Blind Deconvolution-Notes

Han-Wen Kuo

Abstract

I Quick Note

I.I Riemannian Gradient and Hessian

Let sparse signal $\mathbf{x}_0 \in \mathbb{R}^m$, indexed by $\mathbf{x} = \{x_0, \ldots, x_{m-1}\}$, filter $\mathbf{a}_0 \in \mathbb{S}^{k-1}$ with $k \ll m$. The sparse signal $\mathbf{x}_0$ is generated under Bernoulli-Gaussian model while as kernel function $\mathbf{a}_0$ being uniform random on sphere. The observation signal $\mathbf{y} \in \mathbb{R}^m$ is generated by $\mathbf{y} = \tilde{\mathbf{a}}_0 \otimes \mathbf{x}_0$ We are attempting to minimize the following function:

$$\mathbf{\Phi}(\mathbf{a}) = \min_{\mathbf{x}} \mathbf{\Psi}(\mathbf{x}, \mathbf{a}) \tag{I.1}$$

where

$$\mathbf{\Psi}(\mathbf{x}, \mathbf{a}) := \frac{1}{2} \|\tilde{\mathbf{a}} \otimes \mathbf{x} - \mathbf{y}\|_2^2 + \lambda r_\mu(\mathbf{x}) - \frac{1}{2} \|\mathbf{y}\|_2^2 \tag{I.2}$$

For analysis purpose, we assume the sparsity inducing loss function r_μ to be Huber-loss h_μ, namely for $i \in [m]$:

$$h_\mu(\mathbf{x}) = \sum_{i=1}^{m} h_\mu(x_i), \quad h_\mu(x_i) := \begin{cases} x_i^2/(2\mu) & |x_i| \le \mu \\ |x_i| - \mu/2 & |x_i| \ge \mu \end{cases} \tag{I.3}$$

It is apparent that the derivative of Huber loss function can be characterized as:

$$\nabla h_\mu(\mathbf{x})_i = \begin{cases} x_i/\mu & |x_i| \le \mu \\ \operatorname{sgn}(x_i) & |x_i| \ge \mu \end{cases} \tag{I.4}$$

We denote the vector $\nu(\mathbf{x}) \in \{-1, 0, 1\}^m$ to represent the sign of $\mathbf{x}$ such that

$$\nu(\mathbf{x})_i = \begin{cases} 1 & x_i > \mu \\ 0 & -\mu \le x_i \le \mu \\ -1 & x_i < -\mu \end{cases} \tag{I.5}$$

Define the set R_ν as a subset of sphere $\mathbb{S}^{k-1}$ as following:

$$R_\nu := \left\{ \mathbf{a} \in \mathbb{S}^{k-1} : \nu(\operatorname*{argmin}_{\mathbf{x}} \mathbf{\Psi}(\mathbf{x}, \mathbf{a})) = \nu \right\} \tag{I.6}$$

In light of our observation with loss function r_μ being ℓ_1 penalty function on relation between μ and R_0 has shown that when μ increases, the region of R_0 also increases, and have yet spot any saddle point in R_0^c. It seems reasonable to us that we should look at the relation between R_0 and μ under simple setup, say $\mathbf{y} = \tilde{\mathbf{a}}_0$.

First we need to specify the optimality condition of $\mathbf{x_a} := \operatorname{argmin}_{\mathbf{x}} \mathbf{\Psi}(\mathbf{x}, \mathbf{a})$ with $\mathbf{a} \in R_\nu$, let $\operatorname{supp}(\nu) =: \mathcal{I} \subset [m]$, we gain :

$$\mathbf{C}_{\mathbf{a}}^{*}(\tilde{\mathbf{a}} \otimes \mathbf{x_a} - \mathbf{y}) + \lambda \nabla h_{\mu}(\mathbf{x_a}) = \mathbf{0} \tag{I.7}$$

$$\iff (\mathbf{C}_{\mathbf{a}}^{*}\mathbf{C_a} + \frac{\lambda}{\mu}\mathbf{V}_{\mathcal{I}^c}\mathbf{V}_{\mathcal{I}^c}^{*})\mathbf{x_a} = \mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} - \lambda \mathbf{V}_{\mathcal{I}}\mathbf{V}_{\mathcal{I}}^{*}\mathrm{sgn}\,(\mathbf{x_a}) \tag{I.8}$$

$$\iff (\mathbf{C}_{\mathbf{a}}^{*}\mathbf{C_a} + \frac{\lambda}{\mu}\mathbf{I})\mathbf{x_a} = \mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} + \lambda \mathbf{V}_{\mathcal{I}}\mathbf{V}_{\mathcal{I}}^{*}(\frac{\mathbf{x_a}}{\mu} - \mathrm{sgn}\,(\mathbf{x_a})) \tag{I.9}$$

$$\iff (\mathbf{C}_{\mathbf{a}}^{*}\mathbf{C_a} + \frac{\lambda}{\mu}\mathbf{I})\mathbf{x_a} = \mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} + \frac{\lambda}{\mu}\mathrm{SOFT}_{\mu}(\mathbf{x_a}) \tag{I.10}$$

$$\iff \begin{cases} \mathbf{V}_{\mathcal{I}}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{C_a}\mathbf{x_a} = \mathbf{V}_{\mathcal{I}}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} - \lambda \mathbf{V}_{\mathcal{I}}^{*}\mathrm{sgn}\,(\mathbf{x_a}) \\ \mathbf{V}_{\mathcal{I}^c}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{C_a}\mathbf{x_a} = \mathbf{V}_{\mathcal{I}^c}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} - \frac{\lambda}{\mu}\mathbf{V}_{\mathcal{I}^c}^{*}\mathbf{x_a} \end{cases} \tag{I.11}$$

$$\implies \frac{1}{2}\mathbf{x}_{\mathbf{a}}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{C_a}\mathbf{x_a} = \frac{1}{2}\mathbf{x}_{\mathbf{a}}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} - \frac{\lambda}{2}\mathbf{x}_{\mathbf{a}}^{*}(\nabla h_{\mu}(\mathbf{x_a})) \tag{I.12}$$

By (I.8), we define the following two vectors:

$$\mathbf{G} := \mathbf{C}_{\mathbf{a}}^{*}\mathbf{C_a} + \frac{\lambda}{\mu}\mathbf{V}_{\mathcal{I}^c}\mathbf{V}_{\mathcal{I}^c}^{*}, \qquad \xi := \mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} - \lambda \mathbf{V}_{\mathcal{I}}\mathbf{V}_{\mathcal{I}}^{*}\mathrm{sgn}\,(\mathbf{x_a}) \tag{I.13}$$

The optimal vector $\mathbf{x_a}$ can be realized as:

$$\mathbf{x_a} = \mathbf{G}^{-1}\xi \tag{I.14}$$

$$h_{\mu}(\mathbf{x_a}) = \frac{1}{2\mu}\left\|\mathbf{V}_{\mathcal{I}^c}^{*}\mathbf{x_a}\right\|_2^2 + \left\|\mathbf{V}_{\mathcal{I}}^{*}\mathbf{x_a}\right\|_1 - \frac{\mu}{2}\left|\mathcal{I}\right| \tag{I.15}$$

