Learning Sparse Representations of High Dimensional Data on Large Scale Dictionaries (Supplemental Material)

1 Derivation of the dual formulation of the lasso problem

In this section, we prove that given the primal problem (i.e. the lasso problem),

$$\min_{w_1,w_2,\ldots,w_m} \quad \frac{1}{2}\|\mathbf{x}-\sum_{i=1}^{m} w_i\mathbf{b}_i\|_2^2 + \lambda\sum_{i=1}^{m}|w_i|, \tag{1}$$

the dual problem is

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{\lambda^2}{2}\|\boldsymbol{\theta}-\frac{\mathbf{x}}{\lambda}\|_2^2 \\ \text{s.t.} \quad & |\mathbf{b}_i^T\boldsymbol{\theta}| \le 1 \quad \forall i = 1,2,\ldots,m, \end{aligned} \tag{2}$$

and that the relationship between the optimal solution $\tilde{w}_i$ of (1) and the optimal solution $\tilde{\boldsymbol{\theta}}$ of (2) is

$$\mathbf{x} = \sum_{i=1}^{m}\tilde{w}_i\mathbf{b}_i + \lambda\tilde{\boldsymbol{\theta}}, \qquad \mathbf{b}_i^T\tilde{\boldsymbol{\theta}} \in \begin{cases} \{1\} & \text{if } \tilde{w}_i > 0, \\ \{-1\} & \text{if } \tilde{w}_i < 0, \\ [-1,1] & \text{if } \tilde{w}_i = 0. \end{cases} \tag{3}$$

To prove this, we consider a more general problem called the nonnegative lasso problem:

$$\min_{w_i\ge 0} \quad \frac{1}{2}\|\mathbf{x}-\sum_{i=1}^{m} w_i\mathbf{b}_i\|_2^2 + \lambda\sum_{i=1}^{m} w_i. \tag{4}$$

It suffices to prove that the dual problem of the nonnegative lasso problem (4) is

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{\lambda^2}{2}\|\boldsymbol{\theta}-\frac{\mathbf{x}}{\lambda}\|_2^2 \\ \text{s.t.} \quad & \mathbf{b}_i^T\boldsymbol{\theta} \le 1 \quad \forall i = 1,2,\ldots,m, \end{aligned} \tag{5}$$

and that the relationship between the optimal solution $\tilde{w}_i$ of (4) and the optimal solution $\tilde{\boldsymbol{\theta}}$ of (5) is

$$\mathbf{x} = \sum_{i=1}^{m}\tilde{w}_i\mathbf{b}_i + \lambda\tilde{\boldsymbol{\theta}}, \qquad \mathbf{b}_i^T\tilde{\boldsymbol{\theta}} \in \begin{cases} \{1\} & \text{if } \tilde{w}_i > 0, \\ [-\infty,1] & \text{if } \tilde{w}_i = 0. \end{cases} \tag{6}$$

Because if we can prove that (5) is the dual problem of (4) via relationship (6). Then for the standard lasso problem (1) without the nonnegative constraint, we can simply replace the codewords

$\{\mathbf{b}_i\}$ with $\{\pm\mathbf{b}_i\}$ and the weights $\{w_i\}$ with $\{\max\{w_i, 0\}, \max\{-w_i, 0\}\}$. This will transform the standard lasso problem into a nonnegative lasso problem. Applying the results of the nonnegative lasso problem proves that (2) is the dual problem of (1) via relationship (3).

To derive the dual problem of (4), introduce dummy variable $\boldsymbol{\nu}$ with $\lambda\boldsymbol{\nu} = \mathbf{x} - \sum_{i=1}^m w_i\mathbf{b}_i$ and rewrite the primal problem (4) as:

$$\begin{aligned} \min \quad & \frac{\lambda^2}{2}\|\boldsymbol{\nu}\|_2^2 + \lambda\sum_{i=1}^m w_i, \\ \text{s.t.} \quad & -w_i \le 0 \\ & \mathbf{x} - \sum_{i=1}^m w_i\mathbf{b}_i = \lambda\boldsymbol{\nu}. \end{aligned} \tag{7}$$

