Appendix

A Proof of Sparse Coding Stability Theorem

The flow of this section is as follows. We first establish some preliminary notation and summarize important conditions. Several lemmas are then presented to support a key sparsity lemma. This sparsity lemma establishes that the solution to the perturbed problem is sparse provided the perturbation is not too large. Finally, the sparsity of this new solution is exploited to bound the difference of the new solution from the old solution. This flow is embodied by the proof flowchart in Figure 1.

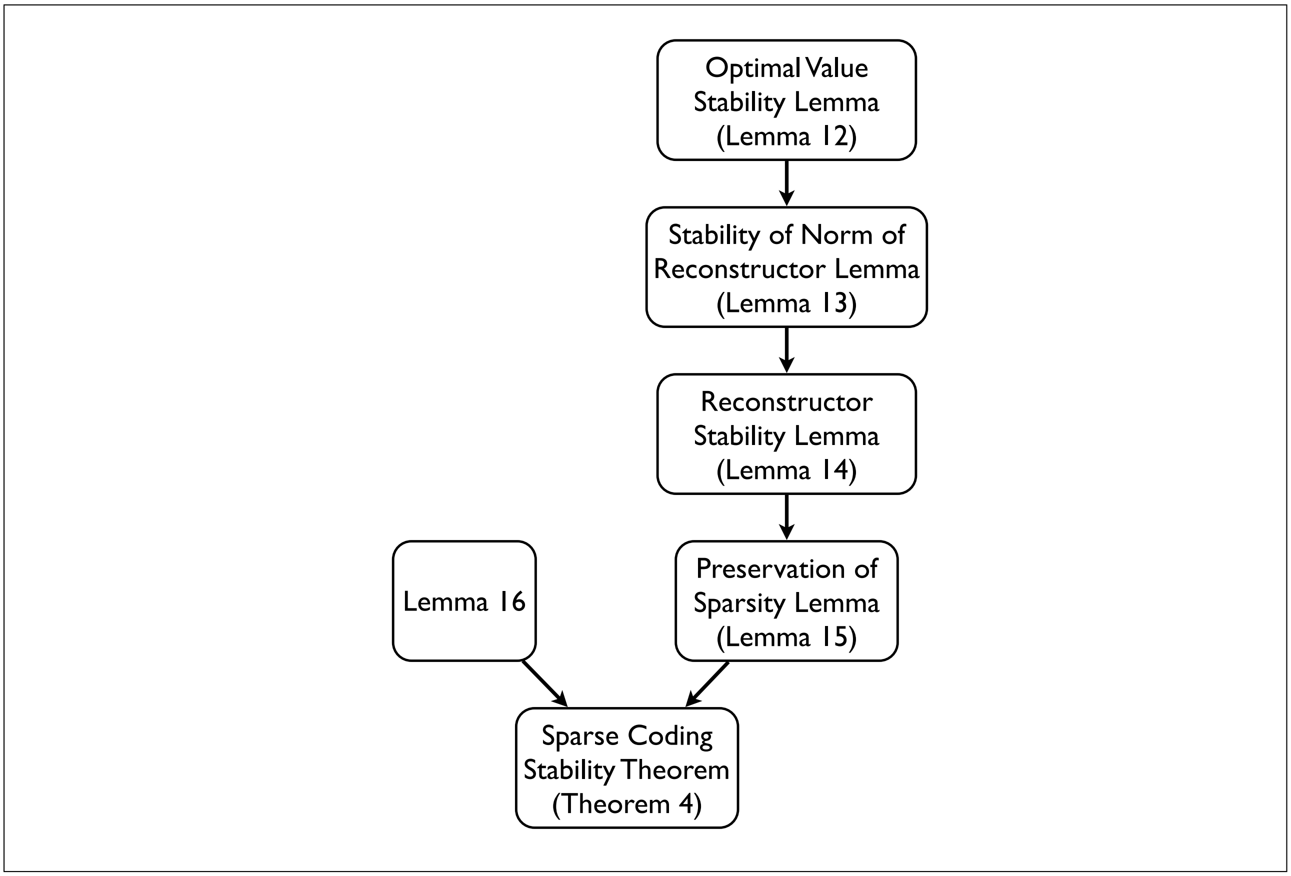
Optimal Value Stability Lemma (Lemma 12)
Stability of Norm of Reconstructor Lemma (Lemma 13)
Reconstructor Stability Lemma (Lemma 14)
Lemma 16
Preservation of Sparsity Lemma (Lemma 15)
Sparse Coding Stability Theorem (Theorem 4)

Figure 1: Proof flowchart for the Sparse Coding Stability Theorem (Theorem 4).

A.1 Notation and Assumptions

Let α and $\tilde{\alpha}$ respectively denote the solutions to the LASSO problems:

$$\alpha = \arg\min_z \frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1 \qquad\qquad \tilde{\alpha} = \arg\min_z \frac{1}{2}\|x - \tilde{D}z\|_2^2 + \lambda\|z\|_1.$$

First, let's review the optimality conditions for the LASSO (see Asif & Romberg, 2010, conditions L1 and L2):

$$\begin{aligned} \langle D_j, x - D\alpha\rangle &= \operatorname{sign}(\alpha_j)\lambda \quad \text{if } \alpha_j \neq 0, \\ |\langle D_j, x - D\alpha\rangle| &< \lambda \quad \text{otherwise.} \end{aligned}$$

Note that the above optimality conditions imply that if $\alpha_j \neq 0$ then

$$|\langle D_j, x - D\alpha\rangle| = \lambda.$$

Assumptions

The statement of the Sparse Coding Stability Theorem (Theorem 4) makes the following assumptions:

(A1) - Closeness D and $\tilde{D}$ are close, as measured by operator norm:

$$\|\tilde{D} - D\|_2 \leq \varepsilon.$$

(A2) - Incoherence There is a $\mu > 0$ such that, for all $J \subseteq [k]$ satisfying $|J| = s$:

$$\sigma_{\min}(D_J) \geq \mu.$$

(A3) - Sparsity with Margin For some fixed $\tau > 0$, there is a $\mathcal{I} \subseteq [k]$ with $|\mathcal{I}| = k - s$ such that for all $i \in \mathcal{I}$:

$$|\langle D_i, x - D\alpha\rangle| < \lambda - \tau.$$

Consequently, all $i \in \mathcal{I}$ satisfy $\alpha_i = 0$.