The condition (I.12) helps us to simplify the objective function $\mathbf{\Phi}$:

$$\mathbf{\Phi}(\mathbf{a}) = \mathbf{\Psi}(\mathbf{x_a}, \mathbf{a}) = \frac{1}{2}\mathbf{x}_{\mathbf{a}}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{C_a}\mathbf{x_a} - \mathbf{x}_{\mathbf{a}}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} + \lambda h_{\mu}(\mathbf{x_a}) \tag{I.16}$$

$$= -\frac{1}{2}\mathbf{x}_{\mathbf{a}}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} + \lambda h_{\mu}(\mathbf{x_a}) - \frac{\lambda}{2}\mathbf{x}_{\mathbf{a}}^{*}(\nabla h_{\mu}(\mathbf{x_a})) \tag{I.17}$$

$$= -\frac{1}{2}\mathbf{x}_{\mathbf{a}}^{*}\mathbf{C}_{\mathbf{a}}^{*}\mathbf{y} + \frac{\lambda}{2}\sum_{i\in\mathcal{I}}(|x_i| - \mu) \tag{I.18}$$

$$= -\frac{1}{2}\mathbf{x}_{\mathbf{a}}^{*}(\mathbf{C}_{\mathbf{a}}^{*}\mathbf{C_a} + \frac{\lambda}{\mu}\mathbf{V}_{\mathcal{I}^c}\mathbf{V}_{\mathcal{I}^c}^{*})\mathbf{x_a} - \frac{\lambda\mu}{2}\left|\mathcal{I}\right| \tag{I.19}$$

$$= -\frac{1}{2}\xi^{*}\mathbf{G}^{-1}\xi - \frac{\lambda\mu}{2}\left|\mathcal{I}\right| \tag{I.20}$$

Now we need to characterize the gradient of $\mathbf{\Phi}$ for $\mathbf{a} \in R_{\nu}$, observe that

$$\frac{\partial \mathbf{\Phi}}{\partial \mathbf{a}} = -\xi^{*}\mathbf{G}^{-1}\frac{\partial \xi}{\partial \mathbf{a}} + \frac{1}{2}\xi^{*}\mathbf{G}^{-1}\frac{\partial \mathbf{G}}{\partial \mathbf{a}}\mathbf{G}^{-1}\xi \tag{I.21}$$

whereas

$$\frac{\partial \xi}{\partial \mathbf{a}} = \check{\mathbf{C}}_{\mathbf{y}}, \quad \frac{\partial \mathbf{G}}{\partial a_i} = \mathbf{C}_{\mathbf{e}_i}^{*}\mathbf{C_a} + \mathbf{C}_{\mathbf{a}}^{*}\mathbf{C}_{\mathbf{e}_i} \tag{I.22}$$

then

$$\frac{\partial \mathbf{\Phi}}{\partial a_i} = -\mathbf{x}_{\mathbf{a}}^{*}\check{\mathbf{C}}_{\mathbf{y}}\mathbf{e}_i + \frac{1}{2}\mathbf{x}_{\mathbf{a}}^{*}\frac{\partial \mathbf{G}}{\partial a_i}\mathbf{x_a} \tag{I.23}$$

$$= -\mathbf{x}_{\mathbf{a}}^{*}\check{\mathbf{C}}_{\mathbf{y}}\mathbf{e}_i + \frac{1}{2}\mathbf{x}_{\mathbf{a}}^{*}(\mathbf{C_a}\check{\mathbf{C}}_{\mathbf{x_a}} + \mathbf{C}_{\mathbf{a}}^{*}\mathbf{C}_{\mathbf{x_a}})\mathbf{e}_i \tag{I.24}$$

$$= -\mathbf{x}_{\mathbf{a}}^{*}\check{\mathbf{C}}_{\mathbf{y}}\mathbf{e}_i + \mathbf{x}_{\mathbf{a}}^{*}\check{\mathbf{C}}_{\mathbf{x_a}}\mathbf{C_a}\mathbf{e}_i \tag{I.25}$$

$$\frac{\partial \mathbf{\Phi}}{\partial \mathbf{a}} = \iota^{*}(\check{\mathbf{C}}_{\mathbf{a}}\check{\mathbf{C}}_{\mathbf{x_a}} - \check{\mathbf{C}}_{\mathbf{y}})\mathbf{x_a} \tag{I.26}$$

$$= \iota^{*}(\check{\mathbf{C}}_{\mathbf{a}\otimes\mathbf{x_a}} - \check{\mathbf{C}}_{\mathbf{y}})\mathbf{x_a} \tag{I.27}$$

Which is the same as ℓ_1 penalty function. Similarly:

$$\langle \mathbf{a}, \nabla \mathbf{\Phi}(\mathbf{a}) \rangle = \tilde{\mathbf{a}}^* \mathbf{C}_\mathbf{a} \mathbf{C}_{\mathbf{x_a}}^* \mathbf{x_a} - \mathbf{y}^* \mathbf{C}_\mathbf{a} \mathbf{x_a} \tag{I.28}$$

$$= \lambda \sum_{i \in \mathcal{I}} (|x_i| - \mu) - 2\lambda h_\mu(\mathbf{x}) \tag{I.29}$$

$$= -\frac{\lambda}{\mu} \left\| \mathbf{V}_{\mathcal{I}^c}^* \mathbf{x_a} \right\|_2^2 - \lambda \left\| \mathbf{V}_\mathcal{I}^* \mathbf{x_a} \right\|_1 \tag{I.30}$$

Now we need to establish the twice derivative of function $\mathbf{\Phi}(\mathbf{a})$:

$$\frac{\partial \mathbf{\Phi}}{\partial a_i \partial a_j} = -\frac{\partial \xi^*}{\partial a_j} \mathbf{G}^{-1} \frac{\partial \xi}{\partial a_i} - \frac{\partial \xi^*}{\partial a_i} \frac{\partial \mathbf{G}^{-1}}{\partial a_j} \xi - \frac{\partial \xi^*}{\partial a_j} \frac{\partial \mathbf{G}^{-1}}{\partial a_i} \xi - \frac{1}{2} \xi^* \frac{\partial \mathbf{G}^{-1}}{\partial a_i \partial a_j} \xi \tag{I.31}$$

$$= -\mathbf{e}_i^* \check{\mathbf{C}}_\mathbf{y} \mathbf{G}^{-1} \check{\mathbf{C}}_\mathbf{y} \mathbf{e}_j + \mathbf{e}_i^* \check{\mathbf{C}}_\mathbf{y} \mathbf{G}^{-1} (\mathbf{C}_\mathbf{a}^* \mathbf{C}_{\mathbf{x_a}} + \mathbf{C}_\mathbf{a} \check{\mathbf{C}}_{\mathbf{x_a}}) \mathbf{e}_j + \mathbf{e}_i^* (\mathbf{C}_{\mathbf{x_a}}^* \mathbf{C}_\mathbf{a} + \check{\mathbf{C}}_{\mathbf{x_a}} \mathbf{C}_\mathbf{a}^*) \mathbf{G}^{-1} \check{\mathbf{C}}_\mathbf{y} \mathbf{e}_j$$
$$- \mathbf{e}_i^* (\mathbf{C}_{\mathbf{x_a}}^* \mathbf{C}_\mathbf{a} + \check{\mathbf{C}}_{\mathbf{x_a}} \mathbf{C}_\mathbf{a}^*) \mathbf{G}^{-1} (\mathbf{C}_\mathbf{a}^* \mathbf{C}_{\mathbf{x_a}} + \mathbf{C}_\mathbf{a} \check{\mathbf{C}}_{\mathbf{x_a}}) \mathbf{e}_j + \mathbf{e}_i^* \mathbf{C}_{\mathbf{x_a}}^* \mathbf{C}_{\mathbf{x_a}} \mathbf{e}_j \tag{I.32}$$

$$\nabla^2 \mathbf{\Phi}(\mathbf{a}) = -\iota^* (\mathbf{C}_{\mathbf{x_a}}^* \mathbf{C}_\mathbf{a} + \check{\mathbf{C}}_{\mathbf{x_a}} \mathbf{C}_\mathbf{a}^* - \check{\mathbf{C}}_\mathbf{y}) \mathbf{G}^{-1} (\mathbf{C}_\mathbf{a}^* \mathbf{C}_{\mathbf{x_a}} + \mathbf{C}_\mathbf{a} \check{\mathbf{C}}_{\mathbf{x_a}} - \check{\mathbf{C}}_\mathbf{y}) \iota + \iota^* \mathbf{C}_{\mathbf{x_a}}^* \mathbf{C}_{\mathbf{x_a}} \iota \tag{I.33}$$

$$= -\iota^* (\check{\mathbf{C}}_\mathbf{a} (\mathbf{C}_{\mathbf{x_a}}^* + \check{\mathbf{C}}_{\mathbf{x_a}}) - \check{\mathbf{C}}_\mathbf{y}) \mathbf{G}^{-1} ((\check{\mathbf{C}}_{\mathbf{x_a}} + \mathbf{C}_{\mathbf{x_a}}) \check{\mathbf{C}}_\mathbf{a} - \check{\mathbf{C}}_\mathbf{y}) \iota + \iota^* \mathbf{C}_{\mathbf{x_a}}^* \mathbf{C}_{\mathbf{x_a}} \iota \tag{I.34}$$