Apparently, the Slater's condition holds because a strictly feasible solution exists (for example, setting $w_i = 1, i = 1, 2, \ldots, m$). Therefore we can use the strong duality and the standard optimization procedure (see [1]). By introducing the Lagrangian multipliers $\boldsymbol{\eta} = (\eta_1, \eta_2, \ldots, \eta_m)$ and $\lambda\boldsymbol{\theta}$, the Lagrangian can be written as:

$$L(\mathbf{w}, \boldsymbol{\nu}, \boldsymbol{\eta}, \boldsymbol{\theta}) = \frac{\lambda^2}{2}\|\boldsymbol{\nu}\|_2^2 + \lambda\sum_{i=1}^m w_i + \sum_{i=1}^m \eta_i(-w_i) + \lambda\boldsymbol{\theta}^T\left(\mathbf{x} - \sum_{i=1}^m w_i\mathbf{b}_i - \lambda\boldsymbol{\nu}\right). \tag{8}$$

Now we solve for the Lagrangian dual function, which is defined as $g(\boldsymbol{\eta}, \boldsymbol{\theta}) = \inf_{\mathbf{w},\boldsymbol{\nu}} L(\mathbf{w}, \boldsymbol{\nu}, \boldsymbol{\eta}, \boldsymbol{\theta})$. Since (8) is a linear function in w_i, $g(\boldsymbol{\eta}, \boldsymbol{\theta})$ is not $-\infty$ only when the coefficient before each w_i is 0, i.e., when $\eta_i = \lambda - \lambda\boldsymbol{\theta}^T\mathbf{b}_i$. And when this is the case,

$$L(\mathbf{w}, \boldsymbol{\nu}, \boldsymbol{\eta}, \boldsymbol{\theta}) = \frac{\lambda^2}{2}\|\boldsymbol{\nu}\|_2^2 + \lambda\boldsymbol{\theta}^T(\mathbf{x} - \lambda\boldsymbol{\nu}) = \frac{\lambda^2}{2}\|\boldsymbol{\nu} - \boldsymbol{\theta}\|_2^2 + \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{\lambda^2}{2}\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda}\|_2^2. \tag{9}$$

To minimize this we also need $\boldsymbol{\nu} = \boldsymbol{\theta}$. Therefore the Lagrange dual function is:

$$g(\boldsymbol{\eta}, \boldsymbol{\theta}) = \begin{cases} \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{\lambda^2}{2}\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda}\|_2^2 & \text{if } \eta_i = \lambda - \lambda\boldsymbol{\theta}^T\mathbf{b}_i, \forall i = 1, 2, \ldots, m \\ -\infty & \text{otherwise} \end{cases} \tag{10}$$

And the dual problem:

$$\begin{aligned} \max \quad & g(\boldsymbol{\eta}, \boldsymbol{\theta}) \\ \text{s.t.} \quad & \eta_i \ge 0, i = 1, 2, \ldots, m, \end{aligned} \tag{11}$$

can be equivalently written as

$$\begin{aligned} \max_{\boldsymbol{\theta}} \quad & \frac{1}{2}\|\mathbf{x}\|_2^2 - \frac{\lambda^2}{2}\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda}\|_2^2 \\ \text{s.t.} \quad & \lambda(1 - \boldsymbol{\theta}^T\mathbf{b}_i) \ge 0, i = 1, 2, \ldots, m, \end{aligned} \tag{12}$$

which is apparently equivalent to (5). The relationship in (6) follows from the optimality condition $\boldsymbol{\nu} = \boldsymbol{\theta}$ and applying complementary slackness $\eta_i w_i = \lambda(1 - \boldsymbol{\theta}^T\mathbf{b}_i)w_i = 0$ on the optimal solutions.

