Supplemental Material to *Conditional Sparse Coding and Grouped Multivariate Regression*

Min Xu
minx [AT] cs.cmu.edu
John Lafferty
lafferty [AT] galton.uchicago.edu

February 15, 2013

1 Proof of Lemma 5.1:

Recall that $R(B) = \mathbb{E}\|Y - BX\|_2^2$ and $\widehat{R}(B) = \frac{1}{n}\|\mathbb{Y} - B\mathbb{X}\|_F^2$ where $\mathbb{X}$ is $p \times n$ matrix and Y is $q \times n$ matrix.

$$
\begin{aligned}
R(B) &= \mathbb{E}\|Y - BX\|_2^2 && (1.1)\\
&= \operatorname{tr}\left\{ \begin{pmatrix} -I_q \\ B \end{pmatrix}^T \mathbb{E}\big[(Y,X)(Y,X)^T\big] \begin{pmatrix} -I_q \\ B \end{pmatrix} \right\} && (1.2)\\
&= \operatorname{tr}\left\{ \begin{pmatrix} -I_q \\ B \end{pmatrix}^T \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{YX}^T & \Sigma_{XX} \end{pmatrix} \begin{pmatrix} -I_q \\ B \end{pmatrix} \right\} && (1.3)\\
&\equiv \operatorname{tr}\left\{C^T \Sigma C\right\}. && (1.4)
\end{aligned}
$$

Similarly, the sample risk can be reexpressed as

$$
\begin{aligned}
\widehat{R}(B) &= \frac{1}{n}\|\mathbb{Y} - \mathbb{X}B\|_F^2 && (1.5)\\
&= \operatorname{tr}\left\{ \begin{pmatrix} -I_q \\ B \end{pmatrix}^T \frac{1}{n}(\mathbb{Y},\mathbb{X})^T(\mathbb{Y},\mathbb{X}) \begin{pmatrix} -I_q \\ B \end{pmatrix} \right\} && (1.6)\\
&= \operatorname{tr}\left\{ \begin{pmatrix} -I_q \\ B \end{pmatrix}^T \begin{pmatrix} \widehat{\Sigma}_{YY} & \widehat{\Sigma}_{YX} \\ \widehat{\Sigma}_{YX}^T & \widehat{\Sigma}_{XX} \end{pmatrix} \begin{pmatrix} -I_q \\ B \end{pmatrix} \right\} && (1.7)\\
&\equiv \operatorname{tr}\left\{C^T \widehat{\Sigma}_n C\right\}. && (1.8)
\end{aligned}
$$

Each of these risks has an "uncontrollable" contribution that does not depend on B. Specifically,

$$
\begin{aligned}
R_u &= \operatorname{tr}\{\Sigma_{YY}\} && (1.9)\\
\widehat{R}_u &= \operatorname{tr}\{\widehat{\Sigma}_{YY}\}. && (1.10)
\end{aligned}
$$

We can express the remaining "controllable" risk as

$$R_c(B) = \mathrm{tr}\left\{ \begin{pmatrix} -2I_q \\ B \end{pmatrix}^T \begin{pmatrix} \Sigma_{YY} & \Sigma_{YX} \\ \Sigma_{YX}^T & \Sigma_{XX} \end{pmatrix} \begin{pmatrix} 0_q \\ B \end{pmatrix} \right\} \tag{1.11}$$

$$\equiv \mathrm{tr}\left\{ D^T \Sigma C \right\} \tag{1.12}$$

$$\widehat{R}_c(B) = \mathrm{tr}\left\{ \begin{pmatrix} -2I_q \\ B \end{pmatrix}^T \begin{pmatrix} \widehat{\Sigma}_{YY} & \widehat{\Sigma}_{YX} \\ \widehat{\Sigma}_{YX}^T & \widehat{\Sigma}_{XX} \end{pmatrix} \begin{pmatrix} 0_q \\ B \end{pmatrix} \right\} \tag{1.13}$$

$$\equiv \mathrm{tr}\left\{ D^T \widehat{\Sigma} C \right\}. \tag{1.14}$$

Thus, we can write

$$R(B) - \widehat{R}(B) = R_u - \widehat{R}_u + R_c(B) - \widehat{R}_c(B) \tag{1.15}$$

$$= R_u - \widehat{R}_u + \mathrm{tr}\left\{ D^T \left(\Sigma - \widehat{\Sigma}_n \right) C \right\}. \tag{1.16}$$

Now, using Hölder's inequality for matrices,

$$\mathrm{tr}AB \leq \|A\|_r \|B\|_{r'} \tag{1.17}$$

where $1/r + 1/r' = 1$, we have with $\|\cdot\|_2$ denoting the spectral norm,

$$R_c(B) - R_c(\widehat{B}) = \mathrm{tr}\left\{ D^T \left(\Sigma - \widehat{\Sigma}_n \right) C \right\} \tag{1.18}$$

$$\leq \|D(\Sigma - \widehat{\Sigma}_n)\|_2 \, \|C\|_* \tag{1.19}$$

$$= \|D(\Sigma - \widehat{\Sigma}_n)\|_2 \, \|B\|_* \tag{1.20}$$

$$\leq \|D\|_2 \, \|\Sigma - \widehat{\Sigma}_n\|_2 \, \|B\|_* \tag{1.21}$$

$$\leq c \max(2, \|B\|_2) \, \|\Sigma - \widehat{\Sigma}_n\|_2 \, \|B\|_* \tag{1.22}$$

$$\leq c \|B\|_*^2 \, \|\Sigma - \widehat{\Sigma}_n\|_2 \tag{1.23}$$

$$\tag{1.24}$$

In (1.22) and (1.23), c denotes a generic constant.

A concentration of measure result from Vershynin (2011) states that with probability at least $1 - \exp(-c_1 p)$, we have that, for subgaussian random variables, $\|\Sigma - \widehat{\Sigma}_n\|_2 \leq c_2 \sqrt{\frac{p+q}{n}}$ where c_1 and c_2 are constants depending on $\|\Sigma\|_2$. Combining this with equation 1.24 and a union bound over the G groups gives us the lemma immediately.

The reason we use the controllable error, and the matrix $C = \begin{pmatrix} 0_q \\ B \end{pmatrix}$ rather than $\begin{pmatrix} -I_q \\ B \end{pmatrix}$, is that the trace norm of the latter is of the order $(p+q)\|B\|_*$ rather than $\|B\|_*$.

2 Proof of Theorem 5.1

Following the usual argument, if $\widehat{B} = \arg\min_{B \in \mathcal{B}} \|\mathbb{Y} - B\mathbb{X}\|_F$ where $\mathcal{B} = \{B : \|B\|_* \leq L\}$.

We have that

$$R(\widehat{B}) - \inf_{B\in\mathcal{B}} R(B) = R(\widehat{B}) - \widehat{R}(\widehat{B}) - (R(B_*) - \widehat{R}(B_*)) + (\widehat{R}(\widehat{B}) - \widehat{R}(B_*)) \tag{2.1}$$

$$\leq R(\widehat{B}) - \widehat{R}(\widehat{B}) - (R(B_*) - \widehat{R}(B_*)) \tag{2.2}$$

$$= R_c(\widehat{B}) - \widehat{R}_c(\widehat{B}) - (R_c(B_*) - \widehat{R}_c(B_*)) \tag{2.3}$$

$$\leq 2 \sup_{B\in\mathcal{B}} \left\{ R_c(B) - \widehat{R}_c(B) \right\} \tag{2.4}$$

$$\tag{2.5}$$

The theorem then follows from Lemma 5.1 and a union bound over the G groups.

