Alternating Minimization for Regression Problems with Vector-valued Outputs

Prateek Jain
Microsoft Research, INDIA
prajain@microsoft.com

Ambuj Tewari
University of Michigan, Ann Arbor, USA
tewaria@umich.edu

Abstract

In regression problems involving vector-valued outputs (or equivalently, multiple responses), it is well known that the maximum likelihood estimator (MLE), which takes noise covariance structure into account, can be significantly more accurate than the ordinary least squares (OLS) estimator. However, existing literature compares OLS and MLE in terms of their asymptotic, not finite sample, guarantees. More crucially, computing the MLE in general requires solving a non-convex optimization problem and is not known to be efficiently solvable. We provide finite sample upper and lower bounds on the estimation error of OLS and MLE, in two popular models: a) Pooled model, b) Seemingly Unrelated Regression (SUR) model. We provide precise instances where the MLE is significantly more accurate than OLS. Furthermore, for both models, we show that the output of a computationally efficient alternating minimization procedure enjoys the same performance guarantee as MLE, up to universal constants. Finally, we show that for high-dimensional settings as well, the alternating minimization procedure leads to significantly more accurate solutions than the corresponding OLS solutions but with error bound that depends only logarithmically on the data dimensionality.

1 Introduction

Regression problems with vector-valued (or, equivalently, multiple) response variables – where we want to predict multiple responses based on a set of predictor variables– is a classical problem that arises in a wide variety of fields such as economics [1, 2, 3], and genomics [4]. In such problems, it is natural to assume that the noise, or error, terms in the underlying linear regression model are correlated across the response variables. For example, in multi-task learning, the errors in different task outputs can be heavily correlated due to similarity of the tasks.

Regression with multiple responses is a classical topic. Textbooks in statistics [5, 6] and econometrics [7] cover it in detail and illustrate practical applications. [2] and [3] provide recent overviews of the Seemingly Unrelated Regressions (SUR) model and the associated estimation procedures. It is well known that for SUR models, the standard Ordinary Least Squares (OLS) estimator may not be (asymptotically) efficient (i.e., may not achieve the Cramer-Rao lower bound on the asymptotic variance) and that efficiency can be gained by using an estimator that exploits noise correlations [3] such as the Maximum Likelihood Estimator (MLE). The two well-known exceptions to this underperformance of OLS are: when the noise across tasks is uncorrelated and when the regressors are shared across tasks. The later is the well-known multivariate regression (MR) setting (see [5, Chapter 6]). However, there are at least two limitations of the existing MLE literature in this context.

First, despite being a classical and widely studied problem, little attention has been paid to the fact that MLE involves solving a non-convex optimization problem in general and is not known to be efficiently solvable. For example, a standard text in econometrics [7, p. 298, footnote 15], when discussing the SUR model, says, “We note, this procedure [i.e., AltMin] produces the MLE when it converges, but it is not guaranteed to converge, nor is it assured that there is a unique MLE.” The
Second, improvement claims for MLE over OLS are based on asymptotic efficiency comparisons [7] Chapter 10] that are valid only in the limit as the sample size goes to infinity. Little is known about the estimation error with a finite number of samples. When discussing the failure of AltMin to converge even after 1,000 iterations, the text [2] says that the “problem with this application may be the very small sample size, 17 observations”. This is consistent with our theoretical results that guarantee error bounds for AltMin once the sample size is large enough (in a quantifiable way).

The main contribution of this paper is quantifying, via finite sample bounds, the improvement in estimation error resulting from joint estimation of the regression coefficients and the noise covariance. Our approach is firmly rooted in the statistical learning theory tradition: we pay attention to efficient computation and use concentration inequalities, rather than limit theorems, to derive finite sample guarantees. In order to have a computationally efficient approach, we adopt an alternating minimization (AltMin) procedure that alternatingly estimates the regression coefficients and the noise covariance matrix, while keeping the other unknown fixed. While both of the individual problems are “easy” to solve and can be implemented efficiently, the general problem is still non-convex and such a procedure might lead to local optima. Whereas practitioners have long recognized that AltMin works well for such problems [1, Chapter 5], we are not aware of any provable guarantees for it in the setting of multiple response regression.

We consider two widely-used vector-output models, namely the Pooled model (Section 2) and the Seemingly Unrelated Regression (SUR) model (Section 3). For both models, we show that the estimation errors reduce by at least a constant factor in each iteration of AltMin.

**Illustrative Example.** To whet the reader’s appetite for what follows, we consider here a simple regression problem with two responses: \(y_{i,1} = X_{i,1}^\top w_* + \eta_i, y_{i,2} = X_{i,2}^\top w_* + \eta_i, 1 \leq i \leq n\), where \(X_{i,1}, X_{i,2} \in \mathbb{R}^d\) are drawn i.i.d. from the spherical normal distribution. The coefficient vector \(w_*\) is shared across the two problems, which holds true in the pooled model studied in Section 2. Later in Section 3 we also consider the SUR model that allows for different coefficient vectors across problems. More importantly, notice that the i.i.d. noise, \(\eta_i\) (say, it is standard Gaussian) is shared across the two problems. If we estimate \(w_*\) using OLS:

\[
  w_{OLS} = \arg \min_w \frac{1}{n} \sum_{i=1}^{n} (y_{i,1} - X_{i,1}^\top w)^2 + \frac{1}{n} \sum_{i=1}^{n} (y_{i,1} - X_{i,1}^\top w)^2
\]

then we will have \(\|w_{OLS} - w_*\|_2^2 = \Omega(1/\sqrt{n})\). However, subtracting the two equations gives:

\[
y_{i,1} - y_{i,2} = (X_{i,1} - X_{i,2})^\top w_*, 1 \leq i \leq n.
\]

That is, as soon as we have \(n \geq d\) samples, we will recover \(w_*\) exactly by solving the above system of linear equations!

Our toy example motivates the fundamental question that this paper answers: how much can we improve OLS by exploiting noise correlations? Let us make the example more realistic by assuming the model: \(y_i = X_i^\top w_* + \eta_i, 1 \leq i \leq n\), where \(X_i \in \mathbb{R}^{m \times d}\) is sampled i.i.d. from the standard Gaussian and noise vector \(\eta_i\) is drawn from \(\mathcal{N}(0, \Sigma_*)\). A corollary of the main result in Section 2 shows that MLE (see (3)) improves upon the OLS parameter error bound by a factor of \(\text{Error}_{OLS}/\text{Error}_{MLE} = \text{tr}(\Sigma_*)/m^2\). This factor can easily be seen to be larger than 1 by using Cauchy-Schwarz inequality: \(\text{tr}(\Sigma_*) \text{tr}(\Sigma_*^{-1})/m^2 = (\sum_j \lambda_j)(\sum_j 1/\lambda_j)/m^2 \geq (\sum_j \sqrt{\lambda_j}/c)^2/m^2 = 1\), where \(\lambda_j\) is the \(j\)-th largest eigenvalue of \(\Sigma_*\). The inequality is tight when \(\sqrt{\lambda_j} = c/\sqrt{\lambda_j}\) for some constant \(c\). That is, when \(\Sigma_* = cI\) which holds true if the noise in each response is mutually independent and has same variance.

The more \(\Sigma_*\) departs from being \(c \cdot I\), the larger the improvement factor. For example, consider \(m = 2\) case again, but rather than \(\eta_{i,1} = \eta_{i,2}\) we have highly correlated \([\eta_{i,1}, \eta_{i,2}]\) with covariance matrix \(\Sigma_* = \begin{bmatrix} 1 & 1 - \epsilon \\ 1 - \epsilon & 1 \end{bmatrix}\). So, \(\Sigma_*^{-1} = \frac{1}{\pi \epsilon^2} \begin{bmatrix} 1 & -1 + \epsilon \\ -1 + \epsilon & 1 \end{bmatrix}\). The improvement factor becomes
Let \( D \) is in itself interesting for several applications, such as query-document rankings. For both MR and SUR models can be thought of as special cases of the pooled model. Moreover, the and tasks (hence the name “pooled” [7]). It may seem that the model is very restrictive compared to our problems. We first consider a pooled model where a single coefficient vector is used across all data points.

