i}^{(\ell)}) \textit{ for all pairs } (\ell, i) \right\} \geq 1 - \sum_{\ell=1}^{L} N_\ell e^{-d_\ell^\beta N_\ell^{1-\beta}}$$

This is extracted from Section-7.2.1 of Soltanolkotabi & Candes (2012). $\kappa_\ell = (N_\ell - 1)/d_\ell$ is the relative number of iid samples. $c(\kappa)$ is some positive value for all $\kappa > 1$ and for a numerical value κ_0, if $\kappa > \kappa_0$, we can take $c(\kappa) = \frac{1}{\sqrt{8}}$. Take $\beta = 0.5$, we get the required bound of r in Theorem 3.

Lemma B.5 (Incoherence bound)**.** *In deterministic subspaces/random sampling setting, the subspace incoherence is bounded from above:*

$$\begin{aligned} Pr\Big\{ \mu(\mathcal{X}_\ell) \leq t\left(\log[(N_{\ell_1} + 1)N_{\ell_2}] + \log L\right) \frac{\operatorname{aff}(S_{\ell_1}, S_{\ell_2})}{\sqrt{d_{\ell_1}}\sqrt{d_{\ell_2}}} \\ \textit{for all pairs}(\ell_1, \ell_2) \textit{ with } \ell_1 \neq \ell_2 \Big\} \\ \geq 1 - \frac{1}{L^2} \sum_{\ell_1 \neq \ell_2} \frac{1}{(N_{\ell_1} + 1)N_{\ell_2}} e^{-\frac{t}{4}} \end{aligned}$$

B.2.1. Proof of Lemma B.5

The proof is an extension of the same proof in Soltanolkotabi & Candes (2012). First we will show that when noise $z_i^{(\ell)}$ is spherical symmetric, and clean data points $y_i^{(\ell)}$ has iid uniform random direction, projected dual directions $v_i^{(\ell)}$ also follows uniform random distribution.

Now we will prove the claim. First by definition,

$$v_i^{(\ell)} = v(x_i^{(\ell)}, X_{-i}^{(\ell)}, \mathcal{S}_\ell, \lambda) = \frac{\mathbb{P}_{S_\ell}\nu}{\|\mathbb{P}_{S_\ell}\nu\|} = \frac{\nu_1}{\|\nu_1\|}.$$

ν is the unique optimal solution of $\mathbf{D}_1$ (A.5). Fix λ, $\mathbf{D}_1$ depends on two inputs, so we denote $\nu(x, X)$ and consider ν a function. Moreover, $\nu_1 = \mathbb{P}_{\mathcal{S}}\nu$ and $\nu_2 = \mathbb{P}_{\mathcal{S}^\perp}\nu$. Let $U \in n \times d$ be a set of orthonormal basis of d-dimensional subspace $\mathcal{S}$ and a rotation matrix $R \in \mathbb{R}^{d \times d}$. Then rotation matrix within subspace is hence URU^T.

$$\begin{aligned} x_1 :=& \mathbb{P}_{\mathcal{S}}x = y + z_1 \sim URU^T y + URU^T z_1 \\ x_2 :=& \mathbb{P}_{\mathcal{S}^\perp}x = z_2 \end{aligned}$$

As y is distributed uniformly on unit sphere of $\mathcal{S}$, and z is spherical symmetric noise(hence z_1 and z_2 are also spherical symmetric in subspace), for any fixed $\|x_1\|$, the distribution is uniform on the sphere. It suffices to show the uniform distribution of ν_1 with fixed $\|x_1\|$.

Since inner product $\langle x, \nu\rangle = \langle x_1, \nu_1\rangle + \langle x_2, \nu_2\rangle$, we argue that if ν is optimal solution of

$$\max_{\nu} \langle x, \nu\rangle - \frac{1}{2\lambda}\nu^T\nu, \quad \text{subject to:} \quad \|X^T\nu\|_\infty \le 1,$$

then the optimal solution of R-transformed optimization

$$\begin{aligned} &\max_{\nu} \langle URU^T x_1 + x_2, \nu\rangle - \frac{1}{2\lambda}\nu^T\nu, \\ &\text{subject to:} \quad \|(URU^T X_1 + X_2)^T\nu\|_\infty \le 1, \end{aligned}$$

is merely the transformed ν under the same R:

$$\begin{aligned} \nu(R) &= \nu(URU^T x_1 + x_2, URU^T X_1 + X_2) \\ &= URU^T\nu_1(x, X) + \nu_2(x, X) = URU^T\nu_1 + \nu_2. \end{aligned} \tag{B.8}$$

To verify the argument, check that $\nu^T\nu = \nu(R)^T\nu(R)$ and

$$\begin{aligned} \langle URU^T x_1 + x_2, \nu(R)\rangle &= \langle URU^T x_1, URU^T\nu_1\rangle + \langle x_1, \nu_2\rangle \\ &= \langle x, \nu\rangle \end{aligned}$$

for all inner products in both objective function and constraints, preserving the optimality.