A.2 Useful Observations

Let v_D^* be the optimal value of the LASSO for dictionary D:

$$\begin{aligned} v_D^* &= \min_z \frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1 \\ &= \frac{1}{2}\|x - D\alpha\|_2^2 + \lambda\|\alpha\|_1 \end{aligned}$$

Likewise, let

$$v_{\tilde{D}}^* = \frac{1}{2}\|x - \tilde{D}\tilde{\alpha}\|_2^2 + \lambda\|\tilde{\alpha}\|_1$$

The first observation is that the values of the optimal solutions are close:

Lemma 12 (Optimal Value Stability) *If* $\|D - \tilde{D}\|_2 \leq \varepsilon$, *then*

$$\left|v_D^* - v_{\tilde{D}}^*\right| \leq \frac{5}{8}\frac{\varepsilon}{\lambda}.$$

Proof The proof is simple:

$$\begin{aligned}
v^*_{\tilde{D}} &\le \frac{1}{2}\|x - \tilde{D}\alpha\|_2^2 + \lambda\|\alpha\|_1 \\
&= \frac{1}{2}\|x - D\alpha + (D - \tilde{D})\alpha\|_2^2 + \lambda\|\alpha\|_1 \\
&\le \frac{1}{2}\left(\|x - D\alpha\|_2^2 + 2\|x - D\alpha\|_2\|(D - \tilde{D})\alpha\|_2 + \|(D - \tilde{D})\alpha\|_2^2\right) + \lambda\|\alpha\|_1 \\
&\le \frac{1}{2}\|x - D\alpha\|_2^2 + \lambda\|\alpha\|_1 + \frac{1}{2}\left(\frac{\varepsilon}{\lambda} + \frac{1}{4}\left(\frac{\varepsilon}{\lambda}\right)^2\right) \\
&\le v^*_D + \frac{5}{8}\frac{\varepsilon}{\lambda}
\end{aligned}$$

for $\frac{\varepsilon}{\lambda} \le 1$. A symmetric argument shows that $v^*_D \le v^*_{\tilde{D}} + \frac{5}{8}\frac{\varepsilon}{\lambda}$. ■

The second observation shows that the norms of the optimal reconstructors are close.

Lemma 13 (Stability of Norm of Reconstructor) *If* $\|D - \tilde{D}\|_2 \le \varepsilon$, *then*

$$\left|\|D\alpha\|_2^2 - \|\tilde{D}\tilde{\alpha}\|_2^2\right| \le \frac{5}{4}\frac{\varepsilon}{\lambda}.$$

Showing this is more involved than the previous observation.
Proof First, we claim (and show) that

$$(x - D\alpha)^T D\alpha = \lambda\|\alpha\|_1. \tag{1}$$

The proof of the claim comes directly from Osborne et al. (2000, circa (2.8)). To see (1), recall that the LASSO objective is

$$\underset{z}{\text{minimize}} \quad \frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1.$$

The subgradient of this objective with respect to z is

$$-D^T(x - Dz) + \lambda v,$$

where $v_j = 1$ if $z_j > 0$, $v_j = -1$ if $z_j < 0$, and $v_j \in [-1, 1]$ if $z_j = 0$. From the definition of v, it follows that

$$v^T z = \|z\|_1.$$

At an optimal point α, $\partial_z \mathcal{L}(\alpha, \lambda) = 0$, and hence

$$\begin{gathered}
D^T(x - D\alpha) = \lambda v \\
\Updownarrow \\
(x - D\alpha)^T D = \lambda v^T \\
\Downarrow \\
(x - D\alpha)^T D\alpha = \lambda v^T \alpha \\
\Updownarrow \\
(x - D\alpha)^T D\alpha = \lambda\|\alpha\|_1,
\end{gathered}$$

as claimed.

Now, we use the fact that the values of the optimal solutions are close (Lemma 12):

$$\left|v_D^* - v_{\tilde{D}}^*\right| \leq \frac{5}{8}\frac{\varepsilon}{\lambda}.$$

But v_D^* is just

$$\begin{aligned}
\frac{1}{2}\langle x - D\alpha, x - D\alpha\rangle + \lambda\|\alpha\|_1 &= \frac{1}{2}\langle x - D\alpha, x - D\alpha\rangle + \langle x - D\alpha, D\alpha\rangle \\
&= \frac{1}{2}\langle x, x - D\alpha\rangle - \frac{1}{2}\langle x - D\alpha, D\alpha\rangle + \langle x - D\alpha, D\alpha\rangle \\
&= \frac{1}{2}\left(\langle x, x - D\alpha\rangle + \langle x - D\alpha, D\alpha\rangle\right) \\
&= \frac{1}{2}\langle x + D\alpha, x - D\alpha\rangle \\
&= \frac{1}{2}\left(\|x\|_2^2 - \|D\alpha\|_2^2\right).
\end{aligned}$$

Consequently,

$$\left|\frac{1}{2}\left(\|x\|_2^2 - \|D\alpha\|_2^2\right) - \frac{1}{2}\left(\|x\|_2^2 - \|\tilde{D}\tilde{\alpha}\|_2^2\right)\right| \leq \frac{5}{8}\frac{\varepsilon}{\lambda}$$

and hence

$$\left|\|D\alpha\|_2^2 - \|\tilde{D}\tilde{\alpha}\|_2^2\right| \leq \frac{5}{4}\frac{\varepsilon}{\lambda}.$$

■

Finally, we prove stability of the optimal reconstructor. Rather than showing that $\|D\alpha - \tilde{D}\tilde{\alpha}\|_2^2$ is $O(\varepsilon)$, it will be more convenient for later purposes to prove the following roughly equivalent result.