3 Proof of Lemma 5.2

Suppose the dictionary $D_1, ..., D_K$ is fixed and let $\|D_k\|_* \leq 1$. Then we define

$$R(\alpha) = \mathbb{E}\|Y - \sum_{k=1}^{K} \alpha_k D_k X\|_2^2$$

and the corresponding

$$\widehat{R}(\alpha) = \frac{1}{n}\|\mathbb{Y} - \sum_{k=1}^{K} \alpha_k D_k \mathbb{X}\|_F^2$$

We get thus that

$$R(\alpha) - \widehat{R}(\alpha) =$$

$$\mathrm{tr}\Sigma_{YY} - \mathrm{tr}\widehat{\Sigma}_{YY} + \sum_{k=1}^{K} \alpha_k \left(\mathrm{tr}(D_k \Sigma_{XY}) - \mathrm{tr}(D_k \widehat{\Sigma}_{XY}) \right) + \sum_{k=1}^{K}\sum_{k'=1}^{K} \alpha_k \alpha_{k'} \left(\mathrm{tr}(D_k \Sigma_{XX} D_{k'}^{\mathsf{T}}) - \mathrm{tr}(D_k \widehat{\Sigma}_{XX} D_{k'}^{\mathsf{T}}) \right)$$

We first consider the term

$$\mathrm{tr}(D_k \Sigma_{XY}) - \mathrm{tr}(D_k \widehat{\Sigma}_{XY}) = \mathbb{E}[Y^{\mathsf{T}} D_k X] - \frac{1}{n}\sum_{i=1}^{n} Y_i^{\mathsf{T}} D_k X_i$$

$$= \sum_{l=1}^{r} \sigma_l \left(\mathbb{E}[Y^{\mathsf{T}} u v^{\mathsf{T}} X] - \frac{1}{n}\sum_{i=1}^{n} Y_i^{\mathsf{T}} u v^{\mathsf{T}} X_i \right)$$

where $\sum_{l=1}^{r} \sigma_l u v^{\mathsf{T}} = D_k$ is the SVD of D_k. Since each $Y^{\mathsf{T}} u$ and $v^{\mathsf{T}} X$ is Gaussin random variable, we use well-known concentration of measure result and get that with probability at least $1 - \frac{1}{n}$, we have $\mathrm{tr}(D_k \Sigma_{XY}) - \mathrm{tr}(D_k \widehat{\Sigma}_{XY}) \leq \|D_k\|_* C \sqrt{\frac{\log(pn)}{n}}$ where C is some constant depending only on variance of X, Y.

We can use identical technique to bound the other terms $\mathrm{tr}(D_k \Sigma_{XX} D_{k'}^{\mathsf{T}}) - \mathrm{tr}(D_k \widehat{\Sigma}_{XX} D_{k'}^{\mathsf{T}})$ and get that, with probability at least $1 - \frac{1}{n}$,

$$R(\alpha) - \widehat{R}(\alpha) \leq R_u + \|\alpha\|_1 C \sqrt{\frac{\log(Kpn)}{n}} + \|\alpha\|_1^2 C \sqrt{\frac{\log(Kpn)}{n}}$$

The lemma then follows from a union bound over all groups.

4 Proof of Theorem 5.3

Following the usual argument, let $\widehat{\alpha} = \arg\min_\alpha \widehat{R}(\alpha) + \lambda\|\alpha\|_1$, and let $\alpha^* = \arg\min_{\alpha:\|\alpha\|_1 \le L} R(\alpha)$

Then we have, with probability at least $1 - \frac{1}{n}$,

$$\begin{aligned} R(\widehat{\alpha}) - R(\alpha^*) &= R(\widehat{\alpha}) - \widehat{R}(\widehat{\alpha}) - (R(\alpha^*) - \widehat{R}(\alpha^*)) + (\widehat{R}(\widehat{\alpha}) - \widehat{R}(\alpha^*)) \\ &\le R(\widehat{\alpha}) - \widehat{R}(\widehat{\alpha}) - (R(\alpha^*) - \widehat{R}(\alpha^*)) + \lambda\|\alpha^*\|_1 \\ &\le C\|\alpha\|_1^2\sqrt{\frac{\log(Knp)}{n}} + CL^2\sqrt{\frac{\log(Knp)}{n}} + CL\sqrt{\frac{\log K}{n}} \end{aligned}$$

5 Proof of Theorem 5.2

We first re-write the excess risk as

$$\begin{aligned} &R(D^{\text{learn}}, \alpha_\lambda^{\text{learn}(g)}) - R(B^{*(g)}) \\ &\quad = R(D^{\text{learn}}, \alpha_\lambda^{\text{learn}(g)}) - \widehat{R}(D^{\text{learn}}, \alpha_\lambda^{\text{learn}(g)}) && (5.1) \\ &\qquad + \widehat{R}(D^{\text{learn}}, \alpha_\lambda^{\text{learn}(g)}) - \widehat{R}(D^{\text{init}}, \alpha_{\text{oracle}}^{\text{init}(g)}) && (5.2) \\ &\qquad + \widehat{R}(D^{\text{init}}, \alpha_{\text{oracle}}^{\text{init}(g)}) - R(D^{\text{init}}, \alpha_{\text{oracle}}^{\text{init}(g)}) && (5.3) \\ &\qquad + R(D^{\text{init}}, \alpha_{\text{oracle}}^{\text{init}(g)}) - R(B^{*(g)}) && (5.4) \end{aligned}$$

where $\alpha_{\text{oracle}}^{\text{init}(g)}$ is as defined in Proposition 5.1 with s set to $\frac{r(p+q)}{2}$.

We then bound (5.1) using Lemma 5.1 because $\|\sum_{k=1}^K \alpha_{\lambda,k}^{\text{learn}(g)} D_k^{\text{learn}}\|_* \le \|\alpha_\lambda^{\text{learn}(g)}\|_1$ since the optimization algorithm enforces a hard nuclear-norm constraint on D_k^{learn}. To control (5.2), we observe that although the dictionary learning procedure is nonconvex, it is guaranteed to improve the objective. Thus, we have immediately that (5.2) is at most $\lambda\|\alpha_{\text{oracle}}^{\text{init}(g)}\|_1$. A bound on (5.4) follows from Proposition 5.1 and a bound on (5.3) follows from Lemma 5.2.

6 Proof of Proposition 5.1

This proof is very involved and we will organize it into several lemmas and sections.

6.1 Linear Algebra Lemmas

Lemma 1. Let u be a unit vector, let v be a vector such that $\text{Proj}_v(u) = \frac{\langle u,v\rangle}{\|v\|_2^2} v = v$.

Then $\|u - v\|_2^2 = 1 - \langle u, v\rangle$.

Proof. Note that

$$\begin{aligned}\|u-v\|_2^2 &= \|u - \frac{u,v}{\|v\|_2^2}v\|_2^2 \\ &= \|u - \langle u, v'\rangle v'\|_2^2 \\ &= 1 - \langle u, v'\rangle^2 \\ &= 1 - \frac{\langle u, v\rangle^2}{\|v\|_2^2}\end{aligned}$$

where $v' = \frac{v}{\|v\|_2}$ is a unit vector.

Since $v = \frac{\langle u,v\rangle}{\|v\|_2^2}v$, we get that $\frac{\langle u,v\rangle}{\|v\|_2^2} = 1$ and hence:

$$\|u-v\|_2^2 = 1 - \langle u, v\rangle$$

□

Lemma 2. Let u, v be two orthonormal vectors. Let u_0, v_0 be two vectors such that

1. $(u_0^\mathsf{T} u) \geq \tau$ and $(v_0^\mathsf{T} v) \geq \tau$
2. $\|u_0\|_2 \leq 1$ and $\|v_0\|_2 \leq 1$
3. Note that the first two items implies that $(u_0^\mathsf{T} v) \leq \sqrt{1-\tau^2}$ and that $(v_0^\mathsf{T} u) \leq \sqrt{1-\tau^2}$

Then we get that $|u_0^\mathsf{T} v_0| \leq 3\sqrt{1-\tau^2}$.

Proof. Let $v_0' = v_0 - \langle v_0, u\rangle u - \langle v_0, v\rangle v$. We note that v_0' is orthogonal to u, v.