By projecting (B.8) to subspace, we show that operator $v(x, X, S)$ is linear *vis a vis* subspace rotation URU^T, i.e.,

$$v(R) = \frac{\mathbb{P}_{S_\ell}\nu(R)}{\|\mathbb{P}_{S_\ell}\nu(R)\|} = \frac{URU^T\nu_1}{\|URU^T\nu_1\|} = URU^T v. \tag{B.9}$$

On the other hand, we know that

$$v(R) = v(URU^T x_1 + x_2, URU^T X_1 + X_2, \mathcal{S}) \sim v(x, X, \mathcal{S}), \tag{B.10}$$

where $A \sim B$ means that the random variables A and B follows the same distribution. When $\|x_1\|$ is fixed and each columns in X_1 has fixed magnitudes, $URU^T x_1 \sim x_1$ and $URU^T X_1 \sim X_1$. Since (x_1, X_1) and (x_2, X_2) are independent, we can also marginalize out the distribution of x_2 and X_2 by considering fixed (x_2, X_2). Combining (B.9) and (B.10), we conclude that for any rotation R,

$$v_i^{(\ell)}(R) \sim URU^T v_i^{(\ell)}.$$

Now integrate the marginal probability of $v_i^{(\ell)}$ over $\|x_{i1}^{\ell}\|$, every column's magnitude of X_{-i1}^{ℓ} and all (x_2, X_2), we showed that the overall distribution of $v_i^{(\ell)}$ is indeed uniformly distributed in the unit sphere of $\mathcal{S}$.

After this key step, the rest is identical to Lemma 7.5 of Soltanolkotabi & Candes (2012). The idea is to use Lemma B.1(upper bound of area of spherical caps) to bound pairwise inner product and Borell's inequality to bound the deviation from expected consine canonical angles, namely, $\|U^{(k)^T}U^{(\ell)}\|_F/\sqrt{d_\ell}$.

B.3. Proof of Theorem 4

The proof of this theorem is also an invocation of Theorem 2 with specific inradii bound and incoherence bound. The bound of inradii is exactly Lemma B.4 with $\beta = 0.5$, $\kappa_\ell = \kappa$, $d_\ell = d$. The bound of incoherence is given by the following Lemma that is extracted from Step 2 of Section 7.3 in Soltanolkotabi & Candes (2012).

Lemma B.6 (Incoherence bound of random subspaces)**.** *In random subspaces setting, the projected subspace incoherence is bounded from above:*

$$Pr\left\{\mu(\mathcal{X}_\ell) \le \sqrt{\frac{6\log N}{n}} \text{ for all } \ell\right\} \ge 1 - \frac{2}{N}.$$

Now that we have shown that projected dual directions are randomly distributed in their respective subspace, as the subspaces themselves are randomly generated, all clean data points y and projected dual direction v from different subspaces can be considered iid generated from the ambient space. The proof of Lemma B.6 follows by simply applying Lemma B.1 and union bound across all N^2 events.

By plug in these expressions into Theorem 2, we showed that it holds with high probability as long as the conditions in Theorem 4 is true.

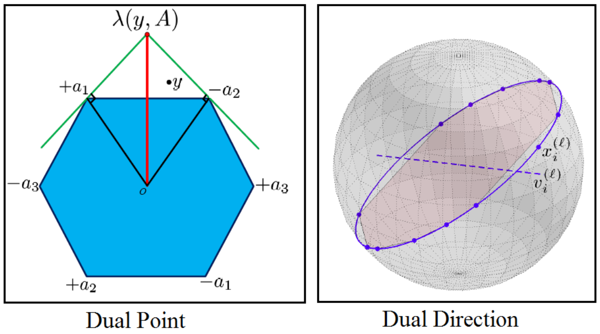
C. Geometric interpretations

In this section, we attempt to give some geometric interpretation of the problem so that the results stated in this paper can be better understood and at the same time, reveal the novelties of our analysis over Soltanolkotabi & Candes (2012). All figures in this section are drawn with "geom3d" (Legland, 2009) and "GBT7.3" (Veres, 2006) in Matlab.