2 The Pooled Model

We first consider a pooled model where a single coefficient vector is used across all data points and the problem as a large regression problem with \( 2 \) The Pooled Model

We first consider a pooled model where a single coefficient vector is used across all data points.
Algorithm 1 AltMin-Pooled: Alternating Minimization for the Pooled Model

| **Require**: $D = \{(X_1, y_1) \ldots (X_{2nT}, y_{2nT})\}$, Number of iterations: $T$
|---|
| 1: Randomly partition $D = \{D^0_w, D^1_w, D^2_w, \ldots, D^T_w, D^T_w\}$, where $|D^t_w| = |D^T_w| = n, \forall t$
| 2: Initialize $w_0 = 0$
| 3: for $t = 0, \ldots, T - 1$ do
| 4: Covariance Estimation: $\hat{\Sigma}_t = \frac{1}{n} \sum_{i \in D^t_w} (y_i - X_iw_t)(y_i - X_iw_t)^T$
| 5: Least-squares Solution: $w_{t+1} = \arg\min_w \frac{1}{n} \sum_{i \in D^t_w} \|\hat{\Sigma}_t^{-\frac{1}{2}}(y_i - X_iw)\|_2^2$
| 6: end for
| 7: Output: $w_T$

Ideally, if $\Sigma_*$ was known, we would like to estimate $w_*$ by decorrelating the noise\(^1\). That is,

$$w_{MLE} = \arg\min_w \frac{1}{n} \sum_{i=1}^n \|\Sigma_*^{-\frac{1}{2}}(y_i - X_iw)\|_2^2. \quad (3)$$

However, $\Sigma_*$ is not known apriori and in general can only be estimated if $w_*$ is known. To avoid this circular requirement, we can jointly estimate $(w_*, \Sigma_*)$ by maximizing the joint likelihood. The joint maximum likelihood estimation (MLE) problem for $(w, \Sigma)$ is given by:

$$(\hat{w}, \hat{\Sigma}) = \arg\max_{w, \Sigma \geq 0} -\log |\Sigma| - \frac{1}{n} \sum_{i=1}^n (y_i - X_iw)^T \Sigma^{-1} (y_i - X_iw). \quad (4)$$

The problem above is non-convex in $(\Sigma, w)$ jointly, and hence standard convex optimization techniques do not apply to the problem. A straightforward heuristic approach is to use alternating minimization (AltMin) where we alternately solve for $\hat{w}$ and $\hat{\Sigma}$ while keeping $\hat{\Sigma}$ and $\hat{w}$ fixed. Note that, each of the above mentioned individual problems are fairly straightforward and can be solved efficiently (see Steps 4, 5 of Algorithm 1). Despite its simplicity and availability of optimal solutions at each iteration, AltMin need not converge to a global optima of the joint problem. Below, we show that despite non-convexity of (4), we can still show that the AltMin procedure has a matching error bound when compared to the optimal MLE solution.

Specifically, we analyze Algorithm 1 which is just the standard AltMin method but uses fresh samples $(y, X)$ for each of the covariance estimation and the least squares step. Practical systems do not perform such re-sampling, but fresh samples at every iteration ensure that errors do not get correlated in adversarial fashion and allows us to use standard concentration bounds. Moreover, since we show convergence at a geometric rate, the number of iterations is not large and hence the sample complexity does not increase by a significant factor.

To prove our convergence results, we require the probability distribution $P_X$ to be a sub-Gaussian distribution with the sub-Gaussian norm $(\|X\|_{\psi_2})$ defined as:

**Definition 1.** Let $X \in \mathbb{R}^{m \times d}$ be a random variable (R.V.) with distribution $P_X$. Then, the sub-Gaussian norm of $X$ is given by:

$$\|X\|_{\psi_2} = \max_{u, \|u\|_2 = 1} \|v^T \Sigma_{Xu}^{-\frac{1}{2}} X^T u\|_{\psi_2}, \text{ where, } \Sigma_{Xu} = \mathbb{E}_{X \sim P_X}[X^T uu^T X].$$

Sub-Gaussian norm of a univariate variable $Q$ is defined as: $\|Q\|_{\psi_2} = \max_{p \geq 1} \frac{1}{\sqrt{p}} \cdot \mathbb{E}[|Q|^p]^\frac{1}{p}$. If $\Sigma_{Xu}$ is not invertible for any fixed $u$ then, we define $\|X\|_{\psi_2} = \infty$.

We pre-multiply $X^T u$ by $\Sigma_{Xu}^{-\frac{1}{2}}$ for normalization, so that for Gaussian $X$, $\|X\|_{\psi_2} = 1$. For bounded variables $X$, s.t., each entry $|X_{ij}| \leq M$, we have: $\|X\|_{\psi_2} \leq M \sqrt{md} \cdot \max_{u, \|u\|_2 = 1} \|\Sigma_{Xu}^{-1}\|_2$.

**Theorem 2** (Result for Pooled Model). Let $X_i \overset{i.i.d.}{\sim} P_X, 1 \leq i \leq n$ with sub-Gaussian norm $\|X_i\|_{\psi_2} < \infty$ and $\eta_i \overset{i.i.d.}{\sim} N(0, \Sigma_*)$ are independent of $X_i$’s. Let $w_* \in \mathbb{R}^d$ be a fixed vector and

---

\(^1\)For simplicity of exposition, throughout the remaining paper, we assume that $\Sigma_*$ is invertible. Non-invertible $\Sigma_*$ can be handled using simple limit arguments and in fact, our results get significantly better if $\Sigma_*$ is not invertible.
n \geq C \cdot (m + d)\|X\|_{\psi_2}. Then, the output \(w_T\) of Algorithm 7 satisfies (w.p. \(\geq 1 - \frac{T}{n^{10}}\)):

\[
E_{X \sim P_X} \left[ \|X(w_T - w_*)\|^2 \right] \leq \frac{Cd \log n}{n} + \frac{1}{\lambda_{\min}} + \frac{\lambda_{\max}^2}{\lambda_{\min}} 2^{-T},
\]

where \(\lambda_{\min} = \lambda_{\min}(\Sigma_*), \lambda_{\max} = \lambda_{\max}(\Sigma_*), \) and \(\Sigma_* = E_{X \sim P_X} \left[ \Sigma_X^{-\frac{1}{2}} X^T \Sigma_*^{-1} X \Sigma_X^{-\frac{1}{2}} \right]. \) Also, \(\Sigma_X = E_{X \sim P_X} \left[ X^T X \right]\) is the covariance of the regressors.

**Remarks:** Using Theorem 15, we also have the following bound for the OLS solution:

\[
E_{X \sim P_X} \left[ \|X(w_{OLS} - w_*)\|^2 \right] \leq \frac{C \cdot d \log n}{n} \cdot \|\Sigma_*\|_2.
\]

The above bound for OLS can be shown to be tight as well (up to log factors) by selecting each \(X_i\) to be sampled i.i.d. from a Gaussian distribution. Let \(w_{\text{MLE}}\) be the MLE solution stated above are in fact tight up to log factors.

**Lemma 3.** Assume the notation of Theorem 2. Let \(w_{t+1}\) be the \((t+1)\)-th iterate of Algorithm 7, then the following holds w.p. \(\geq 1 - 1/n^{10}\):

\[
E_{X \sim P_X} \left[ \|\Sigma_*^{-\frac{1}{2}} X(w_{t+1} - w_*)\|^2 \right] \leq \frac{2Cd \log n}{n} + \frac{1}{2} \cdot E_{X \sim P_X} \left[ \|\Sigma_*^{-\frac{1}{2}} X(w_t - w_*)\|^2 \right].
\]

See Appendix C for detailed proofs of both of the results given above.

### 2.1 Gaussian \(X\): Independent Rows

We first consider a special case where each row of \(X\) is sampled i.i.d. from a Gaussian distribution. That is,

\[
X_i \overset{i.i.d.}{\sim}\mathcal{N}(0, \Lambda), \quad \forall 1 \leq i \leq n, \ 1 \leq j \leq m,
\]

where \(\Lambda \succ 0 \in \mathbb{R}^{d \times d}\) is a covariance matrix and \(\Sigma_X = E_{X \sim P_X} \left[ X^T X \right] = m \cdot \Lambda. \) Let \(\Sigma_* = \sum_{i=1}^{m} \sigma_i (\Sigma_*) u_i u_i^T\) be the eigenvalue decomposition of \(\Sigma_*\). Then,

\[
\Sigma_{X*} = E_{X \sim P_X} \left[ \Sigma_*^{-\frac{1}{2}} X^T \Sigma_*^{-1} X \Sigma_*^{-\frac{1}{2}} \right] = \sum_j E_{X \sim P_X} \left[ \Sigma_*^{-\frac{1}{2}} X^T u_j u_j^T X \Sigma_*^{-\frac{1}{2}} \right] \frac{\sigma_j (\Sigma_*)}{\lambda_i (\Sigma_*)} = \frac{\text{tr}(\Sigma_*)^{-1}}{m} I_{d \times d}.
\]

We now combine the above given observation with Theorem 2 to obtain our error bound for AltMin procedure. Using a slightly stronger version of Theorem 15, we can also obtain the error bound for the OLS (Ordinary Least Squares) solution as well the MLE solution.