We now perform orthogonal decomposition of u_0 in term of u, v, v_0':

$$\begin{aligned}u_0 &= \langle u_0, u\rangle u + \langle u_0, v\rangle v + \frac{\langle u_0, v_0'\rangle}{\|v_0'\|_2^2}v_0' + r \\ u_0' &= \frac{\langle u_0, v_0'\rangle}{\|v_0'\|_2^2}v_0' + r\end{aligned}$$

where $u_0' = u_0 - \langle u_0, u\rangle u - \langle u_0, v\rangle v$ and r is the remainder term. Note that r is orthogonal to v_0', u, v, so by Pythagorean Theorem, we get

$$\begin{aligned}\|u_0'\|_2^2 &= \langle u_0, v_0'\rangle^2 \frac{1}{\|v_0'\|_2^2} + \|r\|_2^2 \\ \|u_0'\|_2^2 &\geq \langle u_0, v_0'\rangle^2 \frac{1}{\|v_0'\|_2^2} \\ \|v_0'\|_2\|u_0'\|_2 &\geq |\langle u_0, v_0'\rangle|\end{aligned}$$

Let $(u,v)^\perp$ denote the orthogonal complement to the span of $\{u, v\}$, then $u_0' = \text{Proj}_{(u,v)^\perp}(u_0)$ and $v_0' = \text{Proj}_{(u,v)^\perp}(v_0)$. By Pythagorean theorem and condition that $\langle u, u_0\rangle \geq \tau$ and $\langle v, v_0\rangle \geq \tau$, we get that $\|u_0'\|_2^2 \leq 1-\tau^2$ and $\|v_0'\|_2^2 \leq 1-\tau^2$.

Now let's consider RHS:

$$\begin{aligned}|\langle u_0, v_0' \rangle| &= |u_0, v_0 - \langle v_0, u \rangle u - \langle v_0, v \rangle v \rangle| \\ &= |\langle u_0, v_0 \rangle - \langle v_0, u \rangle \langle u_0, u \rangle - \langle v_0, v \rangle \langle u_0, v \rangle|\end{aligned}$$

Since $\langle v_0, u \rangle, \langle u_0, v \rangle \in [-\sqrt{1-\tau^2}, \sqrt{1-\tau^2}]$ and $\langle u, u_0 \rangle, \langle v, v_0 \rangle \in [\tau, 1]$, we get that

$$(\langle v_0, u \rangle \langle u_0, u \rangle + \langle v_0, v \rangle \langle u_0, v \rangle) \in [-2\sqrt{1-\tau^2}, 2\sqrt{1-\tau^2}]$$

Combining the two results, we get that

$$\begin{aligned}|\langle u_0, v_0 \rangle| &\leq 2\sqrt{1-\tau^2} + 1 - \tau^2 \\ &\leq 3\sqrt{1-\tau^2}\end{aligned}$$

□

6.2 Probability Theory Lemmas

Lemma 3 (Sampling lemma). Let $v \in \mathbb{R}^p$ be a random unit vector. Then v is sampled uniformly from the unit sphere S^p if and only if v is distributed identically to some random vector $\frac{X}{\|X\|_2}$ where $X = (X_1, ..., X_p) \sim N(0, I_p)$

We also give a useful extension of the Sampling Lemma.

Lemma 4 (Sample-projection lemma). Let $v \in \mathbb{R}^p$ be a random unit vector sampled uniformly from the unit sphere S^p. Let $V \subset \mathbb{R}^p$ be a subspace and let $v' = \text{Proj}_V(v)$ be the orthogonal projection of v onto V, then $\frac{v'}{\|v'\|_2}$ is distributed uniformly at random on the unit sphere $S^q \subset V$ where q is the dimension of V.

Proof. If v is sampled uniformly at random from the unit sphere S^p, then its distribution is independent of the coordinate system. That is, if $\Theta \in \mathbb{R}^{p \times p}$ is an isometry, then Θv is identically distributed as v.

Hence, without loss of generality, we can assume that the projection is onto the first q coordinates of v. The conclusion follows immediately by invoking Lemma 3. □

Lemma 5. Let v_0 be a fixed unit vector and let v be a unit vector sampled uniformly at random. Then with probability at least $c(1-\tau^2)^{p/2}$ (for some constant c), we have that $|v_0^\mathsf{T} v| \geq \tau$.

Proof. Without loss of generality, we can assume that $v_0 = (1, 0, 0, ..., 0)$. By the Sampling Lemma, we can view v as

$$v = \frac{(X_1, ..., X_p)}{\sqrt{X_1^2 + ... + X_p^2}},$$

where $X_1, ..., X_p$ are i.i.d. $N(0, 1)$. We are then interested in $|v_0^\mathsf{T} v| = \sqrt{\frac{X_1^2}{X_1^2 + ... + X_p^2}}$.

If $|v_0^\mathsf{T} v| \geq \tau$, then

$$\frac{X_1^2}{X_1^2 + \underbrace{X_2^2 + ... + X_p^2}_{Z}} \geq \tau^2.$$

Now, $\frac{X_1^2}{X_1^2 + Z} \geq \tau^2$ implies that $(1-\tau^2)X_1^2 \geq \tau^2 Z$, which implies that $\frac{X_1^2}{Z} \geq \frac{\tau^2}{1-\tau^2}$.

Note that X_1^2 has a χ^2 distribution with one degree of freedom and that $Z = X_2^2 + ... + X_p^2$ is a χ^2 distribution with $p-1$ degrees of freedom. Since Z and X_1 are independent, $R = \frac{X_1^2}{(\frac{Z}{p-1})}$ has an F-distribution with $d_1 = 1$ and $d_2 = p-1$.

Our goal now is to use known results about the CDF of the F-distribution to characterize the probability of the event $R \geq (p-1)\frac{\tau^2}{1-\tau^2}$. It is known that the CDF of the F-distribution is the *incomplete regularized Beta function*,

$$\mathbb{P}(R \leq x) = I_{\frac{d_1 x}{d_1 x + d_2}}\left(\frac{d_1}{2}, \frac{d_2}{2}\right)$$

where

$$I_{\frac{d_1 x}{d_1 x + d_2}}(\frac{d_1}{2}, \frac{d_2}{2}) = \frac{\int_0^{\frac{d_1 x}{d_1 x + d_2}} t^{\frac{d_1}{2}-1}(1-t)^{\frac{d_2}{2}-1} dt}{\int_0^1 t^{\frac{d_1}{2}-1}(1-t)^{\frac{d_2}{2}-1} dt}.$$

For our purpose, $d_1 = 1, d_2 = (p-1), x = (p-1)\frac{\tau^2}{1-\tau^2}$. Hence

$$\frac{d_1 x}{d_1 x + d_2} = \frac{\frac{\tau^2}{1-\tau^2}}{\frac{\tau^2}{1-\tau^2} + 1} = \tau^2.$$

Thus,

$$\begin{aligned}\mathbb{P}(R \geq x) &= 1 - I_{\tau^2}(\frac{1}{2}, \frac{p-1}{2}) \\ &= \frac{\int_{\tau^2}^1 t^{-\frac{1}{2}}(1-t)^{\frac{p-3}{2}} dt}{\int_0^1 t^{-\frac{1}{2}}(1-t)^{\frac{p-3}{2}} dt}.\end{aligned}$$

The denominator is a monotonically decreasing function of p for $p \geq 3$ and hence we upper bound it by some constant $0 < \frac{1}{c} \leq 4$. We lower bound the numerator according to

$$\begin{aligned}\int_{\tau^2}^1 t^{-1/2}(1-t)^{(p-3)/2} dt &\geq \int_{\tau^2}^1 (1-t)^{(p-3)/2} dt \\ &= [-(1-t)^{\frac{p-1}{2}}]_{t=\tau^2}^1 \\ &= (1-\tau^2)^{\frac{p-1}{2}} \\ &\geq (1-\tau^2)^{\frac{p}{2}}.\end{aligned}$$

Hence, $\mathbb{P}(R \geq (p-1)\frac{\tau^2}{1-\tau^2}) \geq c(1-\tau^2)^{p/2}$. □

Corollary 6. Let v_0 be a fixed unit vector and let $v_1, ..., v_K$ be K unit vectors sampled independently and uniformly at random. Then

$$\max_{i=1,...,K} |v_0^\mathsf{T} v_i| \geq \tau$$

with probability at least

$$1 - \exp\{-Kc(1-\tau^2)^{p/2}\}.$$

Proof. The probability that $|v_0^\mathsf{T} v_k| \leq \tau$ is at most $1 - c(1-\tau^2)^{p/2}$, as shown before. Since the v_ks are independent, the probability that for all v_k, $|v_0^\mathsf{T} v_k| \leq \tau$ is at most $(1 - c(1-\tau^2)^{p/2})^K$.

