Restricted Isometry Property for κ-sparse Vectors in SLR

1. Recap from Manuscript

A random $\mathbf{\Phi}$ must often satisfy a small *Restricted Isometry Constant* (RIC) (Candes & Tao, 2005), which ensures that $\mathbf{\Phi}$ behaves nearly like an orthonormal system on every κ-sparse subset of coefficients (see below for definition). Concretely, no set of κ columns in $\mathbf{\Phi}$ is nearly linearly dependent, so no sparse vector $\mathbf{z}$ is "collapsed" or "inflated" by $\mathbf{\Phi}$.

Definition 1.1 (Restricted Isometry Constant)**.** Let $\mathbf{\Phi} \in \mathbb{R}^{d\times n}$ be a real matrix. For an integer $\kappa \leq k$, the κ-*Restricted Isometry Constant* δ of $\mathbf{\Phi}$ is the smallest non-negative number such that

$$(1-\delta)\,\|\mathbf{z}\|_2^2 \;\leq\; \|\mathbf{\Phi}\,\mathbf{z}\|_2^2 \;\leq\; (1+\delta)\,\|\mathbf{z}\|_2^2 \tag{1}$$

for *every* vector $\mathbf{z} \in \mathbb{R}^k$ that has at most κ nonzero entries. We abbreviate this condition as *Restricted Isometry Property* (RIP) for future use.

In order to prove Theorem 4.2 from the manuscript we minimize the simplified VAE loss (for a fixed γ) that depends only on $\mathbf{w}$ and is given by:

$$\mathcal{L}(\mathbf{w}) \;=\; \mathbf{x}^\top \mathbf{\Sigma}^{-1}(\mathbf{w})\,\mathbf{x} \;+\; \log\big|\mathbf{\Sigma}(\mathbf{w})\big|. \tag{2}$$

where $\mathbf{\Sigma}(\mathbf{w}) = \mathbf{\Phi}\,\mathrm{diag}[\mathbf{w}]\,\mathbf{\Phi}^\top + \gamma\,\mathbf{I}$. Computing the stationary points by finding $\dfrac{\partial \mathcal{L}}{\partial w_j}$ for each $w_j \in \mathbf{w}$, we obtain:

$$\frac{\partial \mathcal{L}(\mathbf{w})}{\partial w_j} \;=\; -\left(\boldsymbol{\phi}_j^\top \mathbf{\Sigma}^{-1}(\mathbf{w})\,\mathbf{x}\right)^2 \;+\; \boldsymbol{\phi}_j^\top \mathbf{\Sigma}^{-1}(\mathbf{w})\,\boldsymbol{\phi}_j \;=\; 0 \tag{3}$$

$$\implies \left(\boldsymbol{\phi}_j^\top \mathbf{\Sigma}^{-1}(\mathbf{w})\,\mathbf{x}\right)^2 \;=\; \boldsymbol{\phi}_j^\top \mathbf{\Sigma}^{-1}(\mathbf{w})\,\boldsymbol{\phi}_j. \tag{4}$$

This stationarity condition balances the "weighted prediction" for the jth coordinate against its corresponding diagonal element in $\mathbf{\Sigma}(\mathbf{w})^{-1}$. It is to be noted that (4) corresponds to (37), and (2) corresponds to (27) in the manuscript. Thereafter we use Lemma 1 in Appendix (A.2) to show by contradiction, that when $\mathbf{\Phi}$ satisfies the RIP condition, these coupled equations admit no spurious local minima: every minimum of the loss corresponds to a global minimum.

Lemma 1 in Appendix (A.2) stated as follows:

Lemma 1.2. *Suppose there exist two distinct vectors $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$, both satisfying the stationarity conditions over $\mathbf{x}$:*

$$\left(\boldsymbol{\phi}_j^\top \mathbf{\Sigma}(\mathbf{w}^{(1)})^{-1}\mathbf{x}\right)^2 \;=\; \boldsymbol{\phi}_j^\top \mathbf{\Sigma}(\mathbf{w}^{(1)})^{-1}\boldsymbol{\phi}_j, \tag{5}$$

$$\left(\boldsymbol{\phi}_j^\top \mathbf{\Sigma}(\mathbf{w}^{(2)})^{-1}\mathbf{x}\right)^2 \;=\; \boldsymbol{\phi}_j^\top \mathbf{\Sigma}(\mathbf{w}^{(2)})^{-1}\boldsymbol{\phi}_j, \quad \forall\, j. \tag{6}$$

Then $\mathbf{w}^{(1)} = \mathbf{w}^{(2)}$ if $\mathbf{\Phi}$ satisfies the RIP condition with small δ. In other words, there are no "bad" local minima under these stationarity conditions.

Note that (5) and (6) correspond to (38) in the manuscript.

2. Relating RIP δ to No Bad Local Minima

The main requirement for Lemma 1 is the absence of distinct $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ that satisfy (4). With a perfect optimizer, the full rank assumption ensures the presence of a unique inverse $\Sigma^{-1}(\mathbf{w})$ for unique $\mathbf{w}$, leading to no bad local minimas. However, practical optimizers such as SGD might identify distinct $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ as minima for the loss function in (2) that satisfy $\Sigma^{-1}(\mathbf{w}^{(1)}) \approx \Sigma^{-1}(\mathbf{w}^{(2)})$. We use the RIP bound δ to ensure that the difference in inverse terms large enough to be detectable by SGD.