Which is also the same as ℓ_1 penalty objective function. We are ready to establish the Riemannian gradient and Hessian of $\mathbf{\Phi}$ at $\mathbf{a}$:

$$\text{grad}[\mathbf{\Phi}](\mathbf{a}) = P_{\mathbf{a}^\perp} \iota^* (\check{\mathbf{C}}_{\mathbf{a} \otimes \mathbf{x_a}} - \check{\mathbf{C}}_\mathbf{y}) \mathbf{x_a} \tag{I.35}$$

$$\text{Hess}[\mathbf{\Phi}](\mathbf{a}) = P_{\mathbf{a}^\perp} \iota^* \left(\nabla^2 \mathbf{\Phi}(\mathbf{a}) + (\frac{\lambda}{\mu} \left\| \mathbf{V}_{\mathcal{I}^c}^* \mathbf{x_a} \right\|_2^2 + \lambda \left\| \mathbf{V}_\mathcal{I}^* \mathbf{x_a} \right\|_1) \mathbf{I} \right) \iota P_{\mathbf{a}^\perp} \tag{I.36}$$

HK: [TODO] We are going to need to investigate the following thing: (1). Within region R_0 there always exists a negative curvature on saddle points with SMALL enough lambda so that it doesn't engulf the bad point-truncated sinusoid (2). The region of R_0 is large enough so that it within R_0^c the optimal x is always ones sparse by carefully selecting λ, μ

I.II Case R_0

The condition for R_0 can be written as:

$$R_0 = \left\{ \mathbf{a} \in \mathbb{S}^{k-1} : \left\| (\mathbf{C}_\mathbf{a}^* \mathbf{C}_\mathbf{a} - \frac{\lambda}{\mu} \mathbf{I})^{-1} \mathbf{C}_\mathbf{a}^* \mathbf{y} \right\|_\infty < \mu \right\} \tag{I.37}$$

$$= \left\{ \mathbf{a} \in \mathbb{S}^{k-1} : \mathbf{y} \in \mathbf{C}_\mathbf{a}^{-1*} (\mathbf{C}_\mathbf{a}^* \mathbf{C}_\mathbf{a} - \frac{\lambda}{\mu} \mathbf{I}) B_\mu,\ B_\mu = [-\mu, \mu]^m \right\} \tag{I.38}$$

When the gradient is 0, the Hessian becomes:

II Probability of Hitting $\cup_{\text{supp}(\nu) \in \mathcal{I}_{\text{st}}} R_\nu$ with Random Initilization

In this section, we would restrict our attention to a simpler algorithm which employs Riemannian gradient descent with the penalty function being ℓ_1, while as initializing the kernel $\mathbf{a}^{(0)}$ randomly and hope it hits a good region with moderate probability by constituting a sophisticated estimate of smoothing parameter λ. To set up the playground we first introduce basic property of objective function with ℓ_1 penalty term:

$$(P_{\ell_1}) \quad \mathbf{\Phi}(\mathbf{a}) = \min_\mathbf{x} \mathbf{\Psi}(\mathbf{x}, \mathbf{a}) \tag{II.1}$$

$$\mathbf{\Psi}(\mathbf{x}, \mathbf{a}) = \frac{1}{2} \left\| \mathbf{a} \otimes \mathbf{x} - \mathbf{y} \right\|_2^2 + \lambda \left\| \mathbf{x} \right\|_1 - \frac{1}{2} \left\| \mathbf{y} \right\|_2^2 \tag{II.2}$$

where $\mathbf{y} = \tilde{\mathbf{a}}_0 \otimes \mathbf{x}_0$, with $\mathbf{a} \sim \text{Unif}\left(\mathbb{S}^{k-1}\right)$ and $\mathbf{x}_0 \sim \text{BG}(\theta)$. The optimality condition of $\mathbf{x}$ with $\text{supp}\left(\mathbf{x}\right) = I$ given any $\mathbf{a}$ on sphere is:

$$\mathbf{C}_\mathbf{a}^*(\tilde{\mathbf{a}} \otimes \mathbf{x}_\mathbf{a} - \mathbf{y}) + \lambda \partial \left\|\cdot\right\|_1 (\mathbf{x}_\mathbf{a}) = \mathbf{0} \tag{II.3}$$

$$\iff \begin{cases} \mathbf{V}_I^* \mathbf{C}_\mathbf{a}^* \mathbf{C}_\mathbf{a} \mathbf{V}_I \mathbf{x}_I = \mathbf{V}_I^* \mathbf{C}_\mathbf{a}^* \mathbf{y} - \lambda \mathbf{V}_I^* \text{sgn}\left(\mathbf{x}\right) \\ \left\|\mathbf{V}_{I^c}^* \mathbf{C}_\mathbf{a}^* \mathbf{C}_\mathbf{a} \mathbf{V}_I \mathbf{x}_I - \mathbf{V}_{I^c}^* \mathbf{C}_\mathbf{a}^* \mathbf{y}\right\|_\infty \le \lambda \end{cases} \tag{II.4}$$

$$\iff \begin{cases} \mathbf{e}_i^* \mathbf{C}_\mathbf{a}^* \mathbf{C}_\mathbf{a} \mathbf{x}_\mathbf{a} = \mathbf{e}_i^* \mathbf{C}_\mathbf{a}^* \mathbf{y} - \lambda \text{sgn}\left(x_i\right) & i \in I \\ |\mathbf{e}_i^* \mathbf{C}_\mathbf{a}^* \mathbf{C}_\mathbf{a} \mathbf{x}_\mathbf{a} - \mathbf{e}_i^* \mathbf{C}_\mathbf{a}^* \mathbf{y}| \le \lambda & i \notin I \end{cases} \tag{II.5}$$

$$\iff \begin{cases} \left\langle s_i[\tilde{\mathbf{a}}], \sum_{j \in I} x_j s_j[\tilde{\mathbf{a}}] - \mathbf{y} \right\rangle = -\lambda \text{sgn}\left(x_i\right) & i \in I \\ \left| \left\langle s_i[\tilde{\mathbf{a}}], \sum_{j \in I} x_j s_j[\tilde{\mathbf{a}}] - \mathbf{y} \right\rangle \right| \le \lambda & i \notin I \end{cases} \tag{II.6}$$

also we have a good knowledge for the Riemannian gradient:

$$\nabla \mathbf{\Phi}(\mathbf{a}) = \iota^*(\check{\mathbf{C}}_\mathbf{a} \check{\mathbf{C}}_{\mathbf{x}_\mathbf{a}} - \check{\mathbf{C}}_\mathbf{y}) \mathbf{x}_\mathbf{a} \tag{II.7}$$

Now, assume a good support set I_{st} has the following property:

$$\mathcal{I}_{st} = \bigcup_{\tau \in [m]} \left\{ I \subset [m] : I \subset \text{supp}\left(s_{-\tau}[\mathbf{x}_0]\right), I \neq \emptyset \right\} \tag{II.8}$$

which is the support set of shifted thresholded optimal $\mathbf{x}_0$. The reason we consider (II.8) to be a good support set, is that we think Riemannian gradient has good property within this region.