Define a real number t by $c(1-\tau^2)^{p/2} = \frac{t}{K}$. We now use the inequality $(1 - \frac{t}{K})^n \leq \exp(-t)$ for all $t \in [0, K]$ and get

$$(1 - c(1-\tau^2)^{p/2})^K \leq \exp(-cK(1-\tau^2)^{p/2}).$$

We thus conclude that the probability that $|v_0^\mathsf{T} v_k| \geq \tau$ for at least one v_k is at least $1 - \exp(-cK(1-\tau^2)^{p/2})$. □

We can relate the angular deviation between two unit vectors v_0, v_1 to the ℓ_2 distance between v_0 and v_0 projected onto v_1 as

$$\|v_0 - (v_0^\mathsf{T} v_1)v_1\|_2^2 = 1 - \langle v_0, v_1 \rangle^2.$$

Hence, $|v_0^\mathsf{T} v_1| \geq \tau$ if and only if $\|v_0 - \langle v_0, v_1 \rangle v_1\|_2^2 \leq 1 - \tau^2$.

For convenience, we now define $\epsilon = 1 - \tau^2$ and obtain the following proposition describing the sparse approximation error if $s = 1$.

Proposition 7. Let v_0 be a fixed unit vector and let $d_1, ...d_K$ be K iid unit vectors sampled uniformly at random. Then

$$\mathbb{P}\left(\min_{k=1,...,K} \|v_0 - (v_0^\mathsf{T} d_k)d_k\|_2^2 \leq \epsilon \right) \leq 1 - \exp(-cK\epsilon^{p/2}).$$

We now upper bound the error of the optimal sparse-coefficient approximation with the approximation error of an easy-to-analyze algorithm that is similar to Orthogonal Matching Pursuit.

Algorithm 1 Projected Orthogonal Matching Pursuit (POMP)

Let D_t be the set of dictionary entries chosen at iteration t with $D_0 = \emptyset$. Let $d_k \in \mathbb{R}^p$ be a dictionary entry; we will leave its dependency on t implicit

Let r_t be residue at iteration t with $r_0 = w$. Start algorithm with $t = 0$.

For $t = 1, ..., s$

1. Let $\alpha_k = \langle r_{t-1}, d_k \rangle$. Choose $k^* = \arg \max_{k: d_k \notin D_{t-1}} \alpha_k$. Set chosen dictionary $d_t^* = d_{k^*}$ and set $\alpha_t^* = \alpha_{k^*}$

   Note: This is equivalent to choosing $d_t^* = \arg \min_{d_k \notin D_{t-1}} \|r_{t-1} - d_k \alpha_k\|_2^2$

2. Let $D_t = D_{t-1} \cup d_t^*$, let V_t be subspace orthogonal to span of D_t.

3. Project all $d_k \notin D_t$ onto V_t and get d_k'. For all unchosen entries, **set** d_k to be $\frac{d_k'}{\|d_k'\|_2}$.

4. Let $r_t = r_{t-1} - d_t^* \alpha_t^*$.

In Theorem **??**, we show that POMP is very similar to OMP except with slight change in how it picks the new dictionary entry at every iteration.

We would like to repeatedly apply Single Vector Approximation lemma. To do that, we must first show that after we choose d_t^* and projected all remaining dictionary entries, the remaining entries are stochastically independent of each other conditioned on the entries $d_1^*, ..., d_t^*$ already chosen.

Lemma 8. Suppose POMP is at iteration t. Let d_k' be the projection of d_k onto V_t for some d_k not yet chosen. Let $\widetilde{d_k'} = \frac{d_k}{\|d_k\|_2}$.

Then the set $\{\widetilde{d_k'}\}_{d_k \notin D_t}$ is, conditioned on $\{d_1^*, ..., d_t^*\}$ independently and identically distributed as uniformly random unit vectors on sphere $S^m \subset V_t$ where $m = p - t$ is the number of remaining entries.

Proof. Since $s < p$, we get that $m > 0$. The claim then follows directly from induction and Lemma 4 (Sample-Projection Lemma). □

Proposition 9. Let $v \in \mathbb{R}^p$ be a fixed vector. Let $\widehat{v}$ be the estimate of v outputted by the POMP algorithm with sparsity level (number of non-zero entries) $s < p$ and a dictionary of size $K > 2s$.

Then $\|v - \widehat{v}\|_2^2 \leq \|v\|_2^2 \epsilon$ with probability at least $1 - s\exp(-cK\epsilon^{p/2s})$

Proof. With probability at most $\exp(-cK\epsilon^{p/2})$, $\|r_1\|_2^2 \geq \epsilon$ where r_1 is the residue after the first iteration.

Conditioned on $d*_1$, the first dictionary chosen, D_1, the remaining collection of un-chosen dictionaries are still independent and uniformly distributed on the unit sphere in V_1.

We will think of POMP at iteration t as approximating r_{t-1} by a single dictionary entry. We say iteration t is a FAILURE if the approximation differs from r_{t-1} by more than $\|r_{t-1}\|_2^2\epsilon$. Hence, if $\|r_s\|_2^2 \geq \epsilon^s$, then some iterations from 1 to s experienced failure.

Note that by Theorem **??**, r_{t-1} lies in $V_{t-1}^{\perp}$. Note also that $r_t = r_{t-1} - \langle r_{t-1}, d_t^* \rangle$ where d_t^* is chosen amongst $K - t + 1$ dictionary entries uniformly distributed on unit sphere in $V_{t-1}^{\perp}$ (by Lemma 8.

Hence, we know for one iteration, probability of failure, i.e., $\|r_t\|_2^2 \geq \epsilon\|r_{t-1}\|_2^2$, has probability at most $\exp(-c(K - t + 1)\epsilon^{(p-t+1)/2})$, which is less than $\exp(-c\frac{K}{2}\epsilon^{p/2})$ by assumption that $K > 2s$. Instead of carrying the fraction $\frac{K}{2}$ around, we will redefine c here to be old c divided by 2.

By union bound across all iterations then, $\|r_s\|_2^2 \geq \epsilon^s$ with probability at most $s\exp(-cK\epsilon^{p/2})$.

By a change of variable, we get then that $\|r_s\|_2^2 \leq \epsilon$ with probability at least $1 - s\exp(-cK\epsilon^{p/2s})$ □

6.3 Proof of Proposition 5.1

We will follow the proof of Proposition 9 and upper bound the optimal sparse approximation error with the approximation of an algorithm.