2 Proof of Lemma 1

Lemma 1. *If the optimal solution $\tilde{\boldsymbol{\theta}}$ of* (2) *satisfies $\|\tilde{\boldsymbol{\theta}} - \mathbf{q}\|_2 \le r$, then $|\mathbf{b}_i^T\mathbf{q}| < (1-r) \Rightarrow w_i = 0$.*

Proof. Assume that we have $|\mathbf{b}_i^T\mathbf{q}| < (1-r)$. According to (3), in order to assert that $w_i = 0$, we only need to prove that for the optimal solution $\tilde{\boldsymbol{\theta}}$ of (2): $|\mathbf{b}_i^T\tilde{\boldsymbol{\theta}}| < 1$, which can be proved by:

$$\begin{aligned}
|\mathbf{b}_i^T\tilde{\boldsymbol{\theta}}| &= |\mathbf{b}_i^T(\tilde{\boldsymbol{\theta}} - \mathbf{q}) + \mathbf{b}_i^T\mathbf{q}| \\
&\leq |\mathbf{b}_i^T(\tilde{\boldsymbol{\theta}} - \mathbf{q})| + |\mathbf{b}_i^T\mathbf{q}| \\
&\leq \|\mathbf{b}_i\|_2\|\boldsymbol{\theta} - \mathbf{q}\|_2 + |\mathbf{b}_i^T\mathbf{q}| \\
&< r + (1-r) = 1.
\end{aligned} \tag{13}$$

The first inequality is a simple triangle inequality. The second inequality uses the Cauchy-Schwarz inequality. The third inequality uses our assumptions $\|\boldsymbol{\theta} - \mathbf{q}\|_2 \leq r$ and $|\mathbf{b}_i^T\mathbf{q}| < (1-r)$. □

3 Proof of Lemma 2

Lemma 2. *Given* $\lambda_{\max} = \mathbf{x}^T\mathbf{b}_*$, $\|\mathbf{x}\|_2 = \|\mathbf{b}_*\|_2 = 1$. *If* $\boldsymbol{\theta}$ *satisfies*

$$\begin{aligned}
&(a) \qquad \|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda}\|_2 \leq \frac{1}{\lambda} - \frac{1}{\lambda_{max}}, \\
&(b) \qquad \boldsymbol{\theta}^T\mathbf{b}_* \leq 1,
\end{aligned}$$

then $\boldsymbol{\theta}$ *must also satisfy*

$$\begin{aligned}
&(c) \qquad \|\boldsymbol{\theta} - (\frac{\mathbf{x}}{\lambda} - (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*)\|_2 \leq \sqrt{\frac{1}{\lambda_{max}^2} - 1}\left(\frac{\lambda_{\max}}{\lambda} - 1\right), \\
&(d) \qquad \|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda_{\max}}\|_2 \leq 2\sqrt{\frac{1}{\lambda_{max}^2} - 1}\left(\frac{\lambda_{\max}}{\lambda} - 1\right).
\end{aligned}$$

Proof. We first prove (c) by

$$\begin{aligned}
&(\frac{1}{\lambda} - \frac{1}{\lambda_{\max}})^2 \geq \|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda}\|_2^2 \qquad \text{(by assumption (a))} \\
=&\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_* - (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*\|_2^2 \\
=&\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*\|_2^2 + \|(\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*\|_2^2 - 2\left(\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*\right)^T(\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_* \\
=&\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*\|_2^2 + (\frac{\lambda_{\max}}{\lambda} - 1)^2 - 2(\frac{\lambda_{\max}}{\lambda} - 1)\left(\boldsymbol{\theta}^T\mathbf{b}_* - \frac{\mathbf{x}^T\mathbf{b}_*}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\|\mathbf{b}_*\|_2^2\right) \\
=&\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*\|_2^2 + (\frac{\lambda_{\max}}{\lambda} - 1)^2 - 2(\frac{\lambda_{\max}}{\lambda} - 1)\left(\boldsymbol{\theta}^T\mathbf{b}_* - \frac{\lambda_{\max}}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\right) \\
=&\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*\|_2^2 + (\frac{\lambda_{\max}}{\lambda} - 1)^2 + 2(\frac{\lambda_{\max}}{\lambda} - 1)(1 - \boldsymbol{\theta}^T\mathbf{b}_*) \\
\geq&\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*\|_2^2 + (\frac{\lambda_{\max}}{\lambda} - 1)^2 \qquad \text{(by assumption (b))}.
\end{aligned}$$