Case 1: $\text{supp}\left(\mathbf{x}_\mathbf{a}\right) = \text{supp}\left(\mathbf{x}_0\right)$ The simplest case is when we have exact support recovery on the sparse vector. Let's name $I_0 = \text{supp}\left(x_0\right)$, and in this case we can see the Riemannien gradient becomes:

$$P_{\mathbf{a}^\perp} \iota^* \nabla \mathbf{\Phi}(\mathbf{a}) = P_{\mathbf{a}^\perp} \iota^* \sum_{j \in I_0} x_j s_{-j}[\sum_{i \in I_0} x_i s_i[\tilde{\mathbf{a}}] - x_{0i} s_i[\tilde{\mathbf{a}}_0]] \tag{II.9}$$

$$= P_{\mathbf{a}^\perp} \iota^* \left(\sum_{i \neq j \in I_0} x_i x_j s_{i-j}[\tilde{\mathbf{a}}] - x_{0i} x_j s_{i-j}[\tilde{\mathbf{a}}_0] + \sum_{i \in I_0} x_i^2 \tilde{\mathbf{a}} - x_{0i} x_i \tilde{\mathbf{a}}_0 \right) \tag{II.10}$$

$$= \sum_{i \neq j \in I_0,\, |i-j| < k} P_{\mathbf{a}^\perp} \iota^* s_{i-j}[x_i x_j \tilde{\mathbf{a}} - x_{0i} x_j \tilde{\mathbf{a}}_0] - \langle \mathbf{x}_0, \mathbf{x} \rangle P_{\mathbf{a}^\perp} \tilde{\mathbf{a}}_0 \tag{II.11}$$

HK: We need $\langle \mathbf{x}_0, \mathbf{x} \rangle$ to be positive and dominant therefore guarantee we are converging to the desired optimal solution. Here $\mathbf{x}_0$ is random and $\mathbf{x}$ is not. We need to characterize the condition for $\mathbf{x}$

Case 2: $\text{supp}\left(\mathbf{x}_\mathbf{a}\right) = \text{supp}\left(s_{-\tau}[\mathbf{x}_0]\right)$ Let $s_{-\tau}[I_0] = I_{-\tau}$, the Riemannian gradient in this region becomes:

$$P_{\mathbf{a}^\perp} \iota^* \nabla \mathbf{\Phi}(\mathbf{a}) = P_{\mathbf{a}^\perp} \iota^* \sum_{k \in I_{-\tau}} x_k s_{-k}[\sum_{i \in I_{-\tau}} x_i s_i[\tilde{\mathbf{a}}] - \sum_{j \in I_0} x_{0j} s_j[\tilde{\mathbf{a}}_0]] \tag{II.12}$$

$$= P_{\mathbf{a}^\perp} \iota^* \left(\sum_{i,j \notin I_0} x_i x_j s_{i-j}[\tilde{\mathbf{a}}] - x_{0i} x_j s_{i-j+\tau}[\tilde{\mathbf{a}}_0] + \sum_{i \in I_0} x_i^2 \tilde{\mathbf{a}} - x_{0i} x_i s_\tau[\tilde{\mathbf{a}}_0] \right) \qquad (\mathbf{x} \leftarrow s_\tau[\mathbf{x}])$$

$$= \sum_{i \neq j \in I_0, |i-j| < k} x_i x_j P_{\mathbf{a}^\perp} \iota^* s_{i-j}[\tilde{\mathbf{a}}] - \sum_{i \neq j \in I_0, |i-j+\tau| < k} x_{0i} x_j P_{\mathbf{a}^\perp} \iota^* s_{i-j+\tau}[\tilde{\mathbf{a}}_0] - \langle \mathbf{x}_0, \mathbf{x} \rangle P_{\mathbf{a}^\perp} \iota^* s_\tau[\tilde{\mathbf{a}}_0] \tag{II.13}$$

HK: These regions can be quite problematic. Consider then case when $\tau = k - 1$, the desired descent direction $\iota^* s_\tau[\tilde{\mathbf{a}}_0]$ can has a much smaller norm then the other direction such as $\iota^* s_{i-j+\tau}[\tilde{\mathbf{a}}]$.

We assume the following property of random $\mathbf{a}$ and $\mathbf{a}_0$

Conjecture II.1. [Large Correlation with one of Shift Truncation] *Given* $\mathbf{a} \in \mathbb{R}^k \sim N(0,1)$ *and* $\mathbf{a}_0 \in \mathbb{R}^k \sim N(0,1)$. *There exists one* $\tau^* \in [\pm k]$ *such that* $\langle \tilde{\mathbf{a}}, s_{\tau^*}[\tilde{\mathbf{a}}_0] \rangle > \lambda_1$ *while as* $\forall\, \tau \in [\pm k] \backslash \{\tau^*\}$, $|\langle \tilde{\mathbf{a}}, s_\tau[\tilde{\mathbf{a}}_0] \rangle| \leq \lambda_2 < \lambda_1$

Suppose we wisely select a proper λ and enlarge the region of R_ν. With numbers of random initialization there would be some guesses landing on this good region...

II.I Lower Bound of λ

When λ gets smaller, the function landscape becomes more irregular and non-smoother...

II.II Upper Bound of λ

When λ gets larger, the region R_0 occupies more space on sphere, when our objective is with ℓ_1 penalty function, the function is flat and in R_0 and is local maximum of the function, i.e. the gradient is zero. The Riemannian gradient descent method would fail if our initialized $\mathbf{a}^{(0)}$ lands on this region. This provides us the upper bound of smooth parameter λ. We can see that...

III Trivia

III.I Hypercube of in complex domain

For some reason, I'm interested in finding a simple expression of the following set:

$$\hat{B}_\infty^m := \{\mathbf{F}\mathbf{q} : q_i \in [-1,1]\} \tag{III.1}$$

The better choice to describe the set is with hyperplane description. Notice that a n-dimensionoal hypercube has 2^n vertices and $2n$ $n-1$ facet. Since Fourier transform matrix is unitary thus transforming a hypercube in original space remains being a hypercube in transformed space. Intuitively, one can easily see the following relation

$$\begin{cases} \langle \mathbf{F}\mathbf{x}, \mathbf{F}\mathbf{y} \rangle &= \langle \mathbf{x}, \mathbf{F}^*\mathbf{F}\mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle \\ \|\mathbf{F}\mathbf{x}\| &= \|\mathbf{x}\| \end{cases} \tag{III.2}$$

holds when we define the matrix $\mathbf{F}$ to be unitary Fourier matrix:

$$\mathbf{F}_{n,k} = \frac{1}{\sqrt{m}} e^{\frac{i2\pi n}{m} k} \tag{III.3}$$

The hypercube set in original space is

$$B_\infty^m = \{\mathbf{q} : q_i \in [-1,1]\} \tag{III.4}$$

$$= \{\mathbf{q} : \forall\, i \in [m],\ \langle \mathbf{q}, \mathbf{e}_i \rangle \leq 1,\ \langle \mathbf{q}, \mathbf{e}_i \rangle \geq -1\} \tag{III.5}$$

where $\pm\mathbf{e}_i$ are normal vectors of each facet. Now the transformed hypercube can described similarly with intersection of hyperplains:

$$\hat{B}_\infty^m = \{\mathbf{z} : \forall\, i \in [m],\ \langle \mathbf{z}, \mathbf{F}\mathbf{e}_i \rangle \leq 1,\ \langle \mathbf{z}, \mathbf{F}\mathbf{e}_i \rangle \geq -1\} \tag{III.6}$$

We try to link this description with the set R_0, consider the following set

$$R_0 = \left\{ \mathbf{a} \in \mathbb{S}^{k-1} : \exists\, \mathbf{q} \in B_\infty^m \quad s.t. \quad \iota\check{\mathbf{a}} \otimes \tilde{\mathbf{a}}_0 \otimes \mathbf{x}_0 = \lambda \mathbf{q} \right\} \tag{III.7}$$

$$= \left\{ \mathbf{a} \in \mathbb{S}^{k-1} : |\langle \iota\check{\mathbf{a}}, \mathbf{e}_i \otimes \mathbf{a}_0 \otimes \mathbf{x}_0 \rangle| \leq \lambda \quad \forall\, i \in [m] \right\} \tag{III.8}$$

IV Sufficient and Necessary Condition for Stationarity

Our first approach was to segregate the function landscape in terms of the support pattern of correspondent optimal sparse solution $\mathbf{x}$, however after observes several simple cases one may discover such segmentation doesn't match the property of function gradient and Hession well—in the continuous section with distinctive large gradient, one may find the support pattern changing several times, and same phenomenon can be observed with region of large concave direction. Thus this urge us to explore some more compact description of stationarity.