We start with an illustration of the projected dual direction in contrast to the original dual direction(Soltanolkotabi & Candes, 2012).

Dual direction v.s. Projected dual direction:

An illustration of original dual direction is given in Figure C.1 for data point y.



Dual Point

Dual Direction

Figure C.1. The illustration of dual direction in Soltanolkotabi & Candes (2012).

The projected dual direction can be easier understood algebraically. By definition, it is the projected optimal solution of (A.5) to the true subspace. To see it more clearly, we plot the feasible region of ν in Figure C.2 (b), and the projection of the feasible region in Figure C.3. As (A.5) is not an LP (it has a quadratic term in the objective function), projected dual direction cannot be easily determined geometrically as in Figure C.1. Nevertheless, it turns out to be sufficient to know the feasible region and the optimality of the solution.

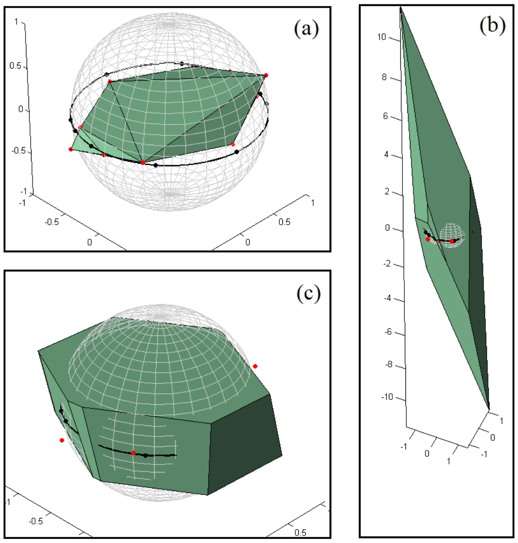


Figure C.2. Illustration of **(a)** the convex hull of noisy data points, **(b)** its polar set and **(c)** the intersection of polar set and $\|\nu_2\|$ bound. The polar set (b) defines the feasible region of (A.5). It is clear that ν_2 can take very large value in (b) if we only consider feasibility. By considering optimality, we know the optimal ν must be inside the region in (c).

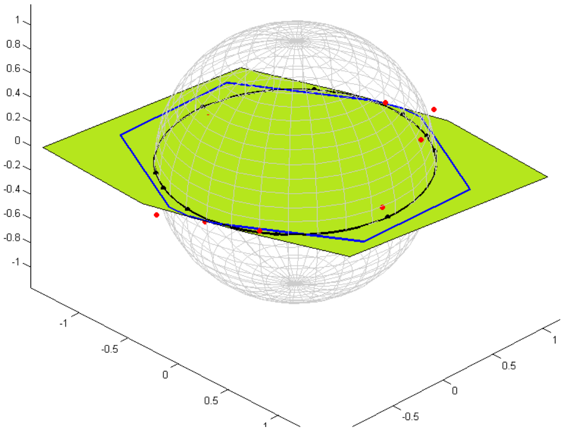


Figure C.3. The projection of the polar set (the green area) in comparison to the projection of the polar set with $\|\nu_2\|$ bound (the blue polygon). It is clear that the latter is much smaller.

Magnitude of dual variable ν:

A critical step of our proof is to bound the magnitude of $\|\nu_1\|$ and $\|\nu_2\|$. This is a simple task in the noiseless case as Soltanolkotabi and Candes merely take the circumradius of the full feasible region as a bound. This is sufficient because the feasible region is a cylinder perpendicular to the subspace and there is no harm choosing only solutions within the intersection of the cylinder and the subspace. Indeed, in noiseless case, we can choose arbitrary ν_2 because $Y^T(\nu_1 + \nu_2) = Y^T\nu_1$.

In the noisy case however, the problem becomes a bit involved. Instead of a cylinder, the feasible region is now a spindle shaped polytope (see Figure C.2(b)) and the choice of ν_2 has an impact on the objective value. That is why we need to consider the optimality condition and give $\|\nu_2\|$ a bound.