This gives us:

$$\|\boldsymbol{\theta} - \frac{\mathbf{x}}{\lambda} + (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*\|_2^2 \leq \sqrt{(\frac{1}{\lambda} - \frac{1}{\lambda_{\max}})^2 - (\frac{\lambda_{\max}}{\lambda} - 1)^2} = \sqrt{\frac{1}{\lambda_{max}^2} - 1}\left(\frac{\lambda_{\max}}{\lambda} - 1\right), \tag{14}$$

which is (c). To prove (d), we first prove an intermediate result:

$$\|\frac{\mathbf{x}}{\lambda}-(\frac{\lambda_{\max}}{\lambda}-1)\mathbf{b}_*-\frac{\mathbf{x}}{\lambda_{\max}}\|_2=\sqrt{\frac{1}{\lambda_{max}^2}-1}\left(\frac{\lambda_{\max}}{\lambda}-1\right). \tag{15}$$

This can be proved by

$$\begin{aligned}
&\|\frac{\mathbf{x}}{\lambda}-(\frac{\lambda_{\max}}{\lambda}-1)\mathbf{b}_*-\frac{\mathbf{x}}{\lambda_{\max}}\|_2^2\\
=&\|\frac{\mathbf{x}}{\lambda}-\frac{\mathbf{x}}{\lambda_{\max}}\|_2^2+\|(\frac{\lambda_{\max}}{\lambda}-1)\mathbf{b}_*\|_2^2-2\left(\frac{\mathbf{x}}{\lambda}-\frac{\mathbf{x}}{\lambda_{\max}}\right)^T(\frac{\lambda_{\max}}{\lambda}-1)\mathbf{b}_*\\
=&(\frac{1}{\lambda}-\frac{1}{\lambda_{\max}})^2\|\mathbf{x}\|_2^2+(\frac{\lambda_{\max}}{\lambda}-1)^2\|\mathbf{b}_*\|_2^2-2\left(\frac{1}{\lambda}-\frac{1}{\lambda_{\max}}\right)(\frac{\lambda_{\max}}{\lambda}-1)\mathbf{x}^T\mathbf{b}_*\\
=&(\frac{1}{\lambda}-\frac{1}{\lambda_{\max}})^2+(\frac{\lambda_{\max}}{\lambda}-1)^2-2\left(\frac{1}{\lambda}-\frac{1}{\lambda_{\max}}\right)(\frac{\lambda_{\max}}{\lambda}-1)\lambda_{\max}=\left(\frac{1}{\lambda_{max}^2}-1\right)\left(\frac{\lambda_{\max}}{\lambda}-1\right)^2,
\end{aligned}$$

which is the square of (15). With (c) and (15), (d) can be proved by a simple triangle inequality:

$$\begin{aligned}
\|\boldsymbol{\theta}-\frac{\mathbf{x}}{\lambda_{\max}}\|_2\leq&\|\boldsymbol{\theta}-\frac{\mathbf{x}}{\lambda}+(\frac{\lambda_{\max}}{\lambda}-1)\mathbf{b}_*\|_2+\|\frac{\mathbf{x}}{\lambda}-(\frac{\lambda_{\max}}{\lambda}-1)\mathbf{b}_*-\frac{\mathbf{x}}{\lambda_{\max}}\|_2\\
\leq&\sqrt{\frac{1}{\lambda_{max}^2}-1}\left(\frac{\lambda_{\max}}{\lambda}-1\right)+\sqrt{\frac{1}{\lambda_{max}^2}-1}\left(\frac{\lambda_{\max}}{\lambda}-1\right)\\
=&2\sqrt{\frac{1}{\lambda_{max}^2}-1}\left(\frac{\lambda_{\max}}{\lambda}-1\right).
\end{aligned} \tag{16}$$

□

4 Proof of Lemma 3

Lemma 3. *When $\lambda_{max}>\sqrt{3}/2$, if ST1/SAFE discards $\mathbf{b}_i$, then ST2 also discards $\mathbf{b}_i$.*

Proof. If ST1/SAFE discards $\mathbf{b}_i$, then we must have $0\leq|\mathbf{x}^T\mathbf{b}_i|<\lambda-1+\lambda/\lambda_{max}$. In order to prove that ST2 also discards $\mathbf{b}_i$, we only need to prove the following inequality:

$$\lambda-1+\frac{\lambda}{\lambda_{max}}<\lambda_{max}\left(1-2\sqrt{\frac{1}{\lambda_{max}^2}-1}\left(\frac{\lambda_{\max}}{\lambda}-1\right)\right). \tag{17}$$