Analytical expression for stationarity Since we consider the algorithm is optimized over a sphere, the natural gradient condition should be altered to Riemannian gradient, which is the proper linear approximation under such constraint, thus:

$$\begin{cases} \operatorname{grad}[\mathbf{\Phi}](\mathbf{a}) = P_{\mathbf{a}^\perp}\iota^*(\check{\mathbf{C}}_{\mathbf{a}\otimes\mathbf{x}} - \check{\mathbf{C}}_{\mathbf{y}})\mathbf{x} = \mathbf{0} \\ \partial_{\mathbf{x}}\mathbf{\Psi}(\mathbf{x},\mathbf{a}) = \mathbf{C}_{\mathbf{a}}^*(\mathbf{C}_{\mathbf{a}}\mathbf{x} - \mathbf{y}) + \lambda\partial r_\mu(\mathbf{x}) = \mathbf{0} \end{cases} \tag{IV.1}$$

For simplicity, we may first assume the sparse regularizing function r_μ to be $\|\cdot\|_1$, then the simplified form can be found (II.6).

Sufficient conditions, ℓ_1-norm The first apparent sufficient condition is when λ is large then all kernels in R_0 are local maximas hence are stationary points, the equivalent expression for $\mathbf{a} \in R_0$ follows as:

$$\mathbf{a} \in R_0 \iff \|\mathbf{C}_{\mathbf{a}}^*\mathbf{y}\|_\infty \le \lambda \iff \left\|\check{\mathbf{C}}_{\mathbf{y}\iota}\mathbf{a}\right\|_\infty \le \lambda \iff \tag{IV.2}$$

HK: todo, the goal is to find for each stationarity points, which is local-minima, local-maxima or saddle points.

V Planted Dictionary Sparse Vector in Random Subspace

In this section we would focus on an simple problem similar to the planted sparse vector problem but with small variation: suppose we have an overcomplete dictionary with k-RIP property with parameter $delta_k$ $\mathbf{D} \in \mathbb{R}^{m\times n}$ where $m < n$, then planted sparse vector subspace are defined as the range of the matrix $\mathbf{U} \in \mathbb{R}^{m\times L}$ defined as

$$\mathbf{U} = [\frac{\mathbf{D}\mathbf{x}_0}{\|\mathbf{D}\mathbf{x}_0\|_2} \mid \mathbf{u}_1 \mid \mathbf{u}_2 \mid \cdots \mid \mathbf{u}_{L-1}] = [\frac{\mathbf{D}\mathbf{x}_0}{\|\mathbf{D}\mathbf{x}_0\|_2} \mid \bar{\mathbf{U}}] \tag{V.1}$$

where $\bar{\mathbf{U}}$ is constructed randomly as long as being column orthonormal and diagonal to $\mathbf{D}\mathbf{x}_0$ thus $\mathbf{U}$ itself is also column orthonormal. The matrix $\bar{\mathbf{U}}$ can be described as orthonomalization of iid random Gaussian vectors $\{\mathbf{g}_i\}_{i=1}^{L-1} \subset \mathbb{R}^m$ with each each entries are also iid Gaussian $g_{ij} \sim \mathcal{N}(0, 1/m)$, that projected away from direction $\mathbf{D}\mathbf{x}_0$, namely:

$$\bar{\mathbf{U}} = \operatorname{orth}\left[P_{\mathbf{D}\mathbf{x}_0^\perp}\{\mathbf{g}_1, \ldots, \mathbf{g}_{L-1}\}\right] = \operatorname{orth}[P_{\mathbf{D}\mathbf{x}_0^\perp}\mathbf{G}] \tag{V.2}$$

where $\mathbf{G} = [\mathbf{g}_1, \ldots, \mathbf{g}_{L-1}]$, and notice that the projection onto range of $\bar{\mathbf{U}}$ can be realized as:

$$P_{\bar{\mathbf{U}}} = \bar{\mathbf{U}}\bar{\mathbf{U}}^* = P_{\mathbf{D}\mathbf{x}_0^\perp}\mathbf{G}[\mathbf{G}^*P_{\mathbf{D}\mathbf{x}_0^\perp}\mathbf{G}]^{-1}\mathbf{G}^*P_{\mathbf{D}\mathbf{x}_0^\perp} \tag{V.3}$$

Then to find such structured sparse vector in subspace, one naturally can formulate the following optimization problem:

$$\min_{\mathbf{q},\mathbf{x}} \frac{1}{2}\|\mathbf{U}\mathbf{q} - \mathbf{D}\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \quad s.t. \quad \mathbf{q} \in \mathbb{S}^{L-1} \tag{V.4}$$

with desired optimal solution $\mathbf{q}^* = \mathbf{e}_1$

Rounding Problem Given the vector $\hat{\mathbf{q}} \in \mathbb{S}^{L-1}$, and supposedly expecting such vector is closed enough to the optimal solution $\mathbf{e}_1$, another sensible approach to solve this problem would be to formulate the problem on domain of tangent space $T_{\hat{\mathbf{q}}}\mathbb{S}^{L-1}$ as follows:

$$\min_{\mathbf{q},\mathbf{x}} \|\mathbf{x}\|_1 \quad s.t. \quad \mathbf{U}\mathbf{q} = \mathbf{D}\mathbf{x}, \quad \langle \hat{\mathbf{q}}, \mathbf{q} \rangle = 1 \tag{V.5}$$

and in this problem, we aim to quantify the relationship between the sparsity of optimal sparse vector $\mathbf{x}_0$ and distance between $\hat{\mathbf{q}}$ and $\mathbf{q}$. Thus in this paragraph we would try to identify the optimal solution of problem (V.5). Notice that the expected optimal solution pair $(\mathbf{q}^*, \mathbf{x}^*)$ is going to be $(\frac{\mathbf{e}_1}{\hat{q}_1}, \frac{\mathbf{x}_0}{\hat{q}_1 \|\mathbf{D}\mathbf{x}_0\|_2})$

We first define a different structured sparsity inducing norm with respect to the column vector of overcomplete matrix $\mathbf{D}$ as $\|\mathbf{y}\|_{\mathbf{D}} := \min_{\mathbf{y}=\mathbf{D}\mathbf{x}} \|\mathbf{x}\|_1$, assuming the matrix $\mathbf{D}$ is full row rank then we can show $\|\cdot\|_{\mathbf{D}}$ is a norm:

$$\|\mathbf{y}\|_{\mathbf{D}} = 0 \iff \mathbf{y} = \mathbf{D}\,\mathbf{0} \iff \mathbf{y} = \mathbf{0} \tag{V.6}$$

$$\|\alpha\,\mathbf{y}\|_{\mathbf{D}} = \min_{\alpha\,\mathbf{y}=\mathbf{D}\mathbf{x}} \|\mathbf{x}\|_1 = \min_{\mathbf{y}=\mathbf{D}\frac{\mathbf{x}}{\alpha}} \alpha \left\|\frac{\mathbf{x}}{\alpha}\right\|_1 = \alpha\,\|\mathbf{y}\|_{\mathbf{D}} \tag{V.7}$$