Lemma 14 (Reconstructor Stability) *If $\|D - \tilde{D}\|_2 \leq \varepsilon$, then*

$$\|D\alpha - D\tilde{\alpha}\|_2^2 \leq \frac{13\varepsilon}{\lambda}.$$

Proof Let $\alpha' := \frac{1}{2}(\alpha + \tilde{\alpha})$. From the optimality of α, it follows that $v_D(\alpha) \leq v_D(\alpha')$, or more explicitly:

$$\frac{1}{2}\|x - D\alpha\|_2^2 + \lambda\|\alpha\|_1 \leq \frac{1}{2}\|x - D\alpha'\|_2^2 + \lambda\|\alpha'\|_1. \tag{2}$$

First, note that $\left|\|D\tilde{\alpha}\|_2^2 - \|\tilde{D}\tilde{\alpha}\|_2^2\right| \leq \frac{7}{4}\frac{\varepsilon}{\lambda}$, because

$$\begin{aligned}
\left|\|D\tilde{\alpha}\|_2^2 - \|\tilde{D}\tilde{\alpha}\|_2^2\right| &\leq 2\left|\langle D\tilde{\alpha}, (\tilde{D} - D)\alpha\rangle\right| + \|(\tilde{D} - D)\tilde{\alpha}\|_2^2 \\
&\leq 2\|D\tilde{\alpha}\|_2\|\tilde{D} - D\|_2\|\tilde{\alpha}\|_2 + \left(\|\tilde{D} - D\|_2\|\tilde{\alpha}\|_2\right)^2 \\
&\leq 2\left(1 + \frac{\varepsilon}{2\lambda}\right)\frac{\varepsilon}{2\lambda} + \frac{1}{4}\left(\frac{\varepsilon}{\lambda}\right)^2 \\
&\leq \frac{7}{4}\frac{\varepsilon}{\lambda},
\end{aligned}$$

assuming $\varepsilon \leq \lambda$. Combining this fact with Lemma 13, $\left|\|D\alpha\|_2^2 - \|\tilde{D}\tilde{\alpha}\|_2^2\right| \leq \frac{5}{4}\frac{\varepsilon}{\lambda}$, yields

$$\left|\|D\alpha\|_2^2 - \|D\tilde{\alpha}\|_2^2\right| \leq \frac{3\varepsilon}{\lambda}.$$

By the convexity of the 1-norm, the RHS of (2) obeys:

$$
\begin{aligned}
&\frac{1}{2}\left\|x - D\left(\frac{\alpha+\tilde{\alpha}}{2}\right)\right\|_2^2 + \lambda\left\|\frac{\alpha+\tilde{\alpha}}{2}\right\|_1 \\
&\leq \frac{1}{2}\left\|x - \frac{1}{2}(D\alpha + D\tilde{\alpha})\right\|_2^2 + \frac{\lambda}{2}\|\alpha\|_1 + \frac{\lambda}{2}\|\tilde{\alpha}\|_1 \\
&= \frac{1}{2}\left(\|x\|_2^2 - 2\langle x, \frac{1}{2}\left(D\alpha + D\tilde{\alpha}\right)\rangle + \frac{1}{4}\|D\alpha + D\tilde{\alpha}\|_2^2\right) + \frac{\lambda}{2}\|\alpha\|_1 + \frac{\lambda}{2}\|\tilde{\alpha}\|_1 \\
&= \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\langle x, D\alpha\rangle - \frac{1}{2}\langle x, D\tilde{\alpha}\rangle + \frac{1}{8}\left(\|D\alpha\|_2^2 + \|D\tilde{\alpha}\|_2^2 + 2\langle D\alpha, D\tilde{\alpha}\rangle\right) + \frac{\lambda}{2}\|\alpha\|_1 + \frac{\lambda}{2}\|\tilde{\alpha}\|_1 \\
&\leq \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\langle x, D\alpha\rangle - \frac{1}{2}\langle x, D\tilde{\alpha}\rangle + \frac{1}{4}\|D\alpha\|_2^2 + \frac{1}{4}\langle D\alpha, D\tilde{\alpha}\rangle + \frac{\lambda}{2}\|\alpha\|_1 + \frac{\lambda}{2}\|\tilde{\alpha}\|_1 + \frac{3}{8}\frac{\varepsilon}{\lambda} \\
&= \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\langle x, D\alpha\rangle - \frac{1}{2}\langle x, D\tilde{\alpha}\rangle + \frac{1}{4}\|D\alpha\|_2^2 + \frac{1}{4}\langle D\alpha, D\tilde{\alpha}\rangle + \frac{1}{2}\langle x - D\alpha, D\alpha\rangle + \frac{1}{2}\langle x - \tilde{D}\tilde{\alpha}, \tilde{D}\tilde{\alpha}\rangle + \frac{3}{8}\frac{\varepsilon}{\lambda} \\
&\leq \frac{1}{2}\|x\|_2^2 - \frac{1}{2}\langle x, D\alpha\rangle - \frac{1}{2}\langle x, D\tilde{\alpha}\rangle + \frac{1}{4}\|D\alpha\|_2^2 + \frac{1}{4}\langle D\alpha, D\tilde{\alpha}\rangle + \frac{1}{2}\langle x, D\alpha\rangle - \frac{1}{2}\|D\alpha\|_2^2 + \frac{1}{2}\langle x, D\tilde{\alpha}\rangle - \frac{1}{2}\|D\alpha\|_2^2 \\
&\quad + \left(\frac{3}{8} + \frac{1}{4} + \frac{5}{8}\right)\frac{\varepsilon}{\lambda}
\end{aligned}
$$