**Corollary 4** (Result for Pooled Model, Gaussian Data, Independent Rows). Let \(X_i\) be sampled s.t. each row \(X_i \overset{i.i.d.}{\sim}\mathcal{N}(0, \Lambda)\) and \(\Lambda \succ 0. \) Also, let \(y_i = X_i w_* + \eta_i, \) where \(\eta_i \sim \mathcal{N}(0, \Sigma_*), \) \(\Sigma_* \succ 0. \) Let \(n \geq C(m + d) \log(m + d). \) Then, the OLS solution (2) and the MLE solution (3) has the following error bounds (w.p. \(\geq 1 - 1/n^{10}\)):

\[
E_{X} \left[ \|X(w_{OLS} - w_*)\|^2 \right] \leq \frac{Cd \log n}{n} \cdot \frac{\text{tr}(\Sigma_*)}{m}, \quad E_{X} \left[ \|X(w_{MLE} - w_*)\|^2 \right] \leq \frac{Cd \log n}{n} \cdot \frac{m}{\text{tr}(\Sigma_*)}.
\]

Moreover, the output \(w_T\) \((T = \log \frac{1}{\epsilon})\) of Algorithm 7 satisfies (w.p. \(\geq 1 - T/n^{10}\)):

\[
\text{Error}_T = E_{X \sim P_X} \left[ \|X(w_T - w_*)\|^2 \right] \leq \frac{8Cd \log n}{n} \cdot \frac{m}{\text{tr}(\Sigma_*)^{-1}} + \epsilon.
\]

**Lower Bound for OLS and MLE:** We now show that the error bounds for both the OLS as well as the MLE solution stated above are in fact tight up to log-factors.

**Lemma 5.** Let the assumptions of Corollary 4 hold. Then, we have (w.p. \(\geq 1 - 1/n^{10} - \exp(-d)\)):

\[
E_{X} \left[ \|X(w_{OLS} - w_*)\|^2 \right] \geq \frac{Cd}{n} \cdot \frac{\text{tr}(\Sigma_*)}{m}, \quad E_{X} \left[ \|X(w_{MLE} - w_*)\|^2 \right] \geq \frac{Cd}{n} \cdot \frac{m}{\text{tr}(\Sigma_*)^{-1}},
\]

where \(C > 0\) is a universal constant.
Remarks: As mentioned in the introduction, \( \frac{m}{\text{tr}(\Sigma_*)} \leq \frac{\text{tr}(\Sigma_*)}{m} \) and the gap becomes larger as \( \Sigma_* \) moves away from \( c \cdot I \). Hence, in the light of the above two lower-bound and upper-bound results, it is clear that AltMin (and MLE) solutions are significantly more accurate than OLS, especially for highly correlated noise vectors. This claim is also borne out from our simulation results (Figure 4).

2.2 Gaussian \( X \): Dependent Rows

We now generalize the above given special case by removing the row-wise independence assumption. That is, \( X = \Sigma_R^{1/2} Z \Lambda^{1/2} \), where \( Z_{ij} \sim \mathcal{N}(0, 1) \) \( \forall \ i,j \) and \( \Sigma_R \in \mathbb{R}^{m \times m} \), \( \Lambda \in \mathbb{R}^{d \times d} \) are the row and the column correlation matrices, respectively. It is easy to see that (see Claim [15]),

\[
\Sigma_* = \mathbb{E}_{X \sim \mathcal{P}_X} [\Sigma_X^{-1/2} X^T \Sigma_*^{-1} X \Sigma_X^{-1/2}] = \frac{\text{tr}(\Sigma_R \Sigma_*^{-1})}{\text{tr}(\Sigma_R)}, \quad I_{d \times d}, \quad \text{where } \Sigma_X = \text{tr}(\Sigma_R) \cdot \Lambda.
\]

Using Theorem 2 with Theorem 15 and (32) (with certain \( A, B \)) we obtain the following corollary.

**Corollary 6** (Result for Pooled Model, Gaussian Data, Dependent Rows). Let \( X_i \) be as defined above. Let \( n \geq C(m + d) \log(m + d) \). Then the following holds (w.p. \( \geq 1 - T/n^{10} \)):

\[
\mathbb{E}_{X \sim \mathcal{P}_X} [\|X(w - w_*)\|^2] \leq \frac{8Cd\log n}{n} \cdot \frac{m}{\text{tr}(\Sigma_R \Sigma_*^{-1})} + \epsilon,
\]

where \( w_T \) is the output of Algorithm 1 with \( T = \log \frac{1}{\epsilon} \).

Similarly, bound for OLS is given by: \( \mathbb{E}_{X \sim \mathcal{P}_X} [\|X(w_{OLS} - w_*)\|^2] \leq \frac{Cd\log n \cdot \text{tr}(\Sigma_X) \cdot m}{\text{tr}(\Sigma_R)^2} \). Here again, it is easy to see that the output of AltMin is significantly more accurate than the OLS solution. \( \Sigma_R \) also plays a critical role here. In fact, if \( \Sigma_R \) is nearly orthogonal to \( \Sigma_*^{-1} \), then the gain over OLS is negligible. To understand this better, consider the following 2-task example:

\[
y_i = \langle x_i, w_* \rangle + \eta_i, \quad y_i = \langle x_i, w_* \rangle + \eta_i.
\]

Note that the noise \( \eta_i \) is perfectly correlated here. However, as rows \( X_i \) are also completely correlated. So, the two equations are just duplicates of each other and hence, AltMin cannot obtain any gains over OLS (as predicted by our bounds as well).

3 Seemingly Unrelated Regression

Seemingly-unrelated regression (SUR) model [21, 22] is a generalization of the basic linear regression model to handle vector valued outputs and has applications in several domains including multi-task learning, economics, genomics etc. Below we present the SUR model and our main result for estimating the coefficients in such a model.

Let \( X_i \in \mathbb{R}^{m \times d} \), \( 1 \leq i \leq n \) be sampled i.i.d. from a fixed distribution \( \mathcal{P}_X \). Let \( W_* \in \mathbb{R}^{m \times d} \) be a fixed matrix of coefficients. The vector-valued output for each data point \( X_i \) is given by:

\[
y_i = X_i \cdot W_* + \eta_i,
\]

where \( X_i \cdot W_* = \text{diag}(X_i W_*^T) \) and \( \eta_i \sim \mathcal{N}(0, \Sigma_*) \) is the noise vector with covariance \( \Sigma_* \).

OLS and MLE solution can be defined similar to the Pooled model:

\[
W_{OLS} = \arg \min_W \frac{1}{n} \sum_{i=1}^{n} \|y_i - X_i \cdot W\|^2, \quad W_{MLE} = \arg \min_W \frac{1}{n} \sum_{i=1}^{n} \|\Sigma_*^{-1/2} (y_i - X_i \cdot W)\|^2.
\]

Here again, we expect MLE to provide significantly better estimation of \( W_* \) by exploiting noise correlation. As \( \Sigma_* \) is not available apriori, both \( \Sigma_* \) and \( W_* \) are estimated by solving the following MLE problem:

\[
(\hat{W}, \hat{\Sigma}) = \arg \max_{W, \Sigma} - \log |\Sigma| - \frac{1}{n} \sum_{i=1}^{n} \|\Sigma_*^{-1/2} (y_i - X_i \cdot W)\|^2
\]

Here again, the MLE problem is non-convex and hence standard analysis does not provide strong convergence guarantees. Still, alternating minimization (of negative log-likelihood) for \( \hat{W}, \hat{\Sigma} \) leads
Moreover, if

\[ P \]

for

Clearly, the error of AltMin decreases to

\[ X \]

example to demonstrate prediction accuracy of AltMin (and MLE) solution vs. OLS solution.

The weight for each individual error term

\[ (\text{Result for SUR Model}) \]

Theorem 7

show that the finite sample error bound of AltMin matches (up to logarithmic factors) the error rate of the MLE solution. Similar to the previous section, we modify the standard AltMin procedure to include fresh samples at each step of the algorithm.

### Theorem 7 (Result for SUR Model)

Let \( X_i \overset{i.i.d.}{\sim} \mathcal{P}_X, 1 \leq i \leq n \), where \( \|X\|_{\psi_2} \) is the sub-Gaussian norm of each \( X_i \). Let \( \eta_i \sim \mathcal{N}(0, \Sigma_i) \), \( \Sigma_i > 0 \), and \( W_e \in \mathbb{R}^{m \times d} \) be a fixed coefficients matrix. Let \( W_T^* \) be the \( T \)-th iterate of Algorithm 2.

Also, let \( n \geq C \cdot md \|X\|_{\psi_2} \), where \( C > 0 \) is a global constant. Then, the following holds (w.p. \( \geq 1 - T/n^{10} \)):

\[
\mathbb{E}_{X \sim \mathcal{P}_X} \left[ \left\| \Sigma_i^{-\frac{1}{2}} (X \bullet W - X \bullet W_e) \right\|_2^2 \right] \leq \frac{4C^2d \log(n)}{n} \cdot m + \| \Sigma_i \|_2^2 \cdot 2^{-T}.
\]

Moreover, if \( \mathcal{P}_X \) is such that each row \( X^p \) is sampled independently and has zero mean, i.e., \( X^p \perp X^q, \forall p, q \) and \( \mathbb{E}_{X \sim \mathcal{P}_X} [X] = 0 \), then the following holds (w.p. \( \geq 1 - T/n^{10} \)):

\[
\sum_{j=1}^{m} (\Sigma^{-1})_{jj} \mathbb{E}_{X \sim \mathcal{P}_X} \left[ (X^j, W^j_T - W^j_e)^2 \right] \leq \frac{4C^2d \log(n)}{n} \cdot m + \| \Sigma_i \|_2^2 \cdot 2^{-T}.
\]

### Remarks

It is easy to obtain error bounds for OLS in this case as it solves each equation independently. In particular, standard single-output linear regression analysis [23] gives:

\[
\mathbb{E}_X \left[ \sum_j \frac{1}{(\Sigma_i)_{jj}} (W^j_{OL} - W^j_e, X^j)^2 \right] \leq \frac{Cd}{n} \cdot m.
\]

The weight for each individual error term \( (W^j_T - W^j_e, X^j)^2 \) in the AltMin error bound is \( (\Sigma^{-1})_{jj} \) while it is \( \frac{1}{(\Sigma_i)_{jj}} \) for OLS. Using Claim 20, \( (\Sigma^{-1})_{jj} \geq \frac{1}{(\Sigma_i)_{jj}} \). Hence, the error terms of AltMin should be significantly smaller than that of OLS. Similar to Section 1, we now provide an illustrative example to demonstrate prediction accuracy of AltMin (and MLE) solution vs. OLS solution.