$$\|\mathbf{y}_1 + \mathbf{y}_2\|_{\mathbf{D}} = \min_{\mathbf{y}_1=\mathbf{D}\mathbf{x}_1, \mathbf{y}_2=\mathbf{D}\mathbf{x}_2} \|\mathbf{x}_1 + \mathbf{x}_2\|_1 \le \min_{\mathbf{y}_1=\mathbf{D}\mathbf{x}_1, \mathbf{y}_2=\mathbf{D}\mathbf{x}_2} \|\mathbf{x}_1\|_1 + \|\mathbf{x}_2\|_1 = \|\mathbf{y}_1\|_{\mathbf{D}} + \|\mathbf{y}_2\|_{\mathbf{D}} \tag{V.8}$$

$$\begin{aligned}
\|\mathbf{y}_1 + \mathbf{y}_2\|_{\mathbf{D}} &= \min_{\mathbf{y}_1=\mathbf{D}\mathbf{x}_1, \mathbf{y}_2=\mathbf{D}\mathbf{x}_2} \left\|\mathbf{x}_1 + \mathbf{x}_2|_{\mathrm{supp}(\mathbf{x}_1)}\right\|_1 + \left\|\mathbf{x}_2|_{\mathrm{supp}(\mathbf{x}_1)^c}\right\|_1 \\
&= \min_{\mathbf{y}_2=\mathbf{D}\mathbf{x}_2} \left\|\mathbf{x}_1^* + \mathbf{x}_2|_{\mathrm{supp}(\mathbf{x}_1^*)}\right\|_1 + \left\|\mathbf{x}_2|_{\mathrm{supp}(\mathbf{x}_1^*)^c}\right\|_1, \quad \mathbf{x}_1^* = \operatorname*{argmin}_{\mathbf{y}_1=\mathbf{D}\mathbf{x}_1} \|\mathbf{x}_1\|_1 \\
&\ge \min_{\mathbf{y}_1=\mathbf{D}\mathbf{x}_1} \|\mathbf{x}_1\|_1 + \min_{\mathbf{y}_2=\mathbf{D}\mathbf{x}_2} (\|\mathbf{x}_2\|_1 - 2\left\|\mathbf{x}_2|_{\mathrm{supp}(\mathbf{x}_1^*)}\right\|_1) \\
&\ge \|\mathbf{y}_1\|_{\mathbf{D}} + \|\mathbf{y}_2\|_{\mathbf{D}} - 2 \max_{\mathbf{y}_2=\mathbf{D}\mathbf{x}_3} \left\|\mathbf{x}_3|_{\mathrm{supp}(\mathbf{x}_1^*)}\right\|_1
\end{aligned} \tag{V.9}$$

Now equivalently we may rewrite the problem (V.5) as

$$\min_{\mathbf{q}} \|\mathbf{U}\mathbf{q}\|_{\mathbf{D}} \quad s.t. \quad \langle \hat{\mathbf{q}}, \mathbf{q} \rangle = 1 \tag{V.10}$$

Let $I = \mathrm{supp}\left(\mathrm{argmin}_{\mathbf{D}\mathbf{x}_0=\mathbf{D}\mathbf{x}} \|\mathbf{x}\|_1\right)$ then

$$\|\mathbf{U}\mathbf{q}\|_{\mathbf{D}} = \left\|q_1 \frac{\mathbf{D}\mathbf{x}_0}{\|\mathbf{D}\mathbf{x}_0\|} + \bar{\mathbf{U}}\bar{\mathbf{q}}\right\|_{\mathbf{D}} \ge \frac{q_1}{\|\mathbf{D}\mathbf{x}_0\|_2} \|\mathbf{D}\mathbf{x}_0\|_{\mathbf{D}} + \min_{\bar{\mathbf{U}}\bar{\mathbf{q}}=\mathbf{D}\mathbf{x}} \|\mathbf{x}\|_1 - 2\,\|\mathbf{x}|_I\|_1 \tag{V.11}$$

instead of considering the equality constraint $\bar{\mathbf{U}}\bar{\mathbf{q}} = \mathbf{D}\mathbf{x}$, we are interested a milder constraint to obtain the lower bound. Consider

$$\min_{\bar{\mathbf{U}}\bar{\mathbf{q}}=\mathbf{D}\mathbf{x}} \|\mathbf{x}\|_1 - 2\,\|\mathbf{x}|_I\|_1 \ge \min_{\mathbf{x}\in\mathrm{null}(P_{\bar{\mathbf{U}}^\perp}\mathbf{D})} \|\mathbf{x}\|_1 - 2\,\|\mathbf{x}_\mathbf{I}\|_1 \tag{V.12}$$

here we are adopting the proof by Candes and Tao **HK: cite**. First we need to investigate the RIP of operator $P_{\bar{\mathbf{U}}^\perp}\mathbf{D}$ as follows:

Lemma V.1. [RIP of random projection of RIP matrix] *Given matrix $\mathbf{D} \in \mathbb{R}^{m\times n}$ such that $\mathbf{D}$ is a k-RIP matrix with parameter δ_k and a random subspace $\bar{\mathbf{U}}$ of subspace dimension $L-1$ generated as* (V.2)*, then with $\delta =$* **TODO: ,** *for all $\mathbf{x} \in \mathbb{R}^n$ such that $|\mathrm{supp}\,(\mathbf{x})| \le k$, we have*

$$(1-\delta)\,\|\mathbf{x}\|_2^2 \le \|P_{\bar{\mathbf{U}}^\perp}\mathbf{D}\mathbf{x}\|_2^2 \le (1+\delta)\,\|\mathbf{x}\|_2^2 \tag{V.13}$$

with probability at least

Proof. For convenience we let $\mathbf{a} = \mathbf{D}\mathbf{x}_0$ within the scope of this proof. First we define the set of vectors in range of $\mathbf{D}$ with sparse inputs:

$$\mathcal{D} := \{\mathbf{D}\mathbf{x} \in \mathbb{R}^m : \mathbf{x} \in \mathbb{R}^n,\ |\mathrm{supp}\,(\mathbf{x})| \le k\} \tag{V.14}$$

then for all vectors in set $\mathcal{D}$ we want to find the maximum difference of the norm between such vector $\mathbf{z} \in \mathcal{D}$ and its projection to algebraic complement subspace of $\bar{\mathbf{U}}$. This can be written as:

$$\max_{\mathbf{z}\in\mathcal{D}} \left| \|\mathbf{z}\|_2^2 - \|P_{\bar{\mathbf{U}}^\perp}\mathbf{z}\|_2^2 \right| = \max_{\mathbf{z}\in\mathcal{D}} \left\|\bar{\mathbf{U}}^*\mathbf{z}\right\|_2^2 = \max_{\mathbf{z}\in\mathcal{D}} \left\|(\mathbf{G}^* P_{\mathbf{a}^\perp}\mathbf{G})^{-1/2}\mathbf{G}^* P_{\mathbf{a}^\perp}\mathbf{z}\right\|_2^2 \tag{V.15}$$

$$\leq \max_{\mathbf{z}\in\mathcal{D}} \left\|(\mathbf{G}^* P_{\mathbf{a}^\perp}\mathbf{G})^{-1/2}\right\|_2^2 \|\mathbf{G}^* P_{\mathbf{a}^\perp}\mathbf{z}\|_2^2 \tag{V.16}$$

$$\leq \frac{1}{\sigma_{\min}(P_{\mathbf{a}^\perp}\mathbf{G})} \max_{\mathbf{z}\in\mathcal{D}} \|\mathbf{G}^* P_{\mathbf{a}^\perp}\mathbf{z}\|_2^2 \tag{V.17}$$

Observe that $\mathbb{E}\,\|\mathbf{G}^* P_{\mathbf{a}^\perp}\mathbf{z}\|_2^2 = \mathbf{z}^* P_{\mathbf{a}^\perp}\mathbb{E}\left[\mathbf{G}\mathbf{G}^*\right] P_{\mathbf{a}^\perp}\mathbf{z} = \frac{L-1}{m}\|P_{\mathbf{a}^\perp}\mathbf{z}\|_2^2$. Furthermore for any given $\mathbf{z} \in \mathcal{D}$, $\|\mathbf{G}^* P_{\mathbf{a}^\perp}\mathbf{z}\|_2$ is a Lipchitz function over Gaussian matrix $\mathbf{G}$:

$$\left| \|\mathbf{G}_1^* P_{\mathbf{a}^\perp}\mathbf{z}\|_2 - \|\mathbf{G}_2^* P_{\mathbf{a}^\perp}\mathbf{z}\|_2 \right| = \tag{V.18}$$

□

Lemma V.2. [Lower Bound on 1-norm with Linear Constraint] *Given a vector $\mathbf{x} \in \mathbb{R}^n$ and linear operator $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^m$ with $2k$-RIP with parameter δ_{2k}, and suppose $\mathbf{x} \in \mathrm{null}(\mathbf{A})$, furthermore we are given a small index set $I \subset [n]$ such that $|I| = k$, then the following inequality holds*

$$\|\mathbf{x}\|_1 - 2\|\mathbf{x}_I\|_1 \geq (1-2c)\|\mathbf{x}\|_2 \tag{V.19}$$

*where $c =$***HK: todo**.