which simplifies to

$$
\begin{aligned}
&\frac{1}{2}\|x\|_2^2 - \frac{3}{4}\|D\alpha\|_2^2 - \frac{1}{2}\langle x, D\alpha\rangle - \frac{1}{2}\langle x, D\tilde{\alpha}\rangle + \frac{1}{4}\langle D\alpha, D\tilde{\alpha}\rangle + \frac{1}{2}\langle x, D\alpha\rangle + \frac{1}{2}\langle x, D\tilde{\alpha}\rangle + \frac{5}{4}\frac{\varepsilon}{\lambda} \\
&= \frac{1}{2}\|x\|_2^2 - \frac{3}{4}\|D\alpha\|_2^2 + \frac{1}{4}\langle D\alpha, D\tilde{\alpha}\rangle + \frac{5}{4}\frac{\varepsilon}{\lambda}.
\end{aligned}
$$

Now, taking the (expanded) LHS of (2) and the newly derived upper bound of the RHS of (2) yields the inequality:

$$
\begin{aligned}
&\frac{1}{2}\|x\|_2^2 - \langle x, D\alpha\rangle + \frac{1}{2}\|D\alpha\|_2^2 + \lambda\|\alpha\|_1 \\
&\leq \frac{1}{2}\|x\|_2^2 - \frac{3}{4}\|D\alpha\|_2^2 + \frac{1}{4}\langle D\alpha, D\tilde{\alpha}\rangle + \frac{5}{4}\frac{\varepsilon}{\lambda}.
\end{aligned}
$$

which implies that

$$
\begin{aligned}
&-\langle x, D\alpha\rangle + \frac{1}{2}\|D\alpha\|_2^2 + \lambda\|\alpha\|_1 \\
&\leq -\frac{3}{4}\|D\alpha\|_2^2 + \frac{1}{4}\langle D\alpha, D\tilde{\alpha}\rangle + \frac{5}{4}\frac{\varepsilon}{\lambda}.
\end{aligned}
$$

Replacing $\lambda\|\alpha\|_1$ with $\langle x - D\alpha, D\alpha\rangle$ yields:

$$
\begin{aligned}
&-\langle x, D\alpha\rangle + \frac{1}{2}\|D\alpha\|_2^2 + \langle x, D\alpha\rangle - \|D\alpha\|_2^2 \\
&\leq -\frac{3}{4}\|D\alpha\|_2^2 + \frac{1}{4}\langle D\alpha, D\tilde{\alpha}\rangle + \frac{5}{4}\frac{\varepsilon}{\lambda},
\end{aligned}
$$

implying that

$$
\frac{1}{4}\|D\alpha\|_2^2 \leq \frac{1}{4}\langle D\alpha, D\tilde{\alpha}\rangle + \frac{5}{4}\frac{\varepsilon}{\lambda}.
$$

Hence,

$$\|D\alpha\|_2^2 \leq \langle D\alpha, D\tilde{\alpha}\rangle + \frac{5\varepsilon}{\lambda}.$$

Now, note that

$$\begin{aligned}\|D\alpha - D\tilde{\alpha}\|_2^2 &= \|D\alpha\|_2^2 + \|D\tilde{\alpha}\|_2^2 - 2\langle D\alpha, D\tilde{\alpha}\rangle \\ &\leq \|D\alpha\|_2^2 + \|D\tilde{\alpha}\|_2^2 - 2\|D\alpha\|_2^2 + 10\frac{\varepsilon}{\lambda} \\ &\leq \|D\alpha\|_2^2 + \|D\alpha\|_2^2 - 2\|D\alpha\|_2^2 + 13\frac{\varepsilon}{\lambda} \\ &= 13\frac{\varepsilon}{\lambda}.\end{aligned}$$

■

A.3 The Sparsity Lemma

We now prove that the solution to the perturbed problem is sparse for sufficiently small ε.

Lemma 15 (Preservation of Sparsity) *Under Assumptions (A1)-(A3), if*

$$\tau \geq \varepsilon\left(1 + \frac{1}{2\lambda}\right) + \sqrt{\frac{13\varepsilon}{\lambda}},$$

then $\tilde{\alpha}_i = 0$ *for all* $i \in \mathcal{I}$.

Proof Let $i \in \mathcal{I}$ be arbitrary. To prove that $\tilde{\alpha}_i = 0$, it is sufficient to show that

$$\left|\langle \tilde{D}_i, x - \tilde{D}\tilde{\alpha}\rangle\right| < \lambda,$$

since $\tilde{\alpha}_i$ is hence zero.