Algorithm 2 Matrix Matching Pursuit (MMP)

Let $M = \sum_j^r \sigma_j \overline{u}_j \underline{u}_j^{\mathsf{T}}$ be input matrix

Let $\{D_1, ..., D_K\}$ be a set of rank 1 dictionary entries with $D_k = \overline{d}_k \underline{d}_k^{\mathsf{T}}$

1. For each $j = 1, ..., r$
   Perform a simultaneous Matching Pursuit as followed:
   Initialize residue as a pair $(\overline{r}_0 = \overline{u}_j, \underline{r}_0 = \underline{u}_j)$
   (a) For each iteration $t = 1, ...s/r$
   Find k^* to maximize $\min(|\overline{r}_{t-1}^{\mathsf{T}} \overline{d}_{k^*}|, |\underline{r}_{t-1}^{\mathsf{T}} \underline{d}_{k^*}|)$
   Add D_{k^*} to set of dictionary entries chosen so far. Project all remaining $\{\overline{d}_k\}$ and $\{\underline{d}_k\}$ to orthogonal complement of $\overline{d}_{k^*}$ and $\underline{d}_{k^*}$ respectively. Update residue $\overline{r}_t = \overline{r}_{t-1} - (\overline{r}_{t-1}^{\mathsf{T}} \overline{d}_{k^*})\overline{d}_{k^*}, \underline{r}_t = \underline{r}_{t-1} - (\underline{r}_{t-1}^{\mathsf{T}} \underline{d}_{k^*})\underline{d}_{k^*}$
   (b) Combine all k^* to get an approximation $(\overline{u}_{A_j}, \underline{u}_{A_j})$ for $(\overline{u}_j, \underline{u}_j)$
2. Output final approximation as $\sum_{j=1}^r \sigma_j \overline{u}_{A_j} \underline{u}_{A_j}^{\mathsf{T}}$

Note: To help explain the analysis, we will introduce a little bit of vocabulary. Suppose matrix M has SVD $M = \sum_{j=1}^r \sigma_j \overline{u}_j \underline{u}_j^{\mathsf{T}}$, then we will say each $\sigma_j \overline{u}_j \underline{u}_j^{\mathsf{T}}$ is an **eigen-atom**. We can think of Matrix Matching Pursuit then as estimating each of the r eigen-atoms separately through a Dual-POMP procedure.

In order word, Matrix Matching Pursuit estimates each eigen-atom $\overline{u}_j \underline{u}_j^\mathsf{T}$ with s/r dictionary entries: $\overline{u}_{A_j} \underline{u}_{A_j}^\mathsf{T}$. We then combine all r of such eigen-atom estimates to get the final approximation.

The first three lemma builds up to a concentration of measure statement saying that for all eigen-atoms $(\overline{u}_i, \underline{u}_i)$, the MMP estimate $(\overline{u}_{A_i}, \underline{u}_{A_i})$ are close to $(\overline{u}_i, \underline{u}_i)$.

Essentially, the following random events must all simultaneously hold.

1. Fix eigen-atom, an iteration of MMP must estimate a pair of residues $(\overline{r}_t, \underline{r}_t)$ well
2. Fix eigen-atom, all iterations must estimate well
3. All eigen-atoms must be estimated well.

Lemma 10 (Dual Angular Deviation). Let $\{(\overline{d}_1, \underline{d}_1), ..., (\overline{d}_K, \underline{d}_K)\}$ be a dictionary of pairs of random vectors.

Let $(\overline{u}, \underline{u})$ be two fixed vectors. Then with probability at least $1 - \exp\{-c^2 K(1-\tau^2)^{(p+q)/2}\}$, there exist $k \in \{1, ..., K\}$ such that

$$(\overline{u}^\mathsf{T} \overline{d}_k)^2 \geq \tau^2 \|\overline{u}\|_2^2 \text{ and } (\underline{u}^\mathsf{T} \underline{d}_k)^2 \geq \tau^2 \|\underline{u}\|_2^2$$

This lemma states that one iteration of MMP must estimate both $\underline{u}_i$ and $\overline{u}_i$ well.

Proof. We first fix a particular k.

By Lemma 5 and by independence of $\overline{d}_k$ and $\underline{d}_k$, we get that with probability at least $c^2(1-\tau^2)^{(p+q)/2}$, $(\overline{u}^\mathsf{T} \overline{d}_k)^2 \geq \tau^2$ and $(\underline{u}^\mathsf{T} \underline{d}_k)^2 \geq \tau^2$.

Therefore, by similar proof technique as Corollary 6, we exploit the independence between $(\overline{d}_k, \underline{d}_k)$ and $(\overline{d_{k'}}, \underline{d_{k'}})$ and get the desired result. □

The following lemma says that for a fixed eigen-atom, all iterations of MMP estimate the corresponding residues well.

Lemma 11. Let $\overline{u}\underline{u}^\mathsf{T}$ be a fixed rank-1 matrix and $\|\overline{u}\|_2 = \|\underline{u}\|_2 = 1$. Let $\overline{u}_A \underline{u}_A^\mathsf{T}$ be estimate of $\overline{u}\underline{u}^\mathsf{T}$ outputted by the OMP+ algorithm with sparsity level s/r and dictionary of size K.

Then $\overline{u}^\mathsf{T} \overline{u}_A \geq \tau$ AND $\underline{u}^\mathsf{T} \underline{u}_A \geq \tau$ with probability at least $1 - \frac{s}{r} \exp\left(-c^2 K (1-\tau)^{r(p+q)/2s}\right)$

Proof. First, note that $\frac{s}{r} < (p+q)$ and that Lemma 8 still applies. At every iteration, the remaining dictionary entries $\underline{d}_k$'s and $\overline{d}_k$'s are still uniformly and independently distributed on the unit sphere in the orthogonal complement of the span of dictionary entries chosen so far.

We will declare iteration t a SUCCESS if, for some $\epsilon > 0$, $\|\underline{r}_t\|_2^2 \leq \|\underline{r}_{t-1}\|_2^2 \epsilon$ AND $\|\overline{r}_t\|_2^2 \leq \|\overline{r}_{t-1}\|_2^2 \epsilon$

Note that $\underline{r}_t = \underline{r}_{t-1} - \langle \underline{r}_{t-1}, \underline{d}_t^* \rangle \underline{d}_t^*$ and

$$\begin{aligned} \|\underline{r}_t\|_2^2 &= \|\underline{r}_{t-1} - \langle \underline{r}_{t-1}, \underline{d}_t^* \rangle \underline{d}_t^*\|_2^2 \\ &= \|\underline{r}_{t-1}\|_2^2 - \langle \underline{r}_{t-1}, \underline{d}_t^* \rangle^2 \end{aligned}$$

And similarly, $\|\overline{r}_t\|_2^2 = \|\overline{r}_{t-1}\|_2^2 - \langle \overline{r}_{t-1}, \overline{d}_t^* \rangle^2$.

By Lemma 10, $\langle \underline{r}_{t-1}, \underline{d}_t^* \rangle^2 \geq \tau^2 \|\underline{r}_{t-1}\|_2^2$ AND $\langle \overline{r}_{t-1}, \overline{d}_t^* \rangle^2 \geq \tau^2 \|\overline{r}_{t-1}\|_2^2$ with probability at least $1 - \exp\{-c^2(K-t+1)(1-\tau^2)^{(p+q-2(t-1))/2}\}$. Using that $\frac{K}{2} > \frac{s}{r}$ and redefining c, we have that $1 - \exp\{-c^2(K-t+1)(1-\tau^2)^{(p+q-2(t-1))/2}\} \geq 1 - \exp\{-c^2 K(1-\tau^2)^{(p+q)/2}\}$.

Hence, with probability at least $1 - \exp\{-c^2 K(1-\tau^2)^{(p+q)/2}\}$,

$$\begin{aligned} \|\underline{r}_t\|_2^2 &= \|\underline{r}_{t-1}\|_2^2 - \langle \underline{r}_{t-1}, \underline{d}_t^* \rangle^2 \\ &\leq \|\underline{r}_{t-1}\|_2^2 (1-\tau^2) \end{aligned}$$

AND $\|\overline{r}_t\|_2^2 \leq \|\overline{r}_{t-1}\|_2^2(1-\tau^2)$. Setting $\epsilon = (1-\tau^2)$ and we have that round t SUCCEEDS with probability at least $1 - \exp\{-c^2 K \epsilon^{(p+q)/2}\}$.