### Illustrative Example

Consider a two-valued output SUR problem, where \( X^1 \sim \mathcal{N}(0, I_{d \times d}) \) and \( X^2 \sim \mathcal{N}(0, I_{d \times d}) \) are sampled independently and the noise covariance by:

\[
\Sigma_i = \begin{bmatrix}
1 & 1 - \epsilon \\
1 - \epsilon & 1
\end{bmatrix}, \quad \Sigma^{-1} = \frac{1}{2\epsilon} \begin{bmatrix}
1 & \epsilon \\
\epsilon & 1
\end{bmatrix},
\]

where \( 0 < \epsilon < 1 \). We sample \( n \geq 2d \) points from this model. That is, \( y_i = X_i \bullet W_e + \eta_i, 1 \leq i \leq n \).

Hence, the estimation error of AltMin and OLS are given by:

\[
\|W_T - W_e\|_F^2 \leq \frac{4C^2d \log n}{n} \cdot \epsilon, \quad \|W_{OL} - W_e\|_F^2 \leq \frac{2Cd}{n}.
\]

Clearly, the error of AltMin decreases to 0 as \( \epsilon \to 0 \) (and \( n \geq Cd \log d \)), i.e., as noise is getting more correlated. In contrast, the error bound of OLS is independent of \( \epsilon \) and remains a large constant even for \( \epsilon = 0 \) and \( n = O(d) \).

### Multivariate Regression Model

We now briefly discuss the popular Multivariate Regression (MR) model (arises in many applications including multitask learning with shared regressors), where each output \( y_i \in \mathbb{R}^m \) is modeled as:

\[
y_i = W_e x_i + \eta_i,
\]

(9)
where $x_i \in \mathbb{R}^d$ are all sampled i.i.d. from a fixed distribution and $\eta_i \sim \mathcal{N}(0, \Sigma_\epsilon)$ is the noise vector. $W_* \in \mathbb{R}^{m \times d}$ is the coefficients matrix. Model (9) can be easily re-written as a SUR problem (5) where each row of $X_i$ is given by $x_i^T$. That is, $X_i = [x_i \mid x_i \mid \ldots \mid x_i]^T$. However, for the MR model, it is well known that the optimal solution to MLE problem is same as the OLS solution [6, Chapter 7]. Naturally, our AltMin/MLE error bounds also do not provide an advantage over OLS bounds. For general $W_*$ in the MR model, the MLE solution is independent of $\Sigma_\epsilon$. But, by imposing certain special structures on $W_*$, MLE indeed leads to significantly more accurate solutions. For example, the Pooled and the SUR model can be posed as special cases of the MR model but with specially structured $W_*$. Similarly, other structures like reduced rank regression [24,25] also allows exploitation of the noise correlation. We leave further investigation of other type of structural assumptions on $W_*$ as a topic of future research.

4 Experiments

In this section, we present results from simulations which were conducted with the following two-fold objective: a) demonstrate that both MLE and AltMin estimators indeed perform significantly better than the Ordinary Least Squares (OLS) estimator when the noise vector has significant dependencies, b) study scaling behavior of the three estimators (OLS, MLE, AltMin) w.r.t. $m, n$.

Solving MLE for the Pooled as well as SUR model is difficult in general. So, we set $\Sigma = \Sigma_\epsilon$ in the MLE optimization (4 for Pooled, 7 for SUR), where $\Sigma_\epsilon$ is the true noise covariance matrix. In this case, the estimator reduces to a least squares problem. We implemented all the three estimators in Matlab and provide results averaged over 20 runs. We run AltMin for at most 50 iterations.

Pooled Model: For the first set of experiments, we generated the data $(X_i, y_i)$ using the Pooled Model (Section 2). We generated $X_i$’s from spherical multi-variate Gaussian and selected $w_*$ to be a random vector. In the first sub-experiment, we considered $m = 2$ and set $\Sigma_\epsilon = \begin{bmatrix} 1 & 1 - \epsilon \\ 1 - \epsilon & 1 \end{bmatrix}$.

Using Theorem 2, AltMin as well as MLE estimator should have error $\|w - w_*\|_2^2 \leq \frac{Cd}{n} \cdot (\epsilon - \epsilon^2)$ while for OLS it is $\|w - w_*\|_2^2 \leq \frac{Cd}{n}$. Figure 4(a) shows that our simulations also exhibit exactly the same trend as predicted by our error bounds. Moreover, errors of both MLE and AltMin are exactly the same, indicating that AltMin indeed converged to the MLE estimate.

Next, we set $\Sigma_\epsilon$ as:

$$\Sigma_\epsilon = \begin{bmatrix} 1 & 1 - \epsilon & 0 \\
1 - \epsilon & 1 & 0 \\
0 & 0 & I_{m-2 \times m-2} \end{bmatrix},$$

with $\epsilon = 0.005$ and measure recovery error $\|w - w_*\|_2$ while varying $m$ and $n$. Note that for AltMin and MLE, the bound for such $\Sigma_\epsilon$ is $\|w - w_*\|_2^2 \leq \frac{Cd}{n} \cdot \frac{\epsilon - \epsilon^2}{(m-2)(\epsilon - \epsilon^2) + 1}$ and hence AltMin and MLE’s error does not change significantly with increasing $m$. But for OLS the error goes down with $m$ as $\|w - w_*\|_2^2 \leq \frac{Cd}{m} \cdot \frac{1}{m}$ which can be observed in the Figure 4(b) as well. Finally, Figure 4(c) clearly indicates that $\|w - w_*\| = O\left(\frac{1}{\sqrt{m}}\right)$ for all the three methods, hence matching our theoretical bounds.

SUR Model: Here we generated data $(X_i, y_i)$ using the SUR model (Section 3) but with $X_i$ sampled from spherical Gaussians. $W_*$ was selected to be a random Gaussian matrix, $\Sigma_\epsilon$ is given by (10). As illustrated in Section 3, the error of MLE/AltMin is at most $O(\epsilon)$ while the error of OLS is independent of $\epsilon$. Figure 4(d) clearly demonstrates the above mentioned error trends.
References


A High-dimensional Setting

We now study the vector-output regression problems in the high-dimensional setting, where $d \gg n$ but the parameter vector is required to be $s$-sparse with $s \ll d$. Our goal is to provide an algorithm with error bounds that are at most logarithmic in $d$ and are linear in $s$. Here, we provide our result for a special case of the Pooled model, where data is sampled from a Gaussian distribution. Our analysis can be easily extended to the SUR model as well.

Let $X_i = \Sigma_i^{1/2} Z_i A^{1/2}$ where each entry of $Z_i$ is sampled i.i.d. from the univariate normal distribution and $\Sigma_R \succ 0$, $A \succ 0$. Let $w_s \in \mathbb{R}^d$ be such that $w_s$ is $s$-sparse, i.e., $\|w_s\|_0 \leq s$. The outputs are given by, $y_i = X_i w_s + \xi_i$, $\xi_i \sim \mathcal{N}(0, \Sigma_s)$, where $\Sigma_s \succ 0$.

For the above setting, we analyze Algorithm 1 but where Least Squares Estimation step (Step 5) is replaced by sparsity constrained optimization:

$$\hat{w} = \arg\min_{\|w\|_0 \leq s} f(w) = \arg\min_{\|w\|_0 \leq s} \frac{1}{n} \sum_{i \in D^w} \|\tilde{\Sigma}_i^{-\frac{1}{2}} (y_i - X_i w)\|_2^2 \quad (11)$$

Note that the above problem is in general NP-hard to solve due to the sparsity constraint. But, we can use the Iterative Hard Thresholding (IHT) method \cite{blumensath2009iterative} to solve (11), if $f(w)$ satisfies the restricted strong smoothness (RSS) and the restricted strong convexity (RSC) properties (defined in (12)). Below, we re-state the IHT convergence result by by \cite{blumensath2009iterative}.