First, note that

$$\begin{aligned}\left|\langle \tilde{D}_i, x - \tilde{D}\tilde{\alpha}\rangle\right| &= \left|\langle D_i + \tilde{D}_i - D_i, x - \tilde{D}\tilde{\alpha}\rangle\right| \\ &\leq \left|\langle D_i, x - \tilde{D}\tilde{\alpha}\rangle\right| + \|\tilde{D}_i - D_i\|_2\|x - \tilde{D}\tilde{\alpha}\|_2 \\ &\leq \left|\langle D_i, x - \tilde{D}\tilde{\alpha}\rangle\right| + \varepsilon \qquad \text{(since } \|x\|_2 \leq 1\text{)}\end{aligned}$$

and

$$\begin{aligned}\left|\langle D_i, x - \tilde{D}\tilde{\alpha}\rangle\right| &= \left|\langle D_i, x - (D + \tilde{D} - D)\tilde{\alpha}\rangle\right| \\ &\leq |\langle D_i, x - D\tilde{\alpha}\rangle| + \left|\langle D_i, (\tilde{D} - D)\tilde{\alpha}\rangle\right| \\ &\leq |\langle D_i, x - D\tilde{\alpha}\rangle| + \|D_i\|_2\|\tilde{D} - D\|_2\|\tilde{\alpha}\|_2 \\ &\leq |\langle D_i, x - D\tilde{\alpha}\rangle| + \frac{\varepsilon}{2\lambda}.\end{aligned}$$

Hence,

$$\left|\langle \tilde{D}_i, x - \tilde{D}\tilde{\alpha}\rangle\right| \leq |\langle D_i, x - D\tilde{\alpha}\rangle| + \varepsilon\left(1 + \frac{1}{2\lambda}\right),$$

and so it is sufficient to show that

$$|\langle D_i, x - D\tilde{\alpha}\rangle| < \lambda - \varepsilon\left(1 + \frac{1}{2\lambda}\right).$$

Now,

$$\begin{aligned}
|\langle D_i, x - D\tilde{\alpha}\rangle| &= |\langle D_i, x - D\tilde{\alpha} + D\alpha - D\alpha\rangle| \\
&\leq |\langle D_i, x - D\alpha\rangle| + |\langle D_i, D\alpha - D\tilde{\alpha}\rangle| \\
&< \lambda - \tau + \|D_i\|_2 \|D\alpha - D\tilde{\alpha}\|_2 \\
&\leq \lambda - \tau + \sqrt{\frac{13\varepsilon}{\lambda}}, \qquad (3)
\end{aligned}$$

where (3) is due to Lemma 14. Consequently, it is sufficient if τ is chosen to satisfy

$$\lambda - \tau + \sqrt{\frac{13\varepsilon}{\lambda}} \leq \lambda - \varepsilon\left(1 + \frac{1}{2\lambda}\right),$$

yielding:

$$\tau \geq \varepsilon\left(1 + \frac{1}{2\lambda}\right) + \sqrt{\frac{13\varepsilon}{\lambda}}.$$

■

A.4 Proof of the Sparse Coding Stability Theorem

Proof (of Theorem 4) Recall that $\varphi_D(x)$ is the unique optimal solution to the problem

$$\min_{z \in \mathbb{R}^k} \frac{1}{2}\|x - Dz\|_2^2 + \lambda\|z\|_1.$$

If not for ℓ_1 penalty, in standard form, the quadratic program is

$$\min_{z \in \mathbb{R}^k} z^T D^T D z - z^T(2Dx) + \lambda\|x\|_1$$

Similarly, let $\tilde{Q}(\cdot)$ be the objective using $\tilde{D}$ instead of D. Denoting $\bar{z} := \begin{pmatrix} z \\ z^+ \\ z^- \end{pmatrix}$ with $z^+, z^- \in \mathbb{R}^k$, an equivalent formulation is

$$\begin{array}{ll}
\underset{\bar{z} \in \mathbb{R}^{3k}}{\text{minimize}} & Q(\bar{z}) := \frac{1}{2}\bar{z}^T \begin{pmatrix} D^T D & \mathbf{0}_{k\times 2k} \\ \mathbf{0}_{2k\times k} & \mathbf{0}_{2k\times 2k} \end{pmatrix} \bar{z} - \frac{1}{2}\bar{z}^T \left(\begin{pmatrix} 2D^T \\ \mathbf{0}_{2k\times d} \end{pmatrix} x \right) + \lambda(\mathbf{0}_k^T \mathbf{1}_{2k}^T)\bar{z} \\
\text{subject to} & z^+ \geq \mathbf{0}_k \qquad z^- \geq \mathbf{0}_k \qquad z - z^+ + z^- = \mathbf{0}_k.
\end{array}$$

For optimal solutions $\bar{z}_* := \begin{pmatrix} z_* \\ z_*^+ \\ z_*^- \end{pmatrix}$ and $\bar{t}_* := \begin{pmatrix} t_* \\ t_*^+ \\ t_*^- \end{pmatrix}$ of Q and $\tilde{Q}$ respectively, from Daniel (1973), we have

$$(\bar{u} - \bar{z}_*)^T \nabla Q(\bar{z}_*) \geq 0 \qquad (4)$$

$$(\bar{u} - \bar{t}_*)^T \nabla \tilde{Q}(\bar{t}_*) \geq 0 \qquad (5)$$

for all feasible $\bar{u} \in \mathbb{R}^{3k}$. Setting $\bar{u}$ to $\bar{t}_*$ in (4) and $\bar{u}$ to $\bar{z}_*$ in (5) and adding (5) and (4) yields

$$(\bar{t}_* - \bar{z}_*)^T (\nabla Q(\bar{z}_*) - \nabla \tilde{Q}(\bar{t}_*)) \geq 0,$$

which is equivalent to

$$(\bar{t}_* - \bar{z}_*)^T (\nabla \tilde{Q}(\bar{t}_*) - \nabla \tilde{Q}(\bar{z}_*)) \leq (\bar{t}_* - \bar{z}_*)^T (\nabla Q(\bar{z}_*) - \nabla \tilde{Q}(\bar{z}_*)) \tag{6}$$