Proof. The proof would composed of two sections, first we shall study the upper bound of $\|\mathbf{x}_I\|_1$ with respect to its full ℓ_1 norm $\|\mathbf{x}\|_1$, by showing the vector $\mathbf{x}$ would not concentrated in certain small support sets. Then similar idea applies in the second part and we would be showing $\|\mathbf{x}\|_1$ much larger then $\|\mathbf{x}\|_2$ by arguing the vectors in null space is fully dense.

Firstly we define the following partition of support set with respect to a given fixed vector $\mathbf{x} \in \mathrm{null}(\mathbf{A})$:

$$I^c = J_1 \uplus J_2 \uplus \cdots \uplus J_p \tag{V.20}$$

$$\forall i \in [p-1], \quad |J_i| = k, \quad |J_p| = n - pk \tag{V.21}$$

where the sets $\{J_i\}_{i=1}^p$ are chosen in regard to absolute values of $\mathbf{x}$ in descending order, namely

$$\forall\, i \in \{2,\ldots,p\}, \quad \max_{k\in J_i} |x_k| \leq \min_{k\in J_{i-1}} |x_k| \tag{V.22}$$

then we have the following property on norm of $\mathbf{x}_{J_i}$:

$$\|\mathbf{x}_{J_i}\|_2 \leq \sqrt{k}\,\|\mathbf{x}_{J_i}\|_\infty \leq \frac{\sqrt{k}}{k}\left\|\mathbf{x}_{J_{i-1}}\right\|_1 = \frac{\left\|\mathbf{x}_{J_{i-1}}\right\|_1}{\sqrt{k}} \tag{V.23}$$

and notice that

$$\mathbf{x} \in \mathrm{null}(\mathbf{A}) \iff \mathbf{A}(\mathbf{x}_I + \mathbf{x}_{J_1} + \cdots \mathbf{x}_{J_p}) = 0 \iff \mathbf{A}(\mathbf{x}_I + \mathbf{x}_{J_1}) = -\mathbf{A}(\mathbf{x}_{J_2} + \cdots + \mathbf{x}_{J_p}) \tag{V.24}$$

and also for vectors $\mathbf{u}, \mathbf{v}$ with both vector k-sparse but disjoint support, we have the lower bound on $\langle \mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{v}\rangle$:**HK: V.21 wrong**

$$2\langle \mathbf{A}\mathbf{e}_i, \mathbf{A}\mathbf{e}_j\rangle \leq (1+\delta_{2k})\|\mathbf{e}_i + \mathbf{e}_j\|_2^2 - \|\mathbf{A}\mathbf{e}_i\|_2^2 - \|\mathbf{A}\mathbf{e}_j\|_2^2 \tag{V.25}$$

$$\leq (1+\delta_{2k})(\|\mathbf{e}_i\|_2^2 + \|\mathbf{e}_j\|_2^2) - (1-\delta_{2k})(\|\mathbf{e}_i\|_2^2 + \|\mathbf{e}_j\|_2^2) \tag{V.26}$$

$$\leq 4\delta_{2k} \tag{V.27}$$

$$\langle \mathbf{A}\mathbf{u}, \mathbf{A}\mathbf{v}\rangle = \left\langle \sum_{i\in I} u_i \mathbf{a}_i, \sum_{j\in J} v_j \mathbf{a}_j \right\rangle \geq -\left| \sum_{i,j} u_i v_j \langle \mathbf{a}_i, \mathbf{a}_j\rangle \right| \geq -2\delta_{2k}\|\mathbf{u}\|_2\|\mathbf{v}\|_2 \tag{V.28}$$

Now we are ready to obtain the upper bound of $\|\mathbf{x}_I\|_1$:

$$\|\mathbf{x}_I\|_1^2 \le k\|\mathbf{x}_I\|_2^2 \le k\|\mathbf{x}_I+\mathbf{x}_{J_1}\|_2^2 \le k(1+\delta_{2k})\|\mathbf{A}(\mathbf{x}_I+\mathbf{x}_{J_1})\|_2^2 \tag{V.29}$$

$$\le k(1+\delta_{2k})\left\langle \mathbf{A}(\mathbf{x}_I+\mathbf{x}_{J_1}), -\sum_{i=2}^{p}\mathbf{A}\mathbf{x}_{J_i}\right\rangle \tag{V.30}$$

$$\le 2k\delta_{2k}(1+\delta_{2k})(\|\mathbf{x}_I\|_2+\|\mathbf{x}_{J_1}\|_2)(\sum_{i=2}^{p}\|\mathbf{x}_{J_i}\|_2) \tag{V.31}$$

$$\le 2\sqrt{k}\delta_{2k}(1+\delta_{2k})(\sqrt{2}\|\mathbf{x}_I+\mathbf{x}_{J_1}\|_2)\|\mathbf{x}\|_1 \tag{V.32}$$

$$\le 8\sqrt{k}\delta_{2k}^2\|\mathbf{x}\|_1^2 \tag{V.33}$$

Now through similar argument we can get the lower bound of $\|\mathbf{x}\|_1$, let $\mathcal{I}$ be a set of disjoint support subsets of $[n]$ with each elements $|I| \le k$ and $\uplus_{I\in\mathcal{I}} = [n]$, then

$$\|\mathbf{x}\|_2^2 = \sum_{I\in\mathcal{I}}\|\mathbf{x}_I\|_2^2 \le 4\frac{n+k}{k}\delta_{2k}(1+\delta_{2k})\|\mathbf{x}\|_1^2 \le \frac{16n}{k}\delta_{2k}^2\|\mathbf{x}\|_1^2 \tag{V.34}$$

Thus we can get the desired property as follows:

$$\|\mathbf{x}\|_1 - 2\|\mathbf{x}_I\|_1 \ge (1-4\sqrt{2k}\delta_{2k})\|\mathbf{x}\|_1 \ge \frac{1-4\sqrt{2k}\delta_{2k}}{4\delta_{2k}}\sqrt{\frac{k}{n}}\|\mathbf{x}\|_2 \tag{V.35}$$

□

VI Finding Structured Sparse vector in Subspace

An intriguing aspect of the sparse blind deconvolution problem is that while provided more information, such as the autocorrelation of optimal kernel $\mathbf{a}_0$, we may obtain a new formulation and escape from the bilinear dilemma from original problem. Such assumption is not overly lack of substance even in application aspect. Suppose we encounter a signal $\mathbf{y}_0$ generated from a fixed short kernel $\mathbf{a}_0$ which convolute with a random sparse vector of extremely high dimension $\mathbf{x}_0$, consider the expectation of circular autocorrelation of observation $\mathbb{E}[\check{\mathbf{y}}_0 \otimes \mathbf{y}_0]$, we may get

$$\mathbb{E}[\check{\mathbf{y}}_0 \otimes \mathbf{y}_0] = \mathbb{E}[\mathbf{C}_{\mathbf{a}_0}^*\mathbf{C}_{\mathbf{x}_0}^*\mathbf{C}_{\mathbf{x}_0}\mathbf{a}_0] = (\mathbb{E}\|\mathbf{x}_0\|^2)\check{\mathbf{a}}_0 \otimes \mathbf{a}_0 = m\cdot\theta\cdot\check{\mathbf{a}}_0\otimes\mathbf{a}_0 \tag{VI.1}$$

Thus when the dimension m grows larger, we may expect the autocorrelation of the observation itself converges to the autocorrelation of the optimal kernel $\mathbf{a}_0$.