By union bound, the probability that at least one iteration FAILS is at most $(s/r)\exp\{-c^2 K \epsilon^{(p+q)/2}\}$. Hence, probability $\|\overline{u} - \overline{u}_A\|_2^2 \geq \epsilon^{s/r}$ or $\|\underline{u} - \underline{u}_A\|_2^2 \geq \epsilon^{s/r}$ is at most $(s/r)\exp\{-c^2 K \epsilon^{(p+q)/2}\}$.

By a change of variable (setting $\epsilon = \epsilon^{s/r}$), we get that $\|\overline{u} - \overline{u}_A\|_2^2 \geq \epsilon$ OR $\|\underline{u} - \underline{u}_A\|_2^2 \geq \epsilon$ with probability at most $(s/r)\exp\{-c^2 K \epsilon^{r(p+q)/2s}\}$

We will now finish the proof by relating the event $\|\overline{u} - \overline{u}_A\|_2^2 \leq \epsilon\|$ AND $\|\underline{u} - \underline{u}_A\|_2^2 \leq \epsilon$ to the event $(\overline{u}^\mathsf{T}\overline{u}_A)^2 \geq \tau^2$ AND $(\underline{u}^\mathsf{T}\underline{u}_A)^2 \geq \tau^2$ for some τ dependent on ϵ.

By Theorem **??**, we know that $\overline{u}_A$ and $\underline{u}_A$ are the orthogonal projections of $\overline{u}$ and $\underline{u}$ respectively to the span of the dictionaries $\{\overline{d}_t^*\}$ and $\{\underline{d}_t^*\}$. Therefore, Lemma 1 applies and $\|\overline{u} - \overline{u}_A\|_2^2 \leq \epsilon$ AND $\|\underline{u} - \underline{u}_A\|_2^2 \leq \epsilon$ implies that $(\overline{u}^\mathsf{T}\overline{u}_A) \geq 1-\epsilon$ AND $(\underline{u}^\mathsf{T}\underline{u}_A) \geq 1-\epsilon$.

Make a change of variable by letting $\tau = 1-\epsilon$ and we complete the proof. □

The following proposition uses previous lemmas and analyzes error of approximating one r-ranked matrix with the MMP procedure.

Proposition 12. Let $M = \sum_{i=1}^r \sigma_i \overline{u}_i \underline{u}_i^\mathsf{T}$ be a $rank$-r matrix where $\{\overline{u}_i\}$ are orthonormal and $\{\underline{u}_i\}$ are orthonormal. Suppose also that $\sum_{i=1}^r \sigma_i^2 = 1$ (i.e. $\|M\|_F^2 = 1$).

Let $\{(\overline{u}_{A_i}, \underline{u}_{A_i})\}_{i=1,...,r}$ be estimates produced by MMP with a dictionary of size K.

Let $M_A = \sum_{i=1}^r \sigma_i \overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T}$, then with probability at least $1 - s\exp(-c^2 K(\frac{\epsilon}{22r})^{r(p+q)/2s})$, $\|M - M_A\|_F^2 \leq \epsilon$.

Proof.

$$\begin{aligned}
\|M - M_A\|_F^2 &= \|\sum_{i=1}^r \sigma_i \overline{u}_i \underline{u}_i^\mathsf{T} - \sum_{i=1}^r \sigma_i \overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T}\|_F^2 \\
&= \|\sum_{i=1}^r \sigma_i (\overline{u}_i \underline{u}_i^\mathsf{T} - \overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T})\|_F^2 \\
&= \|\sum_{i=1}^r \sigma_i \Delta_i\|_F^2 \\
&= \sum_{i,j=1}^r \sigma_i \sigma_j \langle \Delta_i, \Delta_j \rangle
\end{aligned}$$

where $\Delta_i = \overline{u}_i \underline{u}_i^\mathsf{T} - \overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T}$ is a $q \times p$ matrix.

We now divide the rest of the proof into two steps. In the first step we bound $\|\Delta_i\|_F^2$ and in the second step we bound $\langle \Delta_i, \Delta_j \rangle$ for $i \neq j$.

First Step:

$$\begin{aligned}
\|\Delta_i\|_F^2 &= \|\overline{u}_i \underline{u}_i^\mathsf{T} - \overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T}\|_F^2 \\
&= \|\overline{u_1} \underline{u}_i^\mathsf{T}\|_F^2 + \|\overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T}\|_F^2 - 2\langle \overline{u}_i \underline{u}_i^\mathsf{T}, \overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T} \rangle \\
&\leq 2 - 2\mathrm{tr}((\overline{u}_i \underline{u}_i^\mathsf{T})^\mathsf{T} \overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T}) \\
&= 2 - 2(\overline{u}_i^\mathsf{T} \overline{u}_{A_i})(\underline{u}_i^\mathsf{T} \underline{u}_{A_i})
\end{aligned}$$

From previous Lemma 11 and by a union bound across $i = 1, ..., r$, we get that with probability at least $1 - s\exp\left(-c^2 K(1-\tau)^{r(p+q)/2s}\right)$, for all $i = 1, ..., r$, $(\overline{u}_i^\mathsf{T} \overline{u}_{A_i}) \geq \tau$ AND $(\underline{u}_i^\mathsf{T} \underline{u}_{A_i}) \geq \tau$. We will denote this event $\mathcal{S}$.

On event $\mathcal{S}$ then, $\|\Delta_i\|_F^2 \leq 2(1-\tau^2)$.

Second Step: Assume $i \neq j$

$$\begin{aligned}\langle \Delta_i, \Delta_j \rangle &= \langle \overline{u}_i \underline{u}_i^\mathsf{T} - \overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T}, \overline{u}_j \underline{u}_j^\mathsf{T} - \overline{u}_{A_j} \underline{u}_{A_j}^\mathsf{T} \rangle \\ &= \underbrace{\langle \overline{u}_i \underline{u}_i^\mathsf{T}, \overline{u}_j \underline{u}_j^\mathsf{T} \rangle}_{\text{term 1}} - \underbrace{\langle \overline{u}_i \underline{u}_i^\mathsf{T}, \overline{u}_{A_j} \underline{u}_{A_j}^\mathsf{T} \rangle}_{\text{term 2}} - \underbrace{\langle \overline{u}_j \underline{u}_j^\mathsf{T}, \overline{u}_{A_i} \underline{u}_{A_i} \rangle}_{\text{term 3}} + \underbrace{\langle \overline{u}_{A_i} \underline{u}_{A_i}^\mathsf{T}, \overline{u}_{A_j} \underline{u}_{A_j} \rangle}_{\text{term 4}}\end{aligned}$$

Term 1 is $(\overline{u}_i^\mathsf{T} \overline{u}_j)(\underline{u}_i^\mathsf{T} \underline{u}_j) = 0$ by orthonormality.

Term 2 is $(\overline{u}_i^\mathsf{T} \overline{u}_{A_j})(\underline{u}_i^\mathsf{T} \underline{u}_{A_j})$. In event $\mathcal{S}$:

1. $(\overline{u}_j^\mathsf{T} \overline{u}_{A_j}) \geq \tau$ AND $(\underline{u}_j^\mathsf{T} \underline{u}_{A_j}) \geq \tau$ for all j
2. $\|\overline{u}_{A_j}\|_2^2 \leq 1$ AND $\|\underline{u}_{A_j}\|_2^2$ for all j

Hence, because $\overline{u}_j$ and $\overline{u}_i$ are orthogonal for $i \neq j$, we can invoke Pythagorean theorem and get that $|\overline{u}_i^\mathsf{T} \overline{u}_{A_j}| \leq \sqrt{1-\tau^2}$ for all i, j.

Therefore, for all i, j, Term 2 is bounded by $1-\tau^2$ in absolute value. Similarly, Term 3 is bounded by $1-\tau^2$ in absolute value for all i, j.