We calculate the difference of the two sides in (17):

$$\begin{aligned}
&\text{R.H.S. of (17)}-\text{L.H.S. of (17)}\\
=&\lambda_{max}\left(1-2\sqrt{\frac{1}{\lambda_{max}^2}-1}\left(\frac{\lambda_{\max}}{\lambda}-1\right)\right)-(\lambda-1+\frac{\lambda}{\lambda_{max}})\\
=&\lambda_{max}-2\sqrt{1-\lambda_{max}^2}(\frac{\lambda_{max}}{\lambda}-1)-(\lambda-1+\frac{\lambda}{\lambda_{max}})\\
=&\left(\frac{\lambda_{max}-\lambda}{\lambda_{\max}\lambda}\right)\left(\lambda-2\lambda_{max}\sqrt{\frac{1-\lambda_{max}}{1+\lambda_{max}}}\right)
\end{aligned} \tag{18}$$

We need to prove that this is positive. We have already known that $\lambda_{\max} > \lambda$. From $0 < \lambda - 1 + \lambda/\lambda_{max}$ we know that $\lambda > \frac{\lambda_{max}}{\lambda_{max}+1}$. When $\lambda > \sqrt{3}/2$ we have $\frac{\lambda_{max}}{\lambda_{max}+1} > 2\lambda_{max}\sqrt{\frac{1-\lambda_{max}}{1+\lambda_{max}}}$. Therefore $\lambda > \frac{\lambda_{max}}{\lambda_{max}+1} > 2\lambda_{max}\sqrt{\frac{1-\lambda_{max}}{1+\lambda_{max}}}$. So by (18) the R.H.S of (17) is indeed greater than the L.H.S. of (17). □

5 Proof of Lemma 4

Lemma 4. *Given any* $\mathbf{x}, \mathbf{b}_*$ *and* λ*, if ST2 discards* $\mathbf{b}_i$*, then ST3 also discards* $\mathbf{b}_i$*.*

Proof. If ST2 discards $\mathbf{b}_i$, then we have

$$|\mathbf{x}^T\mathbf{b}_i| < \lambda_{\max}\left(1 - 2\sqrt{\frac{1}{\lambda_{max}^2} - 1}\left(\frac{\lambda_{\max}}{\lambda} - 1\right)\right). \tag{19}$$

We can prove that $\mathbf{b}_i$ also satisfies the discarding criteria of ST3:

$$\begin{aligned}
&|\mathbf{x}^T\mathbf{b}_i - (\lambda_{\max} - \lambda)\mathbf{b}_*^T\mathbf{b}_i| \\
=&\lambda\left|\frac{\mathbf{x}^T\mathbf{b}_i}{\lambda} - (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*^T\mathbf{b}_i - \frac{\mathbf{x}^T\mathbf{b}_i}{\lambda_{\max}} + \frac{\mathbf{x}^T\mathbf{b}_i}{\lambda_{\max}}\right| \\
\leq&\lambda\left(\left|\frac{\mathbf{x}^T\mathbf{b}_i}{\lambda} - (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_*^T\mathbf{b}_i - \frac{\mathbf{x}^T\mathbf{b}_i}{\lambda_{\max}}\right| + \left|\frac{\mathbf{x}^T\mathbf{b}_i}{\lambda_{\max}}\right|\right) \\
\leq&\lambda\left(\|\frac{\mathbf{x}^T}{\lambda} - (\frac{\lambda_{\max}}{\lambda} - 1)\mathbf{b}_* - \frac{\mathbf{x}^T}{\lambda_{\max}}\|_2\|\mathbf{b}_i\|_2 + \left|\frac{\mathbf{x}^T\mathbf{b}_i}{\lambda_{\max}}\right|\right) \\
<&\lambda\left(\sqrt{\frac{1}{\lambda_{max}^2} - 1}\left(\frac{\lambda_{\max}}{\lambda} - 1\right) + 1 - 2\sqrt{\frac{1}{\lambda_{max}^2} - 1}\left(\frac{\lambda_{\max}}{\lambda} - 1\right)\right) \\
=&\lambda\left(1 - \sqrt{\frac{1}{\lambda_{max}^2} - 1}\left(\frac{\lambda_{\max}}{\lambda} - 1\right)\right)
\end{aligned} \tag{20}$$