Here,

$$\nabla Q(z) = \frac{1}{2} \begin{pmatrix} D^T D & \mathbf{0}_{k \times 2k} \\ \mathbf{0}_{2k \times k} & \mathbf{0}_{2k \times 2k} \end{pmatrix} z - \frac{1}{2} \begin{pmatrix} 2D^T \\ \mathbf{0}_{2k \times d} \end{pmatrix} x + \lambda \begin{pmatrix} \mathbf{0}_k \\ \mathbf{1}_{2k} \end{pmatrix}.$$

After plugging in the expansions of ∇Q and $\nabla \tilde{Q}$ and incurring cancellations from the zeros, (6) becomes

$$\begin{aligned} (t_* - z_*)^T \tilde{D}^T \tilde{D} (t_* - z_*) &\leq (t_* - z_*)^T \Big((D^T D - \tilde{D}^T \tilde{D}) z_* + 2(\tilde{D} - D)^T x \Big) && (7) \\ &\leq (t_* - z_*)^T (D^T D - \tilde{D}^T \tilde{D}) z_* + 2 \|t_* - z_*\|_2 \|(\tilde{D} - D)^T x\|_2 \\ &\leq (t_* - z_*)^T (D^T D - \tilde{D}^T \tilde{D}) z_* + \|t_* - z_*\|_2 (2\varepsilon) \\ & && (8) \end{aligned}$$

Let us gain a handle on the first term. Below, we will use an operator which we dub the *s-restricted* 2*-norm*: for a dictionary $A \in (B_{\mathbb{R}^d})^k$, the s-restricted 2-norm of A is defined as $\|A\|_{2,s} := \sup_{\{t \in \mathbb{R}^n : \|t\|=1, |\operatorname{supp}(t)| \leq s\}} \|At\|_2$. Now, note that $\tilde{D} = D + E$ for some E satisfying $\|E\|_2 \leq \varepsilon$. Hence,

$$\begin{aligned} &(t_* - z_*)^T (D^T D - \tilde{D}^T \tilde{D}) z_* \\ &= \big| (t_* - z_*)^T (E^T D + D^T E + E^T E) z_* \big| \\ &\leq \big| (t_* - z_*)^T E^T D z_* \big| + \big| (t_* - z_*)^T D^T E z_* \big| + \big| (t_* - z_*)^T E^T E z_* \big| \\ &\leq \|E(t_* - z_*)\|_2 \|D z_*\|_2 + \|D(t_* - z_*)\|_2 \|E z_*\|_2 + \|E(t_* - z_*)\|_2 \|E z_*\|_2 \\ &\leq \|t_* - z_*\|_2 \Big(\|E\|_2 \|D\|_{2,s} \|z_*\|_2 + \|D\|_{2,s} \|E\|_2 \|z_*\|_2 + \|E\|_2^2 \|z_*\|_2 \Big) \\ &\leq \|t_* - z_*\|_2 \left(\frac{\varepsilon \sqrt{s}}{2\lambda} + \frac{\varepsilon \sqrt{s}}{2\lambda} + \frac{\varepsilon^2}{2\lambda} \right) \\ &\leq \|t_* - z_*\|_2 \frac{3}{2} \frac{\varepsilon \sqrt{s}}{\lambda}, \end{aligned}$$

where the penultimate step follows because

1. if $\|z_*\|_0 \leq s$, then Lemma 16 (stated after this proof) implies that $\|D z_*\|_2 \leq \sqrt{s} \|z_*\|_2$ (and $\|z_*\|_2 \leq \|z_*\|_1 \leq \frac{1}{2\lambda}$); and
2. Lemma 15 implies that $\|t_* - z_*\|_0 \leq s$.

Combining this result with the fact that $\tilde{D}$ has s-incoherence lower bounded by μ implies the desired result:

$$\|t_* - z_*\|_2 \leq \frac{3}{2} \frac{\varepsilon \sqrt{s}}{\lambda \mu}.$$

■

Lemma 16 *If $D \in (B_{\mathbb{R}^d})^k$, then $\|D\|_{2,s} \leq \sqrt{s}$.*

Proof Define D_Λ as the submatrix of D that selects the columns indexed by Λ. Similarly, for $t \in \mathbb{R}^k$ define the coordinate projection t_Λ of t.

$$\begin{aligned}
&\sup_{\{t:\|t\|=1,|\operatorname{supp}(t)|\le s\}} \|Dt\|_2 \\
&= \max_{\{\Lambda\subseteq[k]:|\Lambda|\le s\}} \sup_{\{t:\|t\|=1,\operatorname{supp}(t)\subseteq\Lambda\}} \|D_\Lambda t_\Lambda\|_2 \\
&= \max_{\{\Lambda\subseteq[k]:|\Lambda|\le s\}} \sup_{\{t:\|t\|=1,\operatorname{supp}(t)\subseteq\Lambda\}} \left\|\sum_{\omega\in\Lambda} t_\omega D_\omega\right\|_2 \\
&\le \max_{\{\Lambda\subseteq[k]:|\Lambda|\le s\}} \sup_{\{t:\|t\|=1,\operatorname{supp}(t)\subseteq\Lambda\}} \sum_{\omega\in\Lambda} |t_\omega|\|D_\omega\|_2 \\
&\le \max_{\{\Lambda\subseteq[k]:|\Lambda|\le s\}} \sup_{\{t:\|t\|=1,\operatorname{supp}(t)\subseteq\Lambda\}} \sum_{\omega\in\Lambda} |t_\omega| \\
&\le \max_{\{\Lambda\subseteq[k]:|\Lambda|\le s\}} \sup_{\{t:\|t\|=1,\operatorname{supp}(t)\subseteq\Lambda\}} \|t_\Lambda\|_1 \\
&\le \max_{\{\Lambda\subseteq[k]:|\Lambda|\le s\}} \sup_{\{t:\|t\|=1,\operatorname{supp}(t)\subseteq\Lambda\}} \sqrt{s}\|t_\Lambda\|_2 \\
&= \sqrt{s}.
\end{aligned}$$