In fact, noise may tilt the direction of the feasible region (especially when the noise is adversarial). As $\|\nu_2\|$ grows, $\|\nu_1\|$ can potentially get large too. Our bound of $\|\nu_1\|$ reflects precisely the case as it is linearly dependent on $\|\nu_2\|$ (see (A.11)). We remark that in the case of random noise, the dependency on $\|\nu_2\|$ becomes much weaker (see the proof of Lemma B.3).

Geometrically, the bound of ν_2 can be considered a cylinder[1] (ℓ_2 constrained in the $\mathcal{S}^\perp$ and unbounded in $\mathcal{S}$ subspace) that intersect the spindle shaped feasible region, so that we know the optimal ν may never be

[1] In the simple illustration, the cylinder is in fact just the sandwich region $|z| \leq$ some bound.

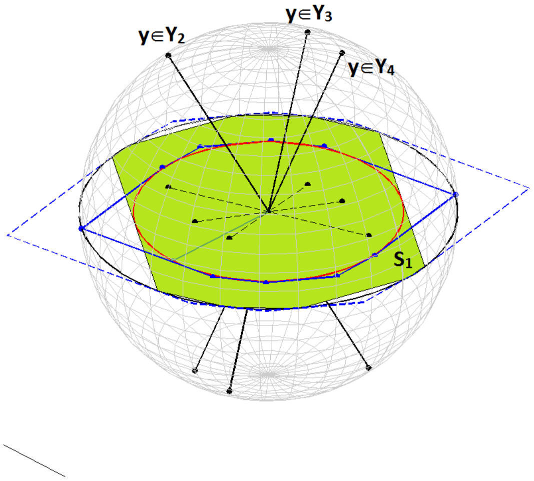


Figure C.4. Noiseless SSC: Theorem 2.5 of Soltanolkotabi & Candes (2012) suggests that the projection of external data points must fall inside the solid blue polygon, which is the intersection of halfspaces defined by dual directions (blue dots) that are tangent planes of the red inscribing sphere.

at the tips of the spindle (see Figure C.2 and C.3). Algebraically, we can consider this as an effect of the quadratic penalty term of ν in the (A.5).

The guarantee in Theorem 1:

The geometric interpretation and comparison of the noiseless guarantee and our noisy guarantee are given in Figure C.4 and C.5. Geometrically, noise reduces the successful region (the solid blue polygon) in two ways. One is subtractive, in a sense that the inradius is smaller (see the bound of $\|\nu_1\|$); the other is multiplicative, as the entire successful region shrinks with a factor related to noise level (something like $1 - f(\delta)$). Readers may refer to (A.16) for an algebraic point of view.

The subtractive effect can also be interpreted in the robust optimization point of view, where the projection of every points inside the uncertainty set (the red balls in Figure C.5) must fall into the successful region (the dashed red polygon).

Either way, it is clear that the error Lasso-SSC can provably tolerate is proportional to the geometric gap $r - \mu$ given in the noiseless case.

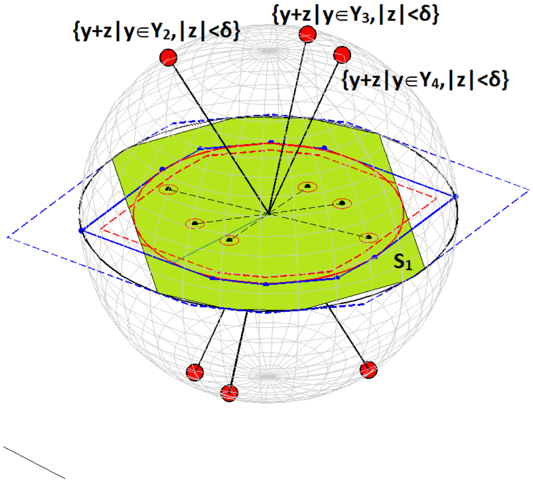

$\{y+z|y\in Y_2,|z|<\delta\}$
$\{y+z|y\in Y_3,|z|<\delta\}$
$\{y+z|y\in Y_4,|z|<\delta\}$
S_1

Figure C.5. Noisy Lasso-SSC: The guarantee of Theorem 1 means that the whole red sphere of each external data points must fall inside the dashed red polygon, which is smaller than the blue polygon by a factor related to the noise level.