Step 1: RIP expression for a κ-sparse vector

Let $\mathbf{z} = \sum_{i\in S} \alpha_i e_i$, where $S \subseteq \{1,\ldots,n\}$ with $|S| = \kappa$, be a κ-sparse vector and let $\mathbf{\Phi} \in \mathbb{R}^{d\times n}$ with columns $\boldsymbol{\phi}_1,\ldots,\boldsymbol{\phi}_n$. Then:

$$\mathbf{\Phi}\mathbf{z} = \sum_{i\in S} \alpha_i \boldsymbol{\phi}_i \tag{7}$$

The squared norm of $\mathbf{\Phi}\mathbf{z}$ becomes:

$$\begin{aligned}\|\mathbf{\Phi}\mathbf{z}\|_2^2 &= \left\|\sum_{i\in S} \alpha_i \boldsymbol{\phi}_i\right\|_2^2 \\ &= \sum_{i\in S} \alpha_i^2 \|\boldsymbol{\phi}_i\|_2^2 + \sum_{\substack{i,j\in S \\ i\neq j}} \alpha_i \alpha_j \langle \boldsymbol{\phi}_i, \boldsymbol{\phi}_j\rangle\end{aligned} \tag{8}$$

The squared norm of $\mathbf{z}$ is:

$$\|\mathbf{z}\|_2^2 = \sum_{i\in S} \alpha_i^2 \tag{9}$$

Step 2: Deviation from isometry and RIP constant

The RIP condition for κ-sparse vectors requires:

$$(1-\delta)\|\mathbf{z}\|_2^2 \leq \|\mathbf{\Phi z}\|_2^2 \leq (1+\delta)\|\mathbf{z}\|_2^2 \tag{10}$$

Subtracting $\|\mathbf{z}\|_2^2$, we obtain the deviation:

$$\|\mathbf{\Phi z}\|_2^2 - \|\mathbf{z}\|_2^2 = \sum_{i\in S} \alpha_i^2(\|\boldsymbol{\phi}_i\|_2^2 - 1) + \sum_{\substack{i,j\in S \\ i\neq j}} \alpha_i\alpha_j\langle\boldsymbol{\phi}_i, \boldsymbol{\phi}_j\rangle \tag{11}$$

Thus, the RIP constant δ is defined as the worst-case deviation over all κ-sparse unit-norm vectors $\mathbf{z}$:

$$\delta = \max_{\substack{S\subseteq\{1,\ldots,n\} \\ |S|=\kappa \\ \sum_{i\in S}\alpha_i^2=1}} \left| \sum_{i\in S} \alpha_i^2(\|\boldsymbol{\phi}_i\|_2^2 - 1) + \sum_{\substack{i,j\in S \\ i\neq j}} \alpha_i\alpha_j\langle\boldsymbol{\phi}_i, \boldsymbol{\phi}_j\rangle \right| \tag{12}$$

Step 3: Relationship between δ and SGD

For a κ-sparse $\mathbf{z}$, the RIP bound δ can be expressed as a weighted sum of activated column norms of $\mathbf{\Phi}$ and cross correlations between them. However, the presence of aligned columns leads to large correlations increasing δ value.

Assume a case with $\mathbf{w}^{(1)} \neq \mathbf{w}^{(2)}$ where both vectors satisfy (5) and (6) respectively. Define the *active set* $\mathcal{S}_m = \{\, j \mid w_j^{(m)} > 0\}$ for $m = 1, 2$. Note that each $\mathbf{w}^{(m)}$ is strictly positive in its active coordinates, and hence corresponds to selecting a certain subset of columns from $\mathbf{\Phi}$. Our goal is to show this situation cannot arise if $\mathbf{\Phi}$ is well-conditioned.

Without loss of generality, pick an index $j \in \mathcal{S}_1$ but $j \notin \mathcal{S}_2$. Thus, $\mathbf{w}^{(1)}$ "turns on" column $\boldsymbol{\phi}_j$ while $\mathbf{w}^{(2)}$ has $w_j^{(2)} = 0$. If $\boldsymbol{\phi}_j$ is co-linear with other columns it will lead to a large δ. Furthermore, it also means a small difference in $\Sigma^{-1}(\mathbf{w}^{(1)})$ and $\Sigma^{-1}(\mathbf{w}^{(2)})$ as $\mathbf{w}^{(1)}$ and $\mathbf{w}^{(2)}$ differ along the indices of the aligned columns having large correlations. It is due to the common large correlation term with $\boldsymbol{\phi}_j$ both in $\Sigma^{-1}(\mathbf{w}^{(1)})$ and $\Sigma^{-1}(\mathbf{w}^{(2)})$, they will have a small difference.

Therefore small δ suggests small correlation and therefore a larger separation between $\Sigma^{-1}(\mathbf{w}^{(1)})$ and $\Sigma^{-1}(\mathbf{w}^{(2)})$. This indicates that a small δ is essential to find the true local/global optimum.

References

Candes, E. J. and Tao, T. Decoding by linear programming. *IEEE transactions on information theory*, 51(12):4203–4215, 2005.