■

B Proof of Symmetrization by Ghost Sample Lemma

Proof (of Lemma 6) Replace $\mathcal{F}(\sigma_n)$ from the notation of Mendelson & Philips (2004) with $\mathcal{F}(\mathbf{z},\mathbf{x}'')$. A modified one-sided version of Mendelson & Philips (2004, Lemma 2.2) that uses the more favorable Chebyshev-Cantelli inequality implies that, for every $t > 0$:

$$\begin{aligned}
&\left(1 - \frac{4\sup_{f\in\mathcal{F}}\operatorname{Var}(l(\cdot,f))}{4\sup_{f\in\mathcal{F}}\operatorname{Var}(l(\cdot,f)) + mt^2}\right)\Pr_{\mathbf{z}\,\mathbf{x}''}\left\{\exists f\in\mathcal{F}(\mathbf{z},\mathbf{x}''),\ (P-P_{\mathbf{z}})l(\cdot,f)\ge t\right\} \\
&\qquad\qquad \le \Pr_{\mathbf{z}\,\mathbf{z}'\mathbf{x}''}\left\{\exists f\in\mathcal{F}(\mathbf{z},\mathbf{x}''),\ (P_{\mathbf{z}'}-P_{\mathbf{z}})l(\cdot,f)\ge \frac{t}{2}\right\}.
\end{aligned}$$

As the losses lie in $[0,b]$ by assumption, it follows that $\sup_{f\in\mathcal{F}}\operatorname{Var}(l(\cdot,f)) \le \frac{b^2}{4}$. The lemma follows since the left hand factor of the LHS of the above inequality is at least $\frac{1}{2}$ whenever $m \ge \left(\frac{b}{t}\right)^2$. ■

C Proof of Learning Bound

Proof (of Theorem 5) Proposition 9 and Lemmas 10 and 11 imply that

$$\begin{aligned}
&\Pr_{\mathbf{z}}\left\{\begin{array}{r} \exists f\in\mathcal{F}_\mu,\ \left[\operatorname{margin}_s(D,\mathbf{x}) > \iota\right] \\ \textbf{and}\ \ ((P-P_{\mathbf{z}})l(\cdot,f) > t) \end{array}\right\} \\
&\qquad \le 2\left(\left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)) + \delta\right).
\end{aligned}$$

Equivalently,

$$
\begin{aligned}
&\Pr_{\mathbf{z}}\left\{\begin{array}{c} \exists f \in \mathcal{F}_\mu,\ \big[\mathrm{margin}_s(D,\mathbf{x}) > \iota\big] \\ \textbf{and}\ \left((P-P_{\mathbf{z}})l(\cdot,f) > 2\left(\varpi + 2L\beta + \frac{b\eta(m,d,k,\varepsilon,\delta)}{m}\right)\right) \end{array}\right\} \\
&\qquad \leq 2\left(\left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)) + \delta\right).
\end{aligned}
$$

Now, expand β and η and replace δ with $\delta/4$:

$$
\begin{aligned}
&\Pr_{\mathbf{z}}\left\{\begin{array}{l} \exists f \in \mathcal{F}_\mu,\ \big[\mathrm{margin}_s(D,\mathbf{x}) > \iota\big]\ \textbf{and} \\ \qquad (P-P_{\mathbf{z}})l(\cdot,f) > 2\left(\varpi + 2L\varepsilon\frac{1}{2\lambda}\left(1+\frac{3r\sqrt{s}}{\mu}\right) + \frac{b(dk\log\frac{1944}{\mathrm{margin}_s^2(D,\mathbf{x})\cdot\lambda} + \log(2m+1) + \log\frac{4}{\delta})}{m}\right) \end{array}\right\} \\
&\leq 2\left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)) + \frac{\delta}{2}.
\end{aligned}
$$

Choosing $\frac{\delta}{4} = \left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k} \exp(-m\varpi^2/(2b^2))$ yields

$$
\begin{aligned}
&\Pr_{\mathbf{z}}\left\{\begin{array}{l} \exists f \in \mathcal{F}_\mu,\ \big[\mathrm{margin}_s(D,\mathbf{x}) > \iota\big]\ \textbf{and} \\ \qquad (P-P_{\mathbf{z}})l(\cdot,f) > 2\left(\begin{array}{l} \varpi + L\varepsilon\frac{1}{\lambda}\left(1+\frac{3r\sqrt{s}}{\mu}\right) + \\ \frac{b(dk\log\frac{1944}{\mathrm{margin}_s^2(D,\mathbf{x})\cdot\lambda} + \log(2m+1) + (d+1)k\log\frac{\varepsilon}{8(r/2)^{1/(d+1)}} + \frac{m\varpi^2}{b^2})}{m} \end{array}\right) \end{array}\right\} \\
&\leq 4\cdot\left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)),
\end{aligned}
$$