D. Numerical algorithm to solve Matrix-Lasso-SSC

In this section we outline the steps of solving the matrix version of Lasso-SSC below ((3.1) in the paper)

$$\begin{aligned}&\min_C \|C\|_1 + \frac{\lambda}{2}\|X - XC\|_F^2\\ &\text{s.t.} \quad \operatorname{diag}(C) = 0,\end{aligned} \tag{D.1}$$

While this convex optimization can be solved by some off-the-shelf general purpose solver such as CVX, such approach is usually slow and non-scalable. An ADMM (Boyd et al., 2011) version of the problem is described here for fast computation. It solves an equivalent optimization program

$$\begin{aligned}&\min_C \|C\|_1 + \frac{\lambda}{2}\|X - XJ\|_F^2\\ &\text{s.t.} \quad J = C - \operatorname{diag}(C).\end{aligned} \tag{D.2}$$

We add to the Lagrangian with an additional quadratic penalty term for the equality constraint and get the augmented Lagrangian

$$\begin{aligned}\mathcal{L} =&\|C\|_1 + \frac{\lambda}{2}\|X - XJ\|_F^2 + \frac{\mu}{2}\|J - C + \operatorname{diag}(C)\|_F^2\\ &+ tr(\Lambda^T(J - C + \operatorname{diag}(C))),\end{aligned}$$

where Λ is the dual variable and μ is a parameter. Optimization is done by alternatingly optimizing over J,

Algorithm 1 Matrix-Lasso-SSC

```
Input: Data points as columns in $X \in \mathbb{R}^{n \times N}$, tradeoff parameter $\lambda$, numerical parameters $\mu_0$ and $\rho$.
Initialize $C = 0$, $J = 0$, $\Lambda = 0$, $k = 0$.
while not converged do
    1. Update $J$ by
       $J = (\lambda X^T X + \mu_k I)^{-1}(\lambda X^T X + \mu_k C - \Lambda).$
    2. Update $C$ by
       $C^{'} = \mathrm{SoftThresh}_{\frac{1}{\mu_k}}\left(J + \Lambda/\mu_k\right),$
       $C = C^{'} - \mathrm{diag}(C^{'}).$
    3. Update $\Lambda$ by
       $\Lambda = \Lambda + \mu_k(J - C)$
    4. Update parameter $\mu_{k+1} = \rho\mu_k$.
    5. Iterate $k = k + 1$;
end while
Output: Affinity matrix $W = |C| + |C|^T$
```

C and Λ until convergence. The update steps are derived by solving $\partial\mathcal{L}/\partial J = 0$ and $\partial\mathcal{L}/\partial C = 0$, it's non-differentiable for C at origin so we use the now standard soft-thresholding operator(Donoho, 1995). For both variables, the solution is in closed-form. For the update of Λ, it is simply gradient descent. For details of the ADMM algorithm and its guarantee, please refer to Boyd et al. (2011). To accelerate the convergence, it is possible to introduce a parameter ρ and increase μ by $\mu = \rho\mu$ at every iteration. The full algorithm is summarized in Algorithm 1.

Note that for the special case when $\rho = 1$, the inverse of $(\lambda Y^T Y + \mu I)$ can be pre-computed, such that the iteration is linear time. Empirically, we found it good to set $\mu = \lambda$ and it takes roughly 50-100 iterations to converge to a sufficiently good points. We remark that the matrix version of the algorithm is much faster than column-by-column ADMM-Lasso especially for the cases when $N > n$. See the experiments.

We would like to point out that Elhamifar & Vidal (2012) had formulated a more general version of SSC to account for not only noisy but also sparse corruptions in the Appendix of their arxiv paper while we were preparing for submission. The ADMM algorithm for Matrix-Lasso-SSC described here can be considered as a special case of the Algorithm 2 in their paper.

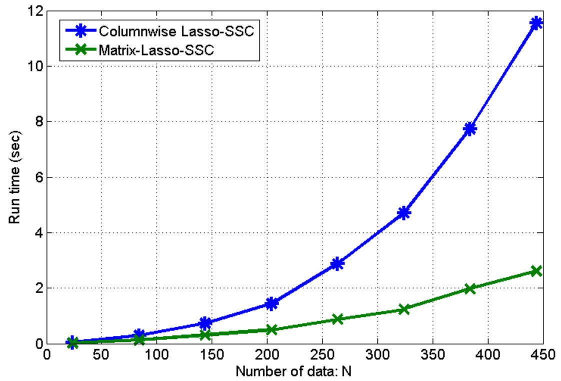


Figure D.1. Run time comparison with increasing number of data. Simulated with $n = 100, d = 4, L = 3, \sigma = 0.2$, κ increases from 2 to 40 such that the number of data goes from 24- 480. It appears that the matrix version scales better with increasing number of data compared to columnwise LASSO.