The first inequality is a simple triangle inequality. The second inequality uses the Cauchy-Schwarz inequality. The third inequality uses the intermediate result (15) in proving Lemma 2, $\|\mathbf{b}_i\|_2 = 1$, and (19). □

6 Proof of Theorem 2

Theorem 2. *Assume that* $\mathcal{X}$ *satisfies SI and has a* κ*-sparse representation using dictionary* $\mathbf{B}$*. Then the projected data* $T(\mathcal{X})$ *satisfies SI if*

$$(2\kappa - 1)M(\mathbf{TB}) < 1, \tag{21}$$

where $M(\cdot)$ *is the mutual coherence of a matrix.*

Proof. If $T(\mathcal{X})$ doesn't satisfy SI, then there exists $(\mathbf{x}_1, \mathbf{x}_2) \in \mathcal{X} \times \mathcal{X}$ and $\gamma \notin \{0, 1\}$ so that $\mathbf{Tx}_1 = \gamma\mathbf{Tx}_2$. Let $\mathbf{x}_1 = \mathbf{Bw}_1$ and $\mathbf{x}_2 = \mathbf{Bw}_2$. We have $\mathbf{TB}(\mathbf{w}_1 - \gamma\mathbf{w}_2) = 0$. Both $\mathbf{w}_1$ and $\mathbf{w}_2$ are κ sparse so $(\mathbf{w}_1 - \gamma\mathbf{w}_2)$ is at most 2κ sparse and nonzero (otherwise contradicting with the SI property of $\mathcal{X}$). However, it's well know that the minimum l_0-norm of vectors in the null space of $\mathbf{TB}$ (i.e. the "spark" of $\mathbf{TB}$) is lower bounded by $1 + 1/M(\mathbf{TB})$ (Lemma 2.1, [2]). So, $2\kappa \geq \|\mathbf{w}_1 - \gamma\mathbf{w}_2\|_0 \geq 1 + 1/M(\mathbf{TB})$, contradicting (21). Therefore $T(\mathcal{X})$ satisfies SI. □

7 Proof of Theorem 3

Theorem 3. *Let the data points lie on a K-dimensional Riemannian submanifold $\mathcal{X} \subset \mathbb{R}^p$ that is compact, has volume V, conditional number $1/\tau$, and geodesic covering regularity R (see [3]). Assume that in the optimal solution of the sparse representation problem for the projected data:*

$$\begin{aligned} \min_{\mathbf{B},\mathbf{W}} \quad & \frac{1}{2}\|\mathbf{TX} - \mathbf{BW}\|_F^2 + \lambda\|\mathbf{W}\|_1 \\ \textit{s.t.} \quad & \|\mathbf{b}_i\|_2^2 \leq 1, \quad \forall i = 1, 2, \ldots, m, \end{aligned} \tag{22}$$

data points $\mathbf{Tx}_1$ and $\mathbf{Tx}_2$ have nonzero weights on the same set of κ codewords. Let $\mathbf{w}_j$ be the new representation of $\mathbf{x}_j$ and $\mu_i = \|\mathbf{Tx}_j - \mathbf{Bw}_j\|_2$ be the length of the residual ($j = 1, 2$). With probability $1 - \rho$:

$$\begin{aligned} \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 &\leq \frac{p}{d}(1+\epsilon_1)(1+\epsilon_2)(\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2 + 2\mu_1^2 + 2\mu_2^2) \\ \|\mathbf{x}_1 - \mathbf{x}_2\|_2^2 &\geq \frac{p}{d}(1-\epsilon_1)(1-\epsilon_2)(\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2, \end{aligned}$$

with $\epsilon_1 = O((\frac{K\ln(NVR\tau^{-1})\ln(1/\rho)}{d})^{0.5-\eta})$ (for any small $\eta > 0$) and $\epsilon_2 = (\kappa - 1)M(\mathbf{B})$.