Here, let's put on a more radical assumption that we have exact information of the autocorrelation $r_{\mathbf{a}_0}$ of optimal kernel $\mathbf{a}_0$, an interesting observation from the Fourier domain of the given data $\mathbf{y}$ and $r_{\mathbf{a}_0}$ can be shown by the following derivation:

$$\mathbf{y}_0 = \mathbf{C}_{\mathbf{a}_0}\mathbf{x}_0 = \sqrt{m}\,\mathbf{F}^*((\mathbf{F}\mathbf{a}_0)\circ(\mathbf{F}\mathbf{x}_0)) \tag{VI.2}$$

$$\Longrightarrow \mathbf{F}\mathbf{x}_0 = \frac{1}{\sqrt{m}}\frac{\mathbf{F}\mathbf{y}_0}{\mathbf{F}\mathbf{a}_0} = \frac{1}{\sqrt{m}}\frac{\mathbf{F}\mathbf{y}_0}{\sqrt{m}(\mathbf{F}^*\mathbf{a}_0)\circ(\mathbf{F}\mathbf{a}_0)}\circ(\mathbf{F}^*\mathbf{a}_0)\cdot\sqrt{m} \tag{VI.3}$$

$$\Longrightarrow \mathbf{F}\mathbf{x}_0 = (\frac{\mathbf{F}\mathbf{y}_0}{\mathbf{F}r_{\mathbf{a}_0}})\circ(\mathbf{F}^*\iota\mathbf{a}_0) = \mathbf{D}_{\mathbf{F}\mathbf{y}_0/\mathbf{F}r_{\mathbf{a}_0}}\mathbf{F}^*\iota\mathbf{a}_0 \tag{VI.4}$$

$$\Longrightarrow \mathbf{x}_0 = \mathbf{F}^*\mathbf{D}_{\mathbf{F}\mathbf{y}_0/\mathbf{F}r_{\mathbf{a}_0}}\mathbf{F}^*\iota\mathbf{a}_0 =: \mathbf{M}\mathbf{a}_0 \tag{VI.5}$$

thus we may view this problem as finding a sparse vector $\mathbf{x}_0$ in the structured subspace $\mathbf{M}$ problem. And naturally to solve such problem one may refer to our previous approach with alternating minimization a.k.a block coordinate descent method:

$$\min_{\mathbf{x}\in\mathbb{R}^m,\mathbf{a}\in\mathbb{S}^{k-1}}\frac{1}{2}\|\mathbf{M}\mathbf{a}-\mathbf{x}\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{VI.6}$$

From the property (VI.5), to show the formulation above make sense, then as we manage to obtain the optimal solution of problem (VI.6), the solution $\mathbf{x}^*$ indeed recovers the correct support of desired optimal sparse vector $\mathbf{x}^*$ while as the values on the support can be differed by λ. One step further, to show the uniqueness result for $\mathbf{x}^*$ as optimal solution, we should resort to the uniqueness result of ℓ_0 analogy:

Lemma VI.1. *Suppose* $\mathbf{M} \in \mathbb{R}^{m\times k}$ *is defined as* (VI.5)*, then the sparsest vector in* $R(\mathbf{M})\backslash\{0\}$ *has the form* $\alpha\,\mathbf{x}_0$ *w.h.p where* $\alpha \in \mathbb{R}\backslash\{0\}$.

Proof. Unlike our previous work on planted sparse vector in random subspace, here the sparse vector is embedded through convolutional operator with kernel vector. To destruct the sparse component in such subspace $R(M)$, we can firstly observe that:

$$\mathbf{M} = \mathbf{F}^*\mathbf{D}_{\mathbf{F}\mathbf{y}_0/\mathbf{F}r_{\mathbf{a}_0}}\mathbf{F}^*\iota = \mathbf{C}_{\mathbf{x}_0}\mathbf{F}^*\mathbf{D}_{\mathbf{F}\mathbf{a}_0/\mathbf{F}r_{\mathbf{a}_0}}\mathbf{F}^*\iota =: \mathbf{C}_{\mathbf{x}_0}\mathbf{A} \tag{VI.7}$$

furthermore

$$\mathbf{A}\mathbf{a}_0 = \mathbf{F}^*\mathbf{D}_{\mathbf{F}\mathbf{a}_0/(\mathbf{F}\mathbf{a}_0)\circ(\mathbf{F}^*\mathbf{a}_0)}\mathbf{F}^*\mathbf{a}_0 = \mathbf{e}_0 \tag{VI.8}$$

$$\Longrightarrow \mathbf{x}_0 \in R(\mathbf{M}) \tag{VI.9}$$

Thus it's natural to see $\alpha\,\mathbf{x}_0$ resides in range of $\mathbf{M}$. Now we would utilize the randomness setup for vector $\mathbf{x}_0$. Let $\mathbf{a} \neq \mathbf{0}$, and suppose $\|\mathbf{a}\|_0 = \ell \geq 3$, and $|\mathrm{supp}\,(\mathbf{x}_0)| = \kappa$, then

$$\|\mathbf{C}_{\mathbf{x}_0}\mathbf{a}\|_0 = \|\mathbf{C}_{\mathbf{a}}\mathbf{x}_0\|_0 = \|a_{i_1}s_{i_1}[\mathbf{x}_0] + \cdots + a_{i_\ell}s_{i_\ell}[\mathbf{x}_0]\|_0 \tag{VI.10}$$

The minimum union of support overlap across different circular shift of $\mathbf{x}_0$ can be calculated as

$$\min_{\tau:=\{i_1,\ldots,i_\ell\},\mathrm{supp}(\mathbf{x}_0)} \left|\bigcup_{j=1}^{\ell} s_{i_j}[\mathrm{supp}\,(\mathbf{x}_0)]\right| \leq \kappa + \ell - 1 \tag{VI.11}$$

where the minimum can be obtained by choosing support form to be $\mathrm{supp}\,(\mathbf{x}_0) = \{0,\ldots,\kappa-1\}$ and $\tau = \{0,\ldots,\ell-1\}$. Then

□

A more realistic scenario, considering $\mathbf{a}_0$ could be nearly non-invertible, or several frequency components are contaminated by noise, then such data model could be further polished when we presume certain entries of $\mathbf{F}_{r_{a_0}}$ to be unmistaken. Let's denote the set of such entries as Ω, then we may structure the formulation as follows:

$$\min_{\mathbf{x}\in\mathbb{R}^m,\mathbf{a}\in\mathbb{S}^{k-1}} \frac{1}{2}\left\|\widehat{\mathbf{M}}_{\mathbf{\Omega}}\mathbf{a} - \mathbf{F}_{\Omega}\mathbf{x}\right\|_2^2 + \lambda\|\mathbf{x}\|_1 \tag{VI.12}$$

where

$$\widehat{\mathbf{M}}_{\Omega} := \mathbf{V}_{\Omega}^*\mathbf{D}_{\mathbf{F}\mathbf{y}_0/\mathbf{F}r_{\mathbf{a}_0}}\mathbf{F}^*\iota \tag{VI.13}$$

$$\mathbf{F}_{\Omega} := \mathbf{V}_{\Omega}^*\mathbf{F} \tag{VI.14}$$

then again we have a Lasso problem as a subproblem when operating alternation minimization, and immediately we may see that the algorithm cannot be trivialized as previous formulation. Since solving Lasso with overcomplete dictionary can be nasty itself.

References