Trace norm regularization for multi-task learning with scarce data

Etienne Boursier ETIENNE.BOURSIER@EPFL.CH
TML Lab, EPFL, Switzerland

Mikhail Konobeev MIKHAIL.KONOBEEV@EPFL.CH
TML Lab, EPFL, Switzerland

Nicolas Flammarion NICOLAS.FLAMMARION@EPFL.CH
TML Lab, EPFL, Switzerland

Editors: Po-Ling Loh and Maxim Raginsky

Abstract

Multi-task learning leverages structural similarities between multiple tasks to learn despite very few samples. Motivated by the recent success of neural networks applied to data-scarce tasks, we consider a linear low-dimensional shared representation model. Despite an extensive literature, existing theoretical results either guarantee weak estimation rates or require a large number of samples per task. This work provides the first estimation error bound for the trace norm regularized estimator when the number of samples per task is small. The advantages of trace norm regularization for learning data-scarce tasks extend to meta-learning and are confirmed empirically on synthetic datasets.

Keywords: Multi-task learning; Meta-learning; Trace norm regularization; Low rank matrix estimation

1. Introduction

Common supervised learning requires a large number of training examples, which are often costly in time and resources to acquire. The available dataset for a single task can be very limited, making impossible to learn solely based on it. Multi-task learning instead estimates a model across multiple tasks, by leveraging structural similarities among them. It jointly uses all datasets and thus learns efficiently as already observed in numerous applications including natural language processing (Ando et al., 2005), image segmentation (Cheng et al., 2011) and medical prediction (Caruana, 1997).

This work considers the problem of multi-task learning, where an unknown linear low-dimensional representation is shared among different tasks (Rohde and Tsybakov, 2011). It studies the following question: *how can we learn across multiple tasks with a very limited number of observations for each of them?* This question is also of fundamental interest to meta-learning and few-shot learning, which aim at aggregating knowledge among multiple tasks to learn a shared representation (Vinyals et al., 2016; Finn et al., 2017).

In spite of the vast multi-task learning literature, the existing results remain unsatisfying. In particular, guarantees on trace norm regularization (Rohde and Tsybakov, 2011) and Burer-Monteiro factorization (Tripuraneni et al., 2021) both assume that the number m of observations per task is large. The former assumes it is larger than the features dimension d, while the latter assumes it scales logarithmically in T, the total number of tasks. Such conditions are not always met in practice, when it is much easier to acquire high dimensional data on new tasks than on existing ones (see e.g. Wang et al., 2017). On the other hand, the Method of Moments (Tripuraneni et al., 2021) learns with a very limited number of observations per task, but requires very specific feature distributions.

Similarly to Rohde and Tsybakov (2011), we study the trace norm regularized estimator. It is a natural choice when estimating low rank matrices, since the trace norm convexifies the rank function. Trace norm based methods have already been successfully used in numerous multi-task learning applications (Amit et al., 2007; Cheng et al., 2011; Harchaoui et al., 2012), but lack theoretical guarantees when the number of observations per task is limited.

Contributions. This work bounds the estimation error of the trace norm regularized estimator with a few observations per task $(m < d)$. The analysis becomes particularly intricate when the number of samples per task is smaller than the features dimension, since no restricted isometry condition holds (Rohde and Tsybakov, 2011). Instead, our analysis uses a weaker restricted strong convexity condition (van de Geer and Bühlmann, 2009). Proving that this restricted strong convexity condition holds is our main technical contribution, besides upper bounding a stochastic term using concentration of heavy tailed distributions. These techniques lead to our main result, of which an informal version is given in Theorem 1.

Theorem 1 (informal). *For any number of observations per task m, the trace norm regularized estimator $\hat{M}$ satisfies with high probability*

$$\|\hat{M} - M^*\|_F \leq \tilde{\mathcal{O}}\left(\sigma\sqrt{r\frac{\frac{d^2}{m} + T}{m}} + \sqrt{rd\frac{d+T}{m^2}}\right),$$

where T is the number of tasks, d is the dimension of the feature space, σ^2 is the variance of the label noise, $M^* \in \mathbb{R}^{d \times T}$ is the ground-truth parameter matrix and r is its rank. The notation $\tilde{\mathcal{O}}$ hides multiplicative constants and logarithmic terms in d, m and T.