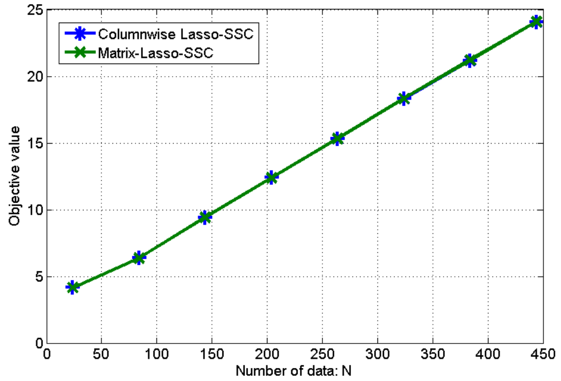


Figure D.2. Objective value comparison with increasing number of data. Simulated with $n = 100, d = 4, L = 3, \sigma = 0.2$, κ increases from 2 to 40 such that the number of data goes from 24- 480. The objective value obtained at stop points of two algorithms are nearly the same.

.

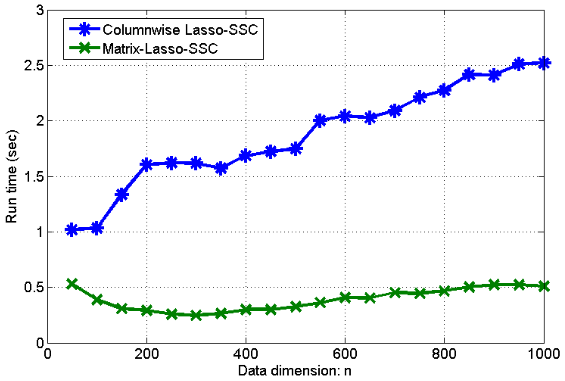


Figure D.3. Run time comparison with increasing number of data. Simulated with $\kappa = 5, d = 4, L = 3, \sigma = 0.2$, ambient dimension n increases from 50 to 1000. Note that the dependence on dimension is weak at the scale due to the fast vectorized computation. Nevertheless, it is clear that the matrix version of SSC runs faster.

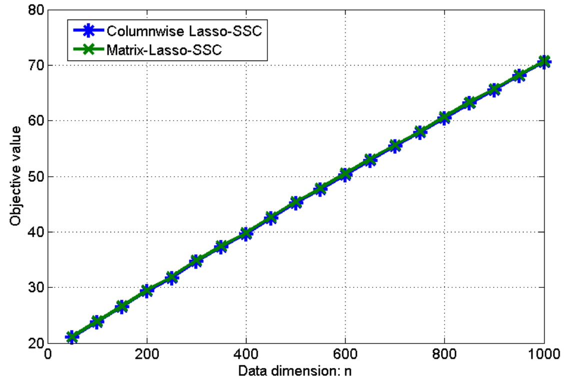


Figure D.4. Objective value comparison with increasing number of data. Simulated with $\kappa = 5, d = 4, L = 3, \sigma = 0.2$, ambient dimension n increases from 50 to 1000. The objective value obtained at stop points of two algorithms are nearly the same.

References

Ball, K. An elementary introduction to modern convex geometry. *Flavors of geometry*, 31:1–58, 1997.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

Dasgupta, S. and Gupta, A. An elementary proof of a theorem of johnson and lindenstrauss. *Random Structures & Algorithms*, 22(1):60–65, 2002.

Donoho, D.L. De-noising by soft-thresholding. *Information Theory, IEEE Transactions on*, 41(3):613–627, 1995.

Elhamifar, E. and Vidal, R. Sparse subspace clustering: Algorithm, theory, and applications. *arXiv preprint arXiv:1203.1005*, 2012.

Legland, David. geom3d toolbox [computer software], 2009. URL http://www.mathworks.com/matlabcentral/fileexchange/24484-geom3d.

Soltanolkotabi, M. and Candes, E.J. A geometric analysis of subspace clustering with outliers. *To appear in Annals of Statistics*, 2012.

Veres, Sandy. Geometric bounding toolbox 7.3 [computer software], 2006. URL http://www.mathworks.com/matlabcentral/fileexchange/11678-polyhedron-and-polytope-computations.