Proof. Using Theorem 3.1 in [3] on random projection $\mathbf{T}$ and the simple fact that $\forall\epsilon < 0.2$: $(1-\epsilon)^2 \geq \frac{1}{1+3\epsilon}, (1+\epsilon)^2 \leq \frac{1}{1-3\epsilon}$, for $d = O(\frac{K\ln(NVR\tau^{-1}\epsilon^{-1})\ln(1/\rho)}{\epsilon^2})$, with probability $1 - \rho$:

$$\frac{1}{(1+3\epsilon)}\frac{d}{p} \leq (1-\epsilon)^2\frac{d}{p} \leq \frac{\|\mathbf{Tx}_1 - \mathbf{Tx}_2\|_2^2}{\|\mathbf{x}_1 - \mathbf{x}_2\|_2^2} \leq (1+\epsilon)^2\frac{d}{p} \leq \frac{1}{(1-3\epsilon)}\frac{d}{p} \tag{23}$$

To bound $\|\mathbf{Tx}_1 - \mathbf{Tx}_2\|_2^2$, let $\mathbf{b}_i$ be a codeword in $\mathbf{B}$ that has nonzero weight, by (3) $(\mathbf{Tx}_1 - \mathbf{Bw}_1)^T\mathbf{b}_i = (\mathbf{Tx}_2 - \mathbf{Bw}_2)^T\mathbf{b}_i = \lambda \operatorname{sign} w_i$. So $(\mathbf{Tx}_1 - \mathbf{Bw}_1) - (\mathbf{Tx}_2 - \mathbf{Bw}_2)$ is orthogonal to any codewords $\mathbf{b}_i$ that has nonzero weight, and therefore is orthogonal to $\mathbf{B}(\mathbf{w}_1 - \mathbf{w}_2)$. Thus:

$$\|\mathbf{Tx}_1 - \mathbf{Tx}_2\|_2^2 = \|\mathbf{B}(\mathbf{w}_1 - \mathbf{w}_2)\|_2^2 + \|\mathbf{T}(\mathbf{x}_1 - \mathbf{x}_2) - \mathbf{B}(\mathbf{w}_1 - \mathbf{w}_2)\|_2^2 \tag{24}$$

Using (24) and the fact that any singular value σ of $\mathbf{B}$ satisfies $1 - (\kappa - 1)M(\mathbf{B}) \leq \sigma^2 \leq 1 + (\kappa - 1)M(\mathbf{B})$ (Proposition 4.3, [4]), we can upper bound and lower bound $\|\mathbf{Tx}_1 - \mathbf{Tx}_2\|_2^2$ by:

$$\begin{aligned} \|\mathbf{Tx}_1 - \mathbf{Tx}_2\|_2^2 &\leq \|\mathbf{B}(\mathbf{w}_1 - \mathbf{w}_2)\|_2^2 + 2(\|\mathbf{Tx}_1 - \mathbf{Bw}_1\|_2^2 + \|\mathbf{Tx}_2 - \mathbf{Bw}_2\|_2^2) \\ &= \|\mathbf{B}(\mathbf{w}_1 - \mathbf{w}_2)\|_2^2 + 2\mu_1^2 + 2\mu_2^2 \leq (1+\epsilon_2)\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2 + 2\mu_1^2 + 2\mu_2^2 \\ \|\mathbf{Tx}_1 - \mathbf{Tx}_2\|_2^2 &\geq \|\mathbf{B}(\mathbf{w}_1 - \mathbf{w}_2)\|_2^2 \geq (1-\epsilon_2)\|\mathbf{w}_1 - \mathbf{w}_2\|_2^2 \end{aligned} \tag{25}$$

Plug these into (23) gives us the desired bounds with $\epsilon_1 = 3\epsilon$ and by $d = O(\frac{K\ln(NVR\tau^{-1}\epsilon^{-1})\ln(1/\rho)}{\epsilon^2})$, $\epsilon_1 = O((\frac{K\ln(NVR\tau^{-1})\ln(1/\rho)}{d})^{0.5-\eta})$ for any small $\eta > 0$. □

References

[1] S.P. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge Univ Pr, 2004.

[2] M. Elad. *Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing*. Springer, 2010.

[3] R.G. Baraniuk and M.B. Wakin. Random projections of smooth manifolds. *Foundations of Computational Mathematics*, 9(1):51–77, 2007.

[4] J.A. Tropp. *Topics in sparse approximation*. PhD thesis, The University of Texas at Austin, 2004.