Note that by using linear regressions on each individual task independently, the estimation error scales as $\mathcal{O}\left(\sigma\sqrt{\frac{dT}{m}}\right)$ (Hsu et al., 2012). In the regime when the number of tasks is large, trace norm regularization thus improves this estimation by a factor $\sqrt{\frac{r}{m}}$, leveraging the low rank structure of the parameter matrix. A gap yet remains with the oracle baseline knowing beforehand the r-dimensional subspace induced by the parameters. This baseline computes linear regressions with r parameters and thus has an error scaling as $\mathcal{O}\left(\sigma\sqrt{\frac{rT}{m}}\right)$.

To our knowledge, Theorem 1 is the first general estimation error bound for a multi-task estimator with an arbitrarily small number of observations per task. As discussed in Section 3, a better bound can be proven for the Method of Moments in this setting[1], but it only holds for a very specific data model (e.g., Gaussian) and behaves much worse in practice as highlighted in Section 6.

Theorem 1 also allows in Section 5 to bound the estimation error for a new, previously unobserved task. This result illustrates the interest of trace norm regularization for meta-learning. Finally, we compare empirically different multi-task regression methods and discuss the practical advantages/drawbacks of trace norm regularization in Section 6.

[1] Such a bound is proven in Appendix A and is a minor contribution of this work.

2. Model

Notations. In the following, $[n] \coloneqq \{1, \ldots, n\}$. For a matrix $M \in \mathbb{R}^{d \times T}$, $M^{(t)} \in \mathbb{R}^d$ denotes its t-th column, $\lambda_i(M)$ its i-th largest singular value and $\|M\|_*$ its trace (or nuclear) norm, i.e., $\|M\|_* = \sum_{i=1}^{\min(d,T)} \lambda_i(M)$. We use the notation $\langle \cdot, \cdot \rangle$ for the canonical inner product both for vectors and matrices.

Model. In the remaining of the paper, we consider the model described in this section. There are T tasks, each of which contains m observation samples $(x_i^t, y_i^t) \in \mathbb{R}^d \times \mathbb{R}$. We consider the linear model

$$y_i^t = \langle {M^*}^{(t)}, x_i^t \rangle + \varepsilon_i^t \qquad \text{for any } (i,t) \in [m] \times [T], \tag{1}$$

where M^* is the matrix of parameters to estimate. We assume in the following that $\operatorname{rank}(M^*) = r$, where $r \ll d$, and that the features and noise variables are well behaved as stated in Assumption 1.

Assumption 1 (Random design). *The (x_i^t) are independent centered 1-sub-Gaussian random variables and the ε_i^t are independent centered σ-sub-Gaussian random variables. Moreover, the features are isotropic, i.e., $\mathbb{E}[(x_i^t)^\top x_i^t] = I_d$.*

We also assume the task diversity condition, which claims that the scale of the parameters is roughly the same for all tasks.

Assumption 2 (Task diversity). *Given some constant C, the parameters matrix M^* verifies:*

$$\max_{t \in [T]} \|{M^*}^{(t)}\|^2 \leq C.$$

The task diversity assumption has been introduced by Tripuraneni et al. (2021) and is also considered in subsequent works (Thekumparampil et al., 2021a,b). It ensures that a single task does not get too significant with respect to the others. However, we do not require any lower bound on the norm of task parameters.

3. Related work

This section discusses the related literature and Table 1 summarizes the available error bounds for the model described in Section 2.