**Theorem 8 (Theorem 1 of \cite{blumensath2009iterative}).** Let $f$ have RSC and RSS parameters given by $L_{\Sigma_s^2} = L$ and $\alpha_{3\Sigma} = \alpha$ respectively. Let IHT algorithm (Algorithm 1, \cite{blumensath2009iterative}) be invoked with $f$, $\bar{s} = \kappa^2 \cdot s$. Then, the $\tau$-th iterate of IHT ($w_{T+1}$), for $\tau = O(\frac{L}{\alpha} \cdot \log \frac{\|w_s\|_2}{\epsilon})$ satisfies: $f(w_{T+1}) \leq f(\hat{w}) + \epsilon$, where $\hat{w}$ is any global optimum of (11).

As the algorithm has only logarithmic dependence on $\epsilon$, we can set $\epsilon$ to be arbitrary small (say $0.001 f(\hat{w})$). For simplicity, we ignore $\epsilon$ for now. Note that the proof of Lemma \cite{candes2008enhancing} only requires that the least squares step satisfies: $\hat{f}(w_{T+1}) \leq f(w_s)$. Moreover, columns of $X$ corresponding to the index set $S_{T+1} = \text{supp}(w_1) \cup \text{supp}(w_{T+1}) \cup \text{supp}(w_s)$ are used by $\tilde{\Sigma}_T$ and the least squares solution. So, Lemma \cite{candes2008enhancing} applies directly but with $d = |S_{T+1}| \leq 3\bar{s}^2$.

Hence, we obtain the following error bound for the $T$-th iterate of Algorithm 1:

$$E_{X,S \sim P_X} \left\| X(w_T - w_s) \right\|_2^2 \leq \frac{8C\bar{s} \log d}{n} \cdot \frac{m}{\text{tr}(\Sigma_R \Sigma_s^{-1})^2} + 2^{-T},$$

Recall that $\bar{s} = \left( \frac{L}{\alpha} \right)^{1/2} \cdot s$, so that $L, \alpha$ are the RSS and the RSC constants of $f$. Hence, we now only need to provide RSS/RSC constants for the above given $f$.

**Lemma 9 (RSC/RSS).** Let $X_i$ be as given above. Also, let $n \geq C\bar{s} \log d$. Then the following holds for all fixed $A$ (w.p. $\geq 1 - \exp(-n)$):

$$0.5 \cdot \lambda_{\min}(A) \text{tr}(A^T \Sigma_R) \|v\|_2^2 \leq \frac{1}{n} \sum_{i=1}^n v^T X_i^T A X_i v \leq 2 \text{tr}(A^T \Sigma_R) \cdot \lambda_{\max}(A) \|v\|_2^2,$$

where $v \in \mathbb{R}^d$ is any $\bar{s}$-sparse vector.

The above lemma shows that for any $\bar{s}$-sparse $w, w'$, we have:

$$\frac{\alpha_{3\Sigma}}{2} \|w - w'\|^2_2 + \langle \nabla f, w - w' \rangle + f(w') \leq f(w) \leq f(w') + \langle \nabla f, w - w' \rangle + \frac{\alpha_{3\Sigma}}{2} \|w - w'\|^2_2, \quad (12)$$

where $L = L_{3\Sigma} = 2\lambda_{\max}(A)$ is the RSS constant of $f$ and $\alpha = \alpha_{3\Sigma} = \frac{\lambda_{\min}(A)}{2}$ is the RSC constant.

That is the error bound for AltMin procedure is given by:

$$E_{X,S \sim P_X} \left\| X(w_T - w_s) \right\|_2^2 \leq \frac{8C\bar{s} \log d}{n} \cdot \left(\frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}\right)^2 \cdot \frac{m}{\text{tr}(\Sigma_R \Sigma_s^{-1})^2} + 2^{-T}. \quad (13)$$

\footnotetext{For simplicity, we ignore a technicality regarding assuming that $S_{T+1}$ is a fixed set. The assumption can be easily removed by taking a union bound over all sets of size $3\bar{s}$.}

10
Proof of Lemma 9 Consider a “fixed” support set $S$ s.t. $|S| = \hat{s}$. Also, let $\Sigma_R^{1/2} A^T A \Sigma_R^{1/2} = \sum_j \lambda_j u_j u_j^T$ be the eigenvalue decomposition of $A^T A$. Then, w.p. $\geq 1 - \exp(-n)$, the following holds for all $v \in \mathbb{R}^d$ s.t. $|v|_0 \leq \hat{s}$ and $\text{supp}(v) \subseteq S$:

$$
\mathbf{v}^T \left( \sum_{i=1}^n X_i A^T A X_i v \right) = \sum_{j=1}^m \lambda_j \mathbf{v}_S^T \left( \sum_{i=1}^n (X_i^T)^S u_j (X_i)_S \right) \mathbf{v}_S,
$$

$$
= \sum_{j=1}^m \lambda_j \mathbf{v}_S^T \Lambda_S^{1/2} \left( \sum_{i=1}^n \Lambda_i^{-1/2}(X_i)_S u_j (X_i)_S \Lambda_i^{-1/2} \right) \Lambda_S^{1/2} \mathbf{v}_S,
$$

$$
\geq \frac{\zeta_1}{2} \sum_{j=1}^m \lambda_j \mathbf{v}_S^T \Lambda_S \mathbf{v}_S = \frac{n}{2} \text{tr}(A^T A \Sigma_R) v^T A v,
$$

where $\zeta_1$ follows by Lemma 10 $(X_i)_S$ denotes the submatrix of $X_i$ corresponding to index set $S$. Similarly, $\mathbf{v}_S$ and $\Lambda_S$ can be defined to be sub-vector and sub-matrix of $\mathbf{v}$ and $\Lambda$, respectively.

The lower bound of the theorem follows by taking a union bound on all $O(d^2)$ sets $S$ and setting $n \geq C \hat{s} \log d$.

The upper bound on $\mathbf{v}^T \left( \frac{1}{n} \sum_{i=1}^n X_i^T A^T A X_i \right) v$ also follows similarly. \qed

B Technical Lemmas

Lemma 10. Let $z_i \overset{i.i.d.}{\sim} \mathcal{P}_z, 1 \leq i \leq n$, where $\mathcal{P}_z$ is such that $E_{z \sim \mathcal{P}_z} [zz^T] = I_d \times d$ and the sub-Gaussian norm of $z$ is given by $\|z\|_{\psi_2}$. Let $n \geq C d \cdot \|z\|_{\psi_2}$. Then, the following holds w.p. $\geq 1 - \exp(-C \cdot n)$:

$$
\left\| \frac{1}{n} \sum_{i=1}^n z_i z_i^T - I_d \times d \right\|_2 \leq \frac{1}{10}.
$$

Proof. Lemma follows directly by Corollary 5.50 of [28]. \qed

Lemma 11 (Corollary 5.35 of [28]). Let $M \in \mathbb{R}^{m \times n}$ be s.t. $M_{ij} \overset{i.i.d.}{\sim} \mathcal{N}(0, 1)$, $\forall i, j$. Also, let $n \geq 4m$, then the following holds w.p. $\geq 1 - \exp(-C \cdot n)$:

$$
\frac{1}{2} \sqrt{n} \leq \sigma_m(M) \leq \sigma_1(M) \leq 2 \sqrt{n},
$$

where $\sigma_i(M)$ is the $i$-th singular value of $M$.

Lemma 12. Let $g = \mathbf{A} \eta$, where $\eta \sim \mathcal{N}(0, I_{n \times n}) \in \mathbb{R}^n$ and $A \in \mathbb{R}^{m \times n}$ is a fixed matrix independent of $\eta$. Also, let $n \geq C m$. Then, w.p. $\geq 1 - 1/n^{10}$, we have:

$$
\|g\|_2^2 \leq C \|A\|_F^2 \log(n).
$$

Proof. First consider the $j$-th coordinate of $g_j = e_j^T \mathbf{A} \eta$. As $\eta$ is a Gaussian vector and $A$ is a fixed matrix, $g_j \sim \|e_j^T A\|_2 \cdot \mathcal{N}(0, 1)$. Hence, w.p. $1 - 1/n^{11}$, $g_j \leq C \|e_j^T A\|_2 \sqrt{\log n}$. Lemma now follows by combining the above observation with $\|g\|_2^2 = \sum_j (e_j^T \mathbf{A} \eta)^2$ and the union bound. \qed

Lemma 13. Let $g = \sum_i A_i \eta_i$, where $A_i \in \mathbb{R}^{d \times m}$, $\eta_i \overset{i.i.d.}{\sim} \mathcal{N}(0, I_{m \times m}) \in \mathbb{R}^m$. Also, let $n \geq C m$. Then, w.p. $\geq 1 - 1/n^{10}$, we have:

$$
\|g\|_2^2 \leq C \left( \sum_{i=1}^n \|A_i\|_F^2 \right) \log(n).
$$
Theorem 15. Let $X_i \sim \mathcal{P}_X, 1 \leq i \leq n$ be sub-Gaussian random variables. Also, let $n \geq C d \|X\|_{\psi_2} \log d$, where $\|X\|_{\psi_2}$ is the sub-Gaussian norm of each $X_i$ (see Definition 7). Let $A$ be any fixed matrix. Then, w.p. $1 - m \exp(-C \cdot n)$, the following holds for all $v \in \mathbb{R}^d$:

$$\frac{1}{2} v^T \left( E_{X \sim \mathcal{P}_X} [X^T A^T AX] \right) v \leq v^T \left( \frac{1}{n} \sum_{i=1}^{n} X_i^T A^T AX_i \right) v \leq 2 v^T \left( E_{X \sim \mathcal{P}_X} [X^T A^T AX] \right) v.$$

Proof. Let $A^T A = \sum_{j} \lambda_j(A^T A) u_j u_j^T$ be the eigenvalue decomposition of $A^T A$. Then, $\forall v \in \mathbb{R}^d$:

$$v^T \sum_{i=1}^{n} X_i A^T AX_i v = \sum_{j=1}^{m} \lambda_j(A^T A) v^T \left( \sum_{i=1}^{n} X_i^T u_j u_j^T X_i \right) v = \sum_{j=1}^{m} \lambda_j(A^T A) v^T \Sigma_{X u_j}^{\frac{1}{2}} \left( \sum_{i=1}^{n} \Sigma_{X u_j}^{-\frac{1}{2}} X_i^T u_j u_j^T X_i \Sigma_{X u_j}^{-\frac{1}{2}} \right) \Sigma_{X u_j}^{\frac{1}{2}} v,$$

where $\Sigma_{X u_j} = E_{X \sim \mathcal{P}_X} [X^T u_j u_j^T X]$. Let $z_{ij} = \Sigma_{X u_j}^{-\frac{1}{2}} X_i^T u_j$. Then, by definition of $\Sigma_{X u_j}$, we have:

$$E[z_{ij} z_{ij}^T] = I_d \times d.$$ 

Moreover, $\|z_{ij}\|_{\psi_2} \leq \|X\|_{\psi_2}$ by definition (see Definition 1). Hence, using Lemma 10 and the union bound for $m$ $u_j$'s (recall that $A$ and hence $u_j$'s are fixed), w.p. $1 - m \exp(-Cn)$ the following holds for all $v \in \mathbb{R}^d$:

$$v^T \sum_{i=1}^{n} X_i A^T AX_i v \geq \frac{n}{2} \sum_{j=1}^{m} \lambda_j(A^T A) v^T \Sigma_{X u_j}^{\frac{1}{2}} v = \frac{n}{2} \sum_{j=1}^{m} \lambda_j(A^T A) v^T E_{X \sim \mathcal{P}_X} [X u_j u_j^T X] v = \frac{n}{2} v^T \left( E_{X \sim \mathcal{P}_X} [X^T A^T AX] \right) v.$$  

The upper bound on $v^T \left( \frac{1}{n} \sum_{i=1}^{n} X_i^T A^T AX_i \right) v$ also follows similarly. \( \square \)

C Proofs of Claims from Section 2

We first provide analysis for a general estimator that decorrelates noise using certain fixed $A, B$ matrices. Our bounds for OLS, MLE follow directly using the below given general theorem.

Theorem 15. Let $X_i \sim \mathcal{P}_X, 1 \leq i \leq n$ where $\mathcal{P}_X$ is a sub-Gaussian distribution with sub-Gaussian norm $\|X\|_{\psi_2} < \infty$ (see Definition 7). Also, let $\eta_i \sim \mathcal{N}(0, I_{m \times m})$. Let $w_\star \in \mathbb{R}^d$ be a fixed weight vector and $A, B$ be fixed matrices. Let,

$$\hat{w} = \arg \min_w \frac{1}{n} \sum_{i=1}^{n} \|AX_i (w - w_\star) - B \eta_i\|_2^2.$$ 

Also, let $n \geq C \cdot (m + d) \log(m + d) \cdot \|X\|_{\psi_2}$, where $C > 0$ is a global constant. Then, the following holds (w.p. $1 - 1/n^{10}$):

$$E_{X \sim \mathcal{P}_X} \left[ \|AX (\hat{w} - w_\star)\|_2^2 \right] \leq \frac{C^2 d \log(n)}{n} \cdot \|B\|_2^2.$$
Proof. As \( \hat{w} \) is the optimal solution to (17), we have:

\[
\sum_{i=1}^{n} \|A_i (\hat{w} - w_*) - B \eta_i \|^2_2 \leq \sum_{i=1}^{n} \|B \eta_i \|^2_2,
\]

\[
\sum_{i=1}^{n} \|A_i (\hat{w} - w_*) \|^2_2 \leq 2(\hat{w} - w_*)^T F^{\frac{1}{2}} F^{-\frac{1}{2}} \sum_{i=1}^{n} X_i^T A^T B \eta_i,
\]

\[
\|F^{\frac{1}{2}} (\hat{w} - w_*) \|^2_2 \leq 2\|F^{\frac{1}{2}} (\hat{w} - w_*)\|_2 \|F^{-\frac{1}{2}} \sum_{i=1}^{n} X_i^T A^T B \eta_i\|_2,
\]  

(18)

where \( F = \sum_{i=1}^{n} X_i^T A^T A X_i \) and \( \zeta_i \) follows from Cauchy-Schwarz inequality. Also, using Lemma 14 \( \lambda_{\min}(F) \geq \frac{1}{2} \lambda_{\min}(\mathbb{E}_{X \sim \mathcal{P}_X} [X^T A^T A X]) \). Using the fact that \( \|X\|_{\infty} < \infty \) (see Definition 1), we have \( \lambda_{\min}(F) \geq \frac{1}{2} \lambda_{\min}(\mathbb{E}_{X \sim \mathcal{P}_X} [X^T A^T A X]) > 0 \). Hence, \( F^{-1/2} \) is well-defined.

Note that \( g = \sum_{i=1}^{n} F^{-\frac{1}{2}} X_i^T A^T B \eta_i \). Using Lemma 13, we have (w.p. \( \geq 1 - 1/n^{10} \)):

\[
\|g\|_2^2 \leq \log n \cdot \sum_{i=1}^{n} \|F^{-\frac{1}{2}} X_i^T A^T B\|_F^2 = \log n \cdot \text{tr}\left( \sum_{i=1}^{n} X_i (F^{-\frac{1}{2}} X_i^T A^T B B^T A)^{-1} \right),
\]

\[
\leq \log n \cdot \|B\|_2^2 \text{tr}\left( F^{-\frac{1}{2}} \sum_{i=1}^{n} X_i^T A^T A X_i \right),
\]

\[
= d \log n \cdot \|B\|_2^2,
\]  

(19)

where the last equality follows from the definition of \( F \).

Now, using Lemma 14 we have (w.p. \( \geq 1 - m \exp(-Cn) \)):

\[
(\hat{w} - w_*)^T \sum_{i=1}^{n} X_i^T A^T A X_i (\hat{w} - w_*) \geq \sum_{i=1}^{n} \frac{n}{2} (\hat{w} - w_*)^T (\mathbb{E}_{X \sim \mathcal{P}_X} [X^T A^T A X]) (\hat{w} - w_*) \].  

(20)

Theorem now follows by combining (18), (19) and (20).

We now provide proofs of both Theorem 2 as well as Lemma 3 which is the key component used by our proof of the main theorem.

Proof of Theorem 2. Theorem follows using Lemma 3 and observing that:

\[
\mathbb{E}_{X \sim \mathcal{P}_X} \left[ \|\Sigma_{\hat{X}}^{-\frac{1}{2}} X (w_T - w_*)\|_2 \right]^2 = \mathbb{E}_{X \sim \mathcal{P}_X} \left[ \|\Sigma_{\hat{X}}^{-\frac{1}{2}} X \Sigma_{\hat{X}}^{-\frac{1}{2}} \Sigma_{\hat{X}}^T (w_T - w_*)\|_2 \right]^2
\]

\[
\geq \lambda_{\min} (\Sigma_{\hat{X}}) \|\Sigma_{\hat{X}}^{\frac{1}{2}} (w_T - w_*)\|_2^2,
\]

\[
= \lambda_{\min}(\Sigma_{\hat{X}}) \|\Sigma_{\hat{X}}^{\frac{1}{2}} (w_T - w_*)\|_2^2,
\]

(21)

where the second inequality follows from the definition of \( \Sigma_{\hat{X}} \) and the last equality follows by using \( \Sigma_{\hat{X}} = \mathbb{E}_{X \sim \mathcal{P}_X} [X^T X] \).

Proof of Lemma 3. Recall that,

\[
\hat{\Sigma}_t = \frac{1}{n} \sum_{i \in D_t^C} (y_i - X_i w_t) (y_i - X_i w_t)^T, \ \Sigma_t = \Delta + \Sigma_*
\]

where \( \Delta = \frac{1}{n} \sum_{i \in D_t^C} X_i (w_* - w_t) (w_* - w_t)^T X_i^T \). Now, using Lemma 14 \( \lambda_{\min}(\hat{\Sigma}_t) \geq \frac{1}{2} \lambda_{\min}(\Delta) > 0 \) as \( \Sigma_* > 0 \). So, \( \hat{\Sigma}_t \) is invertible.

Note that the samples of set \( D_t^C \) are independent of \( w_t \) as well as \( D_t^w \). Moreover, \( D_t^w \) is independent of \( w_t \) and hence is independent of \( \Delta \). Also,

\[
w_{t+1} = \arg \min_{w} \frac{1}{n} \sum_{i \in D_t^w} \|\hat{\Sigma}_t^{-\frac{1}{2}} X_i (w - w_*) - \hat{\Sigma}_t^{-\frac{1}{2}} \eta_i\|_2^2.
\]

(22)
Let $A = \hat{\Sigma}_t^{-\frac{1}{2}}$ and $B = \hat{\Sigma}_t^{-\frac{1}{2}} \Sigma_{\varepsilon}^2$. Note that, $A, B$ are both fixed matrices w.r.t. $D^w_t$ as $\Delta$ itself is independent of $D^w_t$. Now, applying Theorem 15 with the above mentioned $A, B$, we get (w.p. $\geq 1 - 1/n^{10}$):

$$
\mathbb{E}_{X \sim P_X} \left[ \left\| \hat{\Sigma}_t^{-\frac{1}{2}} X (w_{t+1} - w_+) \right\|_2^2 \right] \leq \frac{C^2 d \log(n)}{n} \cdot \left\| \Sigma_{\varepsilon}^2 \hat{\Sigma}_t^{-1} \Sigma_{\varepsilon}^2 \right\|_2.
$$

Now, the following holds (w.p. $\geq 1 - \exp(-cn)$):

$$
\| \Sigma_{\varepsilon}^2 \hat{\Sigma}_t^{-1} \Sigma_{\varepsilon}^2 \|_2 = \max_{v_i \|v_i\|_2 = 1} \left\| v_i^T \Sigma_{\varepsilon}^2 \hat{\Sigma}_t^{-1} \Sigma_{\varepsilon}^2 v_i \right\|_2 \leq \max_{v_i \|v_i\|_2 = 1} 2 \| v_i^T \Sigma_{\varepsilon}^2 \Sigma_{\varepsilon}^{-\frac{1}{2}} \Sigma_{\varepsilon}^{-\frac{1}{2}} v_i \|_2,
$$

where $\zeta_1$ follows from Lemma 16 and the fact that $D^w_t$ is independent of $w_t$. The last inequality follows as $\Delta$ is a p.s.d. matrix, so, $\lambda_{\min} \left( I_{m \times m} + \Sigma_{\varepsilon}^{-\frac{1}{2}} \Delta \Sigma_{\varepsilon}^{-\frac{1}{2}} \right) \geq 1$.

Next, we have:

$$
\mathbb{E}_{X \sim P_X} \left[ \left\| \hat{\Sigma}_t^{-\frac{1}{2}} X (\hat{w} - w_+) \right\|_2^2 \right] \geq \frac{1}{2} \mathbb{E}_{X \sim P_X} \left[ \left\| \hat{\Sigma}_t^{-\frac{1}{2}} X (\hat{w} - w_+) \right\|_2 \right] = \frac{1}{2} \mathbb{E}_{X \sim P_X} \left[ \left\| \Sigma_{\varepsilon}^{-\frac{1}{2}} \Sigma_{\varepsilon}^{-\frac{1}{2}} \Sigma_{\varepsilon}^{-\frac{1}{2}} X (\hat{w} - w_+) \right\|_2 \right],
$$

where $\zeta_1$ follows from Lemma 16 and $\zeta_2$ follows from the definition of $\Sigma_t$, and the last equation follows from $\lambda_{\min} \left( I_{m \times m} + \Sigma_{\varepsilon}^{-\frac{1}{2}} \Delta \Sigma_{\varepsilon}^{-\frac{1}{2}} \right) \geq \frac{1}{\lambda_{\max} \left( I_{m \times m} + \Sigma_{\varepsilon}^{-\frac{1}{2}} \Delta \Sigma_{\varepsilon}^{-\frac{1}{2}} \right)^{-1}}$.

Now,

$$
\lambda_{\max} \left( I_{m \times m} + \Sigma_{\varepsilon}^{-\frac{1}{2}} \Delta \Sigma_{\varepsilon}^{-\frac{1}{2}} \right) = 1 + \left\| \Sigma_{\varepsilon}^{-\frac{1}{2}} \Delta \Sigma_{\varepsilon}^{-\frac{1}{2}} \right\|_2
$$

where $\zeta_1$ follows from Lemma 14.

Using (23), (24), (25), and (26), we have:

$$
\mathbb{E}_{X \sim P_X} \left[ \left\| \Sigma_{\varepsilon}^{-\frac{1}{2}} X (w_{t+1} - w_+) \right\|_2^2 \right] \leq \frac{2C^2 d \log(n)}{n} + \frac{4C^2 d \log(n)}{n} \cdot \mathbb{E}_{X \sim P_X} \left[ \left\| \Sigma_{\varepsilon}^{-\frac{1}{2}} X (w_t - w_+) \right\|_2 \right].
$$

Theorem now follows, as $n \geq 16 C^2 d \log d$.

**Proof of Lemma 5.** We show the error bound for a general estimator:

$$
\hat{w} = \arg \min_w \frac{1}{n} \sum_{i=1}^n \left\| A X_i (w - w_+) - B \eta_i \right\|_2^2,
$$

where $A, B \in \mathbb{R}^{m \times m}$ are fixed p.s.d. matrices.
Now, the optimal solution is given by:

\[
\left( \frac{1}{n} \sum_{i=1}^{n} X_i^T A X_i \right) (\bar{w} - w_*) = \frac{1}{n} \sum_{i=1}^{n} X_i^T A B_i,
\]

\[
\left( \frac{1}{n} \sum_{i=1}^{n} \Lambda^{-\frac{1}{2}} X_i^T A X_i \Lambda^{-\frac{1}{2}} \right) \Lambda^{\frac{1}{2}} (\bar{w} - w_*) = \frac{1}{n} \sum_{i=1}^{n} \Lambda^{-\frac{1}{2}} X_i^T A B_i.
\]  \hspace{1cm} (27)

Note that, \( X_i = Z_i \Lambda^{\frac{1}{2}} \) where \( Z_i \sim N(0, I_{d \times d}) \). Also, let \( A^T A = \sum_j \lambda_j (A^T A) u_j u_j^T \) be the eigenvalue decomposition of \( A^T A \). Then,

\[
\frac{1}{n} \sum_{i=1}^{n} \Lambda^{-\frac{1}{2}} X_i^T A^T A X_i \Lambda^{-\frac{1}{2}} = \sum_{j=1}^{m} \lambda_j (A^T A) \left( \frac{1}{n} \sum_{i=1}^{n} Z_i^T u_j u_j^T Z_i \right).
\]  \hspace{1cm} (28)

Also, note that \( Z_i^T u_j \sim N(0, I_{d \times d}) \). Hence, using the standard Gaussian concentration result (see Lemma [11] along with the assumption that \( n \geq C d \log d \), we have:

\[
\lambda_{\text{max}} \left( \frac{1}{n} \sum_{i=1}^{n} \Lambda^{-\frac{1}{2}} X_i^T A X_i \Lambda^{-\frac{1}{2}} \right) \leq 2 \text{tr}(A^T A).
\]  \hspace{1cm} (29)

We now consider RHS of (27). Note that,

\[
\frac{1}{n} \sum_{i=1}^{n} \Lambda^{-\frac{1}{2}} X_i^T A^T B_i \sim N(0, \beta I_{d \times d}),
\]  \hspace{1cm} (30)

where \( \beta^2 = \frac{1}{n^2} \sum_{i=1}^{n} ||A^T B_i||_2^2 = \frac{1}{n^2} \text{tr}(\Sigma_i^2 B_i A^T B_i) \sum_{i=1}^{n} \eta_i \eta_i^T \), where \( \eta_i \sim N(0, I_{m \times m}) \). Here again, using Lemma [11] we have (w.p. \( \geq 1 - 1/n^{10} \)):

\[
\beta^2 \geq \frac{1}{2n} \text{tr}(B^T A A^T B \Sigma_a).
\]

Hence, w.p. \( \geq 1 - 1/n^{10} - \exp(-d) \), we have:

\[
\left\| \frac{1}{n} \sum_{i=1}^{n} \Lambda^{-\frac{1}{2}} X_i^T A T_i \right\|_2 \geq \frac{\sqrt{d}}{2n} \sqrt{\text{tr}(B^T A A^T B \Sigma_a)}.
\]  \hspace{1cm} (31)

We obtain the following by combining (27), (29), and (31):

\[
E_{X_i \sim P_X} ||X(\bar{w} - w_*)||^2_2 \geq \frac{d}{16n} \frac{m \text{tr}(B^T A A^T B \Sigma_a)}{\text{tr}(A^T A)^2}.
\]  \hspace{1cm} (32)

Note that the “\( m \)” term on RHS appears as \( E_{X_i \sim P_X} ||X(\bar{w} - w_*)||^2_2 = m ||\Lambda^{\frac{1}{2}} (\bar{w} - w_*)||^2_2 \).