which is equivalent to

$$
\begin{aligned}
&\Pr_{\mathbf{z}}\left\{\begin{array}{l} \exists f \in \mathcal{F}_\mu,\ \big[\mathrm{margin}_s(D,\mathbf{x}) > \iota\big]\ \textbf{and} \\ \qquad (P-P_{\mathbf{z}})l(\cdot,f) > 2\left(\begin{array}{l} \varpi + L\varepsilon\frac{1}{\lambda}\left(1+\frac{3r\sqrt{s}}{\mu}\right) + \\ \frac{b(dk\log\frac{1944}{\mathrm{margin}_s^2(D,\mathbf{x})\cdot\lambda} - (d+1)k\log\frac{8}{\varepsilon} + k\log\frac{2}{r} + \log(2m+1) + \frac{m\varpi^2}{b^2})}{m} \end{array}\right) \end{array}\right\} \\
&\leq 4\cdot\left(\frac{8(r/2)^{1/(d+1)}}{\varepsilon}\right)^{(d+1)k} \exp(-m\varpi^2/(2b^2)),
\end{aligned}
$$

Let δ (a new variable, not related to the previous incarnation of δ) be equal to the upper bound, and solve for ϖ, yielding:

$$
\varpi = b\sqrt{\frac{2((d+1)k\log\frac{8}{\varepsilon} + k\log\frac{r}{2} + \log\frac{4}{\delta})}{m}}
$$

and hence

$$
\begin{aligned}
&\Pr_{\mathbf{z}}\left\{\begin{array}{l} \exists f = (D,w) \in \mathcal{F}_\mu,\ \big[\mathrm{margin}_s(D,\mathbf{x}) > \iota\big]\ \textbf{and} \\ \qquad (P-P_{\mathbf{z}})l(\cdot,f) > 2\left(\begin{array}{l} b\sqrt{\frac{2((d+1)k\log\frac{8}{\varepsilon} + k\log\frac{r}{2} + \log\frac{4}{\delta})}{m}} + L\varepsilon\frac{1}{\lambda}\left(1+\frac{3r\sqrt{s}}{\mu}\right) + \\ \frac{b(dk\log\frac{1944}{\mathrm{margin}_s^2(D,\mathbf{x})\cdot\lambda} + \log(2m+1) + \log\frac{4}{\delta})}{m} \end{array}\right) \end{array}\right\} \\
&\leq \delta,
\end{aligned}
$$

If we set $\varepsilon = \frac{1}{m}$, then provided that $m > \frac{243}{\text{margin}_s^2(D,\mathbf{x})\cdot\lambda}$:

$$\Pr_{\mathbf{z}}\left\{\begin{array}{l} \exists f \in \mathcal{F}_\mu,\ \left[\text{margin}_s(D,\mathbf{x}) > \iota\right] \textbf{ and} \\ \qquad (P - P_{\mathbf{z}})l(\cdot, f) > 2\left(\begin{array}{l} b\sqrt{\frac{2((d+1)k\log(8m)+k\log\frac{r}{2}+\log\frac{4}{\delta})}{m}} + \frac{L}{m}\left(\frac{1}{\lambda}(1+\frac{3r\sqrt{s}}{\mu})\right) + \\ \frac{b}{m}\left(dk\log\frac{1944}{\text{margin}_s^2(D,\mathbf{x})\cdot\lambda} + \log(2m+1) + \log\frac{4}{\delta}\right) \end{array}\right) \end{array}\right\}$$
$$\leq \delta.$$

It remains to distribute a prior across the bounds for each choice of s and μ. To each choice of $s \in [k]$ assign prior probability $\frac{1}{k}$. To each choice of $i \in \mathbb{N} \cup \{0\}$ for $2^{-i} \leq \mu$ assign prior probability $(i+1)^{-2}$. For a given choice of $s \in [k]$ and $2^{-i} \leq \mu$ we use $\delta(s,i) := \frac{6}{\pi^2}\frac{1}{(i+1)^2}\frac{1}{k}\delta$ (since $\sum_{i=1}^{\infty}\frac{1}{i^2} = \frac{\pi^2}{6}$). Then, provided that

$$m > \frac{243}{\text{margin}_s(D,\mathbf{x})^2\lambda},$$

the generalization error $(P - P_{\mathbf{z}})l(\cdot, f)$ is bounded by:

$$\begin{aligned} &2b\sqrt{\frac{2\left((d+1)k\log(8m) + k\log\frac{r}{2} + \log\frac{2\pi^2\left(\log_2\frac{4}{\mu_s(D)}\right)^2 k}{3\delta}\right)}{m}} \\ &+ \frac{2b}{m}\left(dk\log\frac{1944}{\text{margin}_s^2(D,\mathbf{x})\cdot\lambda} + \log(2m+1) + \log\frac{2\pi^2\left(\log_2\frac{4}{\mu_s(D)}\right)^2 k}{3\delta}\right) \\ &+ \frac{2L}{m}\left(\frac{1}{\lambda}(1 + \frac{6r\sqrt{s}}{\mu_s(D)})\right). \end{aligned}$$

■

References

Asif, M. Salman and Romberg, Justin. On the LASSO and Dantzig selector equivalence. In *Annual Conference on Information Sciences and Systems (CISS)*, pp. 1–6. IEEE, 2010.

Daniel, James W. Stability of the solution of definite quadratic programs. *Mathematical Programming*, 5 (1):41–53, 1973.

Mendelson, Shahar and Philips, Petra. On the importance of small coordinate projections. *Journal of Machine Learning Research*, 5:219–238, 2004.

Osborne, Michael R., Presnell, Brett, and Turlach, Berwin A. On the lasso and its dual. *Journal of Computational and Graphical Statistics*, pp. 319–337, 2000.