Different structural assumptions have been considered in the multi-task literature. For example, Denevi et al. (2019); Cesa-Bianchi et al. (2021) assume that the task parameters all lie in a small Euclidean ball and Argyriou et al. (2008); Lounici et al. (2009) assume that each parameter vector is sparse and its support is shared among the tasks. In the latter, the parameter matrix M^* has a small $\ell_{2,1}$ norm. This paper studies a classical structural assumption generalizing the sparse setting: the parameter matrix has a small rank. In that case, it seems natural to consider the following estimator

$$\operatorname*{argmin}_{\substack{M \in \mathbb{R}^{d \times T} \\ \operatorname{rank}(M) \leq r}} \frac{1}{mT} \sum_{(i,t) \in [m] \times [T]} \left(y_i^t - \langle M^{(t)}, x_i^t \rangle \right)^2 . \tag{2}$$

When the features are shared among the tasks, i.e., $x_i^t = x_i^{t'}$, the considered model is equivalent to multivariate regression (Izenman, 1975; Obozinski et al., 2008) and a closed form solution of Equation (2) is known (Bunea et al., 2011). Maurer et al. (2016) bound the error of this estimator in a general multi-task setting, using Gaussian complexity arguments. Besides holding only for bounded Lipschitz loss functions, this bound is weaker than what can be obtained for the squared loss, since it does not use any smoothness property on the loss function.

Computing the above optimization program yet becomes intractable when the features differ among the tasks, which corresponds to the setting of interest. A natural approach replaces the rank constraint by a trace norm constraint, since it convexifies the rank function (similarly to the ℓ_1 norm that convexifies the ℓ_0 norm). Equivalently, a regularized problem can be considered:

$$\operatorname*{argmin}_{M \in \mathbb{R}^{d \times T}} \frac{1}{mT} \sum_{(i,t) \in [m] \times [T]} \left(y_i^t - \langle M^{(t)}, x_i^t \rangle \right)^2 + \lambda \|M\|_*.$$

Multi-task learning can be considered as a particular case of matrix completion. Candes and Plan (2011); Rohde and Tsybakov (2011) studied low-rank matrix completion, using restricted isometry conditions. In particular, Rohde and Tsybakov (2011) bound the error of the trace norm regularized estimator described above for multi-task learning. However, they assume a restricted isometry condition, which only holds when the number of samples per task is larger than the features dimension $(m \geq d)$, limiting the interest of this result in practice.

The only known error bounds for trace norm based methods when the number of observations per task is small $(m < d)$ derive from Rademacher complexity arguments (Pontil and Maurer, 2013; Yousefi et al., 2018). For the same reasons as Maurer et al. (2016), they only hold for Lipschitz loss functions and are weaker than what can be obtained for the squared loss.

Other approaches yet manage to provide near tight error bounds when the number of observations per task is smaller than the features dimension. In particular, the Burer-Monteiro factorization considers the problem

$$\operatorname*{argmin}_{\substack{U \in \mathbb{R}^{d \times r} \\ V \in \mathbb{R}^{T \times r}}} \frac{1}{mT} \sum_{(i,t) \in [m] \times [T]} \left(y_i^t - \langle (UV^\top)^{(t)}, x_i^t \rangle \right)^2.$$

In words, low-rank matrices are factorized as $M = UV^\top$. This optimization problem is equivalent to Equation (2). It is not convex in its arguments (U, V), but only bilinear. As a consequence, we can only aim at computing a local minimum of the objective, for example with first order optimization methods. Tripuraneni et al. (2021) nevertheless bounded the estimation error of any local minimum of the above optimization program. Their work is the closest in spirit to ours and is motivated by the theoretical study of meta-learning.

Thekumparampil et al. (2021a,b) recently improved the error guarantees of the Burer-Monteiro factorization, in terms of the distance between the estimated r-dimensional features subspace and the ground-truth one. It is achieved using an alternate minimization algorithm. This allows to provide tighter estimation bounds on a new task in the meta-learning setting considered in Section 5, but does not improve the existing multi-task bounds. However, all the bounds for Burer-Monteiro factorization require that the number of samples per task scales with $\log(T)$. Since we might consider a very large number of tasks in practice, along with a limited number of samples per task, this requirement is a major drawback.