Lemma now follows by using \( A = B = I_{m \times m} \) for OLS and \( A = B = \Sigma_a^{-\frac{1}{2}} \) for MLE.

**Lemma 16.** Let \( y_i, X_i, \eta_i, w_i \) be as defined in Theorem [13] and let \( w_i \in \mathbb{R}^d \) be any fixed vector independent of \( (X_i, \eta_i) \). Also, let \( \bar{X}_i = \frac{1}{n} \sum (y_i - X_i w_i) (y_i - X_i w_i)^T, \Sigma_i = \Delta + \Sigma_a, \) where,

\[
\Delta = \frac{1}{n} \sum_i X_i (w_* - w_i) (w_* - w_i)^T X_i^T,
\]

and \( X_i \)‘s are independent of \( w_i \). Then, w.p. \( \geq 1 - \exp(-C \cdot n) \), the following holds \( \forall v \in \mathbb{R}^d \):

\[
\frac{1}{2} \Sigma_i v \leq v^T \bar{X}_i v \leq 2 \Sigma_i v.
\]

**Proof.** Let \( v \in \mathbb{R}^m \) be any vector. Also, \( \Sigma_i = \Delta + \Sigma_a \succ 0 \) as \( \Sigma_a \succ 0 \). Hence,

\[
v^T \bar{X}_i v = v^T \Sigma_i^{-\frac{1}{2}} \left( \sum_i z_i z_i^T \right) \Sigma_i^{-\frac{1}{2}} v,
\]

where \( z_i = \Sigma_i^{-\frac{1}{2}} X_i (w_* - w_i) \) and \( \Sigma_i^{-\frac{1}{2}} \eta_i \) is an “uncentered” Gaussian vector and hence, \( ||z_i||_{\psi_2} \leq C \) for a global constant \( C > 0 \). Also, \( E[z_i z_i^T] = I \). Lemma now follows using standard Gaussian concentration similar to Lemma [10].
Claim 17. Assume the notation of Theorem 2. Then the following holds:

\[
\frac{1}{\lambda_{\text{min}}} \leq \|\Sigma_w\|_2.
\]

Proof. Let \(\Sigma_w = \sum_j \lambda_j(\Sigma_w)u_ju_j^T\) be the eigenvalue decomposition of \(\Sigma_w\). Then, we have:

\[
\lambda_{\text{min}} = \min_{v, \|v\|=1} v^T \mathbb{E}_{X \sim P_X} \left[\Sigma_w^{-\frac{1}{2}} X^T \Sigma_w^{-1} X \Sigma_w^{-\frac{1}{2}}\right] v, \\
\geq \min_{v, \|v\|=1} \sum_j \frac{1}{\lambda_j(\Sigma_w)} v^T \mathbb{E}_{X \sim P_X} \left[\Sigma_w^{-\frac{1}{2}} X^T u_j u_j^T X \Sigma_w^{-\frac{1}{2}}\right] v \\
\geq \min_{\|v\|=1} \frac{1}{\|\Sigma_w\|_2} \min_{v, \|v\|=1} v^T \mathbb{E}_{X \sim P_X} \left[\Sigma_w^{-\frac{1}{2}} X^T X \Sigma_w^{-\frac{1}{2}}\right] v = \frac{1}{\|\Sigma_w\|_2}.
\]

Hence proved. \(\Box\)

Claim 18. Assume the notation of Section 2. Then, the following holds:

\[
\Sigma_w = \mathbb{E}_{X \sim P_X} \left[\Sigma_w^{-\frac{1}{2}} X^T \Sigma_w^{-1} X \Sigma_w^{-\frac{1}{2}}\right] = \frac{\text{tr}(\Sigma_R^{-1})}{\text{tr}(\Sigma_R)} \cdot I_{d \times d}, \text{ where } \Sigma_w = \text{tr}(\Sigma_R) \cdot \Lambda.
\]

Proof.

\[
\Sigma_w = \mathbb{E}_{X \sim P_X} \left[ X^T X \right] = \Lambda^T \cdot \mathbb{E}_{Z \sim N(0,1)} \left[ Z^T \Sigma_R Z \right] \cdot \Lambda \\
\Sigma_w = \mathbb{E}_{X \sim P_X} \left[ \Sigma_w^{-\frac{1}{2}} X^T \Sigma_w^{-1} X \Sigma_w^{-\frac{1}{2}}\right] = \frac{\text{tr}(\Sigma_R^{-1})}{\text{tr}(\Sigma_R)} \cdot I_{d \times d}.
\]

\(\Box\)

Corollary 19 (Result for Pooled Model, Gaussian Data, Dependent Rows). Let \(X_i\) be as defined above. Let \(n \geq C(m + d) \log(m + d)\). Then the following holds (w.p. \(\geq 1 - T/n^2\)):

\[
\frac{C'd}{n} \cdot \frac{m \cdot \text{tr}(\Sigma_R \Sigma_w)}{\text{tr}(\Sigma_R^2)} \leq \mathbb{E}_{X \sim P_X} \left[ \|X(w_{OLS} - w_*)\|_2^2 \right] \leq \frac{C'd}{n} \cdot \frac{m \cdot \text{tr}(\Sigma_R \Sigma_w)}{\text{tr}(\Sigma_R^2)}, \\
\frac{C'd}{n} \cdot \frac{m \cdot \text{tr}(\Sigma_R^{-1})}{\text{tr}(\Sigma_R^{-1})} \leq \mathbb{E}_{X \sim P_X} \left[ \|X(w_{MLE} - w_*)\|_2^2 \right] \leq \frac{C'd}{n} \cdot \frac{m \cdot \text{tr}(\Sigma_R^{-1})}{\text{tr}(\Sigma_R^{-1})}, \\
\mathbb{E}_{X \sim P_X} \left[ \|X(w_T - w_*)\|_2^2 \right] \leq \frac{8Cd n}{n} \cdot \frac{m}{\text{tr}(\Sigma_R \Sigma_w)} + \epsilon,
\]

where, \(w_T\) is the output of Algorithm 1 with \(T = \log \frac{1}{\epsilon}\).

D Proof of Claims from Section 3

Proof of Theorem 2 Let \(\widetilde{X}_i \in \mathbb{R}^{m \times m-d}\) be defined as:

\[
\widetilde{X}_i = \begin{bmatrix}
X_i \frac{1}{n} & 0 & \cdots & 0 \\
0 & X_i^2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & X_i^m
\end{bmatrix}.
\]

(33)

Also, let \(w_* = \text{vec}(W_*) \in \mathbb{R}^{md \times 1}\) and similarly, \(w_t = \text{vec}(W_t), \forall t\).

Then, the observations \(y_i\) can be re-written as:

\[
y_i = \widetilde{X}_i w_* + \eta_i.
\]

Similarly, we can rewrite the Step 4 in Algorithm 2 as:

\[
\text{vec}(W_{t+1}) = w_{t+1} = \arg \min_{w \in \mathbb{R}^{md}} \frac{1}{n} \sum_{i=1}^n \|\Sigma_i^{-\frac{1}{2}} (\widetilde{X}_i w - y_i)\|_2^2.
\]
That is, the above problem is a special case of the problem discussed in Section 2. First part of the theorem now follows using Lemma 3.

We now consider the second part of the theorem. Using the above given notation, we have:

\[
E_X \left[ \left\| \Sigma^{-1/2}_s (X \bullet W_T - X \bullet W_e) \right\|_2^2 \right] = E_X \left[ \sum_{j,k} (\Sigma^{-1}_s)_{jk} \left( W^j_T - W^k_T, W^j_e - W^k_e, X^j, X^k \right) \right],
\]

\[
\leq \zeta_1 E_X \left[ \sum_j (\Sigma^{-1}_s)_{jj} \left( W^j_T - W^j_e, X^j \right)^2 \right],
\]

where \( \zeta_1 \) follows from the fact that \( X^j \) and \( X^k \) are independent 0-mean vectors. Theorem now follows by using the above observation with the first part of the theorem.

**Claim 20.** Assume hypothesis and notation of Theorem 7. Then, we have: \((\Sigma^{-1}_s)_{jj} \geq \frac{1}{(\Sigma_s)_{jj}} \forall j\).

**Proof.** Let \( \Sigma_s = \sum_{k=1}^m \lambda_k (\Sigma_s) u_k u_k^T \) be the eigenvalue decomposition of \( \Sigma_s \). Now,

\[
1 = \sum_{k=1}^m (e_j^T u_k)^2 = \sum_{k=1}^m \frac{1}{\lambda_k(\Sigma_s)} \cdot \sqrt{\lambda_k(\Sigma_s)} (e_j^T u_k)^2 \leq (\Sigma^{-1}_s)_{jj} (\Sigma_s)_{jj},
\]

where the last inequality follows using Cauchy-Schwarz inequality. Hence, \((\Sigma^{-1}_s)_{jj} \geq \frac{1}{(\Sigma_s)_{jj}} \forall j\).

Moreover, equality holds only when \( \Sigma_s \) is a diagonal matrix.