Term 4 is $(\overline{u}_{A_i}^\mathsf{T} \overline{u}_{A_j})(\underline{u}_{A_i}^\mathsf{T} \underline{u}_{A_j})$. In event $\mathcal{S}$, we can invoke Lemma 2 and get that for all i, j, Term 4 is bounded by $9(1-\tau^2)$ in absolute value.

Hence, $|\langle \Delta_i, \Delta_j \rangle| \leq 11(1-\tau^2)$.

To finish the proof, we pick up from the beginning and note that, on event $\mathcal{S}$:

$$\begin{aligned}\|M - M_A\|_F^2 &= \sum_{i,j=1}^{r} \sigma_i \sigma_j \langle \Delta_i, \Delta_j \rangle \\ &\leq \sum_{i,j=1}^{r} \sigma_i \sigma_j 11(1-\tau^2) \\ &= (\sum_{i=1}^{r} \sigma_i)^2 11(1-\tau^2) \\ &\leq 11\|M\|_*^2(1-\tau^2) \\ &\leq 22\|M\|_*^2(1-\tau)\end{aligned}$$

where for the last step, we used that $(1-\tau^2) = (1-\tau)(1+\tau) \leq 2(1-\tau)$ for $\tau \in [0,1]$.

The probability of event $\mathcal{S}$ is at least $1 - s\exp(-c^2K(1-\tau)^{r(p+q)/2s})$. By letting $\epsilon = 22\|M\|_*^2(1-\tau)$, we get that:

$\|M - M_A\|_F^2 \leq \epsilon$ with probability at least $1 - s\exp(-c^2K(\frac{\epsilon}{22\|M\|_*^2})^{r(p+q)/2s})$.

□

By definition of $B^{0(g)}$ as the optimal dictionary approximation with s-sparse coefficients, and by taking union bound, we know that with probability at least $1 - sG\exp(-cK\frac{\epsilon}{22\|B^{*(g)}\|_*^2}^{r(p+q)/2s})$,

$$\max_{g=1,\ldots,G} \|B^{*(g)} - B^{0(g)}\|_F^2 \leq \epsilon$$

Setting probability $\delta = \frac{1}{K} = sG\exp(-cK\frac{\epsilon}{22\|B^{*(g)}\|_*^2}^{r(p+q)/2s})$ and we get that with probability $1 - \frac{1}{K}$,

$$\max_{g=1,\ldots,G} \|B^{*(g)} - B^{0(g)}\|_F^2 \leq 22\|B^{*(g)}\|_*^2 \left(\frac{1}{c_2 K}\log(GKs)\right)^{2s/r(p+q)}$$

Proposition 5.1 then follows in a straightforward manner by plugging in desired value of K.

7 Dictionary Optimization Algorithm

In this section, we give the details for the optimization algorithms that solve the following:

$$\min_{D \in C_D(\tau)} f(\alpha, D)$$

where we have

$$f(\alpha, D) = \frac{1}{G}\sum_{g=1}^{G}\left\{\frac{1}{n}\|Y^{(g)} - \left(\sum_{k=1}^{K}\alpha_k^{(g)} D_k\right) X^{(g)}\|_F^2 + \lambda\|\alpha^{(g)}\|_1\right\}$$

and

$$C_D(\tau) = \left\{D \in \mathbb{R}^{q\times p} : \|D\|_* \leq \tau \text{ and } \|D\|_2 \leq 1\right\}$$

7.1 Projected Gradient Descent

The projected gradient descent algorithm, although simple, is too slow for many applications.

In this algorithm, Q_L is defined by

$$Q_L(D', D) = f(\alpha, D) + \sum_{k=1}^{K}\langle D'_k - D_k, \nabla_k\rangle + \frac{L}{2}\sum_{k=1}^{K}\|D'_k - D_k\|_F^2.$$

The `SimulProject`(Σ, τ) function performs a simultaneous projection of the diagonal of the matrix Σ onto the intersection of the ℓ_1-ball of radius τ and the ℓ_∞-ball of radius one. We give details on how to perform this project in section 7.3.

7.2 FISTA

The Fast Iterative Shrinkage and Thresholding Algorithm, based on Nesterov's optimal first-order optimization algorithm, proposed by Beck & Teboulle (2009a), is generally faster than the projected gradient descent algorithm.

FISTA is not guaranteed to improve the objective at every iteration and this can lead to instability and divergence. We adapt the monotonic version of FISTA described in Beck & Teboulle (2009b) for dictionary learning.

Algorithm 3 Projected Gradient Descent for Dictionary Learning

Input: $Y^{(g)} \in \mathbb{R}^{q\times n}$, $X^{(g)} \in \mathbb{R}^{p\times n}$, and $\alpha^{(g)} \in \mathbb{R}^K$ for $g = 1, \ldots, G$; $\tau \in \mathbb{R}$, $\gamma > 1$.

Output: $D_1, ..., D_K \in \mathbb{R}^{q\times p}$

1. For $k = 1, ..., K$, generate random unit vectors $\overline{u}, \underline{u}$, and set $D_k = \overline{u}\underline{u}^\mathsf{T}$. Set $L = 1$.
2. Iterate until convergence:
   Precompute

$$\nabla_{all}^{(g)} = \frac{1}{n} Y^{(g)} - \frac{1}{n} \sum_{k=1}^{K} D_k \left(\alpha_k^{(g)} X^{(g)} \right)$$

   (a) For each k, compute the gradient $\nabla_k = -\frac{1}{G} \sum_{g=1}^{G} \nabla_{all} \left(\alpha_k^{(g)} X^{(g)} \right)^\mathsf{T}$

   (b) For each k:

   compute the SVD $D_k - \frac{1}{L}\nabla_k = U_k \Sigma_k V_k^\mathsf{T}$;
   project $\Sigma_k' = \texttt{SimulProject}(\Sigma_k, \tau)$;
   update $D_k' = U_k \Sigma_k' V_k^\mathsf{T}$.

   (c) If $f(\alpha, D') > Q_L(D', D)$, set $L \leftarrow \gamma L$ and repeat from (b).

   (d) Set $D \leftarrow D'$.

7.3 `SimulProject`

Proof of correctness of Algorithm 5. Let $B_1(\tau)$ denote the l_1-ball of radius τ and let $B_\infty(1)$ denote the l_∞-ball of radius 1.

We first note that the instructions given in computing λ in step 2 is correct as shown in Duchi et al. (2008).

We proceed by induction on the dimensionality of the input vector. If $v \in \mathbb{R}$, then clearly the projection is just $v^{new} = \min(\tau, 1)$ and would be outputted by either step 3 or 4 of the algorithm.

Suppose then we have a vector of dimension p. We now perform case analysis:

In case 1, step 3 terminates. In this case, we projected to l_1-ball and have also landed in the l_∞-ball. We claim that v^{new} is the correct projection onto the intersection. Let $u \in B_1(\tau) \cap B_\infty(1)$, by definition, $\|v - v^{new}\|_2 \leq \|u - v\|_2$. Since $v^{new} \in B_\infty(1)$ as well, we get that v^{new} is the correct projection.

In case 2, step 4 terminates. In this case, we projected to the l_∞-ball and have also landed in the l_1-ball. By same argument as before, v^{new} is the correct projection.

In case 3: we go into step 5. We must now project to boundary of both the l_∞-ball and l_1-ball. Thus, we need to find a v^{new} to minimize $||v - v^{new}||_2^2$ and such that BOTH $\max_i v_i^{new} = 1$ and $\sum_i v_i^{new} = \tau$.

v_1^{new} must equal 1 since $v_1 = \max_i v_i$. The remaining v' must then both be in the l_1-ball of radius $\tau - 1$

Algorithm 4 Monotonic FISTA for Dictionary Learning

Input: $Y^{(g)} \in \mathbb{R}^{q\times n}$, $X^{(g)} \in \mathbb{R}^{p\times n}$, and $\alpha^{(g)} \in \mathbb{R}^K$ for $g = 1, \ldots, G$; $\tau \in \mathbb{R}$, $\gamma > 1$.

Output: $D_1, ..., D_K \in \mathbb{R}^{q\times p}$

1. For $k = 1, ..., K$, generate random unit vectors $\overline{u}, \underline{u}$, and set $D_{k,1} = \overline{u}\underline{u}^{\mathsf{T}}$. Set $L = 1$.
   Set $D'_{k,1} = D_{k,1}$, set $c_1 = 1$
2. Iterate $t = 1, ..., T$
   Precompute

$$\nabla^{(g)}_{all} = \frac{1}{n}Y^{(g)} - \frac{1}{n}\sum_{k=1}^{K} D'_{k,t}\left(\alpha^{(g)}_k X^{(g)}\right)$$

   (a) For each k, compute the gradient $\nabla_k = -\frac{1}{G}\sum_{g=1}^{G} \nabla^{(g)}_{all}\left(\alpha^{(g)}_k X^{(g)}\right)^{\mathsf{T}}$

   (b) For each k:

   compute the SVD $D'_{k,t} - \frac{1}{L}\nabla_k = U_k\Sigma_k V_k^{\mathsf{T}}$;

   project $\Sigma'_k = \texttt{SimulProject}(\Sigma_k, \tau)$;

   update $D''_{k,t} = U_k\Sigma'_k V_k^{\mathsf{T}}$.

   (c) If $f(\alpha, D''_{k,t}) > Q_L(D''_{k,t}, D'_{k,t})$, set $L \leftarrow \gamma L$ and repeat from (b).

   (d) Set $c_{t+1} = \frac{1+\sqrt{1+4c_t^2}}{2}$

   (e) For each k:

   Set $D_{k,t} = \arg\min\{f(\alpha, D) : D = D'_{k,t}, D_{k,t-1}\}$

   Set $D'_{k,t+1} = D_{k,t} + \frac{c_t}{c_{t+1}}(D''_{k,t} - D_{k,t}) + \frac{c_t - 1}{c_{t+1}}(D_{k,t} - D_{k,t-1})$

Algorithm 5 `SimulProject`, simultaneous projection onto l_1-ball of radius τ and l_∞-ball of radius 1

- IN: Σ diagonal matrix, $\tau > 0$
- OUT: Σ'

1. Let $v \in \mathbb{R}^p$ be the diagonal of Σ, suppose without loss of generality that $v_1 \geq v_2 \geq ...v_p$.
2. Compute $\lambda \geq 0$, with the following steps, such that soft-thresholding by λ projects v onto l_1-ball of radius τ
   (a) let $k = \max\left\{j = 1, ..., p : v_j - \frac{1}{j}\left(\sum_{r=1}^{j} u_r - \tau\right) > 0\right\}$
   (b) Let $\lambda = \frac{1}{k}\left(\sum_{i=1}^{j} v_j - \tau\right)$
3. Soft-threshold v by λ to get v^{new}. If $v_1^{new} \leq 1$, set diagonal of Σ' as v^{new} and return.
4. Set all $v_i \geq 1$ to be 1 to get v^{new}. If $\sum_i v_i^{new} \leq \tau$, set diagonal of Σ' as v^{new} and return.
5. Set v_1 as 1. Let $v' = (v_2, ..., v_n)$. Recursively call $\texttt{SimulProject}(v', \tau - 1)$

and be in the l_∞-ball of 1. The correctness of the algorithm then follows by inductive hypothesis. □

8 Discussion of Lower Bound for Proposition 5.1

Theorem 13. (Jeong and Kim Jeong & Kim (2009)) For any dictionary $D = d_1, ..., d_K$, we have

$$\liminf_{p\to\infty} \left[\log \overline{d_s}(D) + \frac{2s \log K}{p-s} + \log \frac{p}{p-s} + \frac{s}{p-s} \log \frac{p}{s} \right] \geq 0$$

where $\overline{d_s}(D) = \mathbb{E}_{v\sim \text{Unif}(S^{p-1})} \|v - v^0\|_2^2$ with v drawn uniformly in the p-dimensional unit sphere and v^0 being its optimal s-sparse approximation with respect to dictionary D.

From this, we can get that for any dictionary D, it must be that

$$\overline{d_s}(D) = \Omega \left[(\frac{1}{K})^{2s/(p-s)} (\frac{p-s}{p}) (\frac{s}{p})^{s/(p-s)} \right]$$

If we let $p - s$ be $\Theta(p)$, we get that $\lim_{p\to\infty} (\frac{s}{p})^{s/(p-s)}$ and $\lim_{p\to\infty} \frac{p-s}{p}$ are both constants.

Our analysis assumes throughout that $p - s \geq \frac{p}{2}$ and hence, the lower bound becomes $(\frac{1}{K})^{4s/p}$ compared to our upper bound of $(\frac{\log K}{K})^{2s/p}$ (extra 2 because we are consider squared error). This is tight enough for our purpose because any constant multiplier of the exponent becomes a constant when we apply our approximation error result to derive the statistical rate of convergence.

9 Simulation Experiments on Overlapping Groups

We consider the following simulation setting:

- We let $X \in \mathbb{R}^p$ and $Y \in \mathbb{R}^q$ and estimate 10×10 matrices. We let each group have $n = 50$ samples and we run our algorithm with $K = 40$ dictionary entries.
- We create 50 groups where group 1 to group 20 all completely overlap. Group 30 to group 50 also completely overlap although are completely different from group 1 to group 20. Group 20 to group 30 overlap with both group 1 to 20 and group 30 to 50; half of data in group 20 to group 30 are from group 1 to 20 and half are from group 30 to 50.
- Group 1 to 20 and group 30 to 50 each have a distinct oracle regression matrix. That is, we generate data as $Y = BX + \epsilon$ where B is the oracle regression matrix.

As shown in Figure 1, we observe that our algorithm does consider group 1 and 20 to be highly related: these use the same set of learned dictionary entries. The same is true for group 30 to 50. Group 20 to 30, which overlap in part with group 1 to 20 and with group 30 to 50, is shown to use some of the dictionary entries used by group 1 to 20 and some of the dictionary entries used to group 30 to 50 as expected. Note that because of the sparsity penalty on the coefficients, many of the extraneous dictionary entries are not used by any of the groups, which is appropriate because the groups exhibit high level of overlap and hence does not require all $K = 40$ dictionary entries to model.

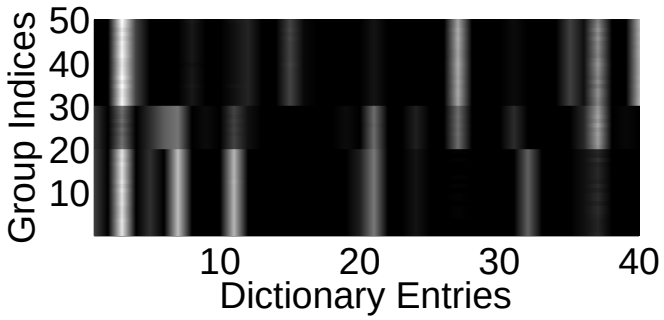

Figure 1: Group relatedness in the overlap simulation experiments. Each value indicates how strongly each group utilizes each dictionary entries. Ligher color correspond to stronger coefficients.

References

Beck, Amir and Teboulle, Marc. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sciences*, 2(1):183–202, 2009a.

Beck, Amir and Teboulle, Marc. Fast gradient-based algorithms for constrained total variation image denoising and deblurring problems. *IEEE Trans. Image Process*, 2009b.

Duchi, John, Shalev-Shwartz, Shai, Singer, Yoram, and Chandra, Tushar. Efficient projections onto the l_1-ball for learning in high dimensions. *ICML*, 2008.

Jeong, H. and Kim, Y-H. Sparse linear representation. *ISIT*, 2009.

Vershynin, Roman. *Introduction to the Non-asymptotic analysis of Random Matrices*, chapter 5. Cambridge University Press, 2011.