The Method of Moments introduced by Tripuraneni et al. (2021) is actually the only estimator that provides satisfying bounds with a very limited number of observations per task. It directly estimates the r-dimensional features subspace. Yet, only a bound on the error of the subspace estimation is known, besides requiring the feature distribution to be Gaussian. In Appendix A, we extend these results to an error bound on the whole estimated parameters matrix and to any spherically

Estimator	**Error bound** $\|\hat{M} - M^*\|_F$	**Samples per task**	**Extra assumption**
Trace norm regularization (Rohde and Tsybakov, 2011)	$\sigma\sqrt{r\frac{d+T}{m}}$	$\Omega(d)$	Deterministic features
Burer-Monteiro factorization (Tripuraneni et al., 2021)	$\sigma\sqrt{r\frac{d+T}{m}}$	$\Omega(r^4 \log(T))$	-
Method of Moments Theorem 3 adapted from (Tripuraneni et al., 2021)	$\sigma\sqrt{r\frac{\sigma^2 rd+T}{m}} + r\sqrt{\frac{d}{m}}$	$\Omega(r\log(r))$	Spherically symmetric feature distribution
Trace norm regularization **Theorem 2**	$\sigma\sqrt{r\frac{\frac{d^2}{m}+T}{m}} + \sqrt{rd\frac{d+T}{m^2}}$	$\Omega(1)$	**-**

Table 1: Different bounds for multi-task learning. Only the dependencies in σ, r, d, m, T are provided and eventual logarithmic terms are omitted. Our main result is highlighted in bold.

symmetric feature distribution. This assumption on the feature distribution is yet often unverified in practice, and the Method of Moments might fail to learn the features subspace without it as shown in Section 6. Moreover, it empirically performs poorly with respect to the other estimators even for Gaussian distributions, as observed in Section 6.

Multi-task classification has also been studied in previous works (Maurer, 2006; Cavallanti et al., 2010), but is not further discussed as it is beyond the scope of this paper.

4. Bound on the estimation error

In this section, we provide error guarantees for the estimator

$$\hat{M} = \operatorname*{argmin}_{M \in \mathcal{W}} \frac{1}{mT} \sum_{(i,t)\in[m]\times[T]} (y_i^t - \langle x_i^t, M^{(t)} \rangle)^2 + \lambda \|M\|_* \tag{3}$$

where $\mathcal{W} = \left\{ M \in \mathbb{R}^{d\times T} \mid \max_{t\in[T]} \|M^{(t)}\|^2 \leq C \right\}$ and C is the constant introduced in Assumption 2. We restrict the estimator to the ball $\mathcal{W}$ for analysis purpose, but we do not need to enforce this constraint in practice, i.e. we empirically obtain good results when solving the unconstrained problem.

Our proof relies on the decomposability of the trace norm (Negahban et al., 2012). Since the restricted isometry condition does not hold with a limited number of observations per task, we instead prove a restricted strong convexity condition. Its proof is particularly difficult, since the condition is non-uniform and considered for an intricate subset of matrices. On the other hand, bounding the effective noise level is challenging because of the random design model. As a consequence, our analysis uses concentration on heavy tailed distributions, while previous works on trace norm regularization use Bernstein inequalities that only hold for sub-exponential distributions.

Showing both restricted strong convexity and noise level conditions is the main technical challenge of this work. These conditions are respectively presented in Sections 4.1 and 4.2. We now state our main result, which bounds the error of the estimator defined in Equation (3). Its proof is given in Section 4.3.

Theorem 2. *Assume $T = \Omega(d)$, for $\lambda = 4\tau$ where $\tau = \frac{c_2\sigma}{\sqrt{T}}\sqrt{\frac{T+d^2/m}{mT}}$, with probability at least $1-(2T+c_0)e^{-c_1 d} - 2e^{-c_1 r(d+T)}$:*

$$\|\hat{M} - M^*\|_F \leq c\sigma\sqrt{r\frac{\frac{d^2}{m}+T}{m}} + c\sqrt{Crd\frac{d+T}{m^2}\ln\left(\frac{dT}{m}\right)}, \tag{4}$$

where c, c_0, c_1 and c_2 are universal positive constants.

This bound is of the same order as the known error bound when the number of samples per task is larger than the dimension (Rohde and Tsybakov, 2011). This result is of great significance when $m < \min(d, \log(T))$ since it provides the first estimation guarantees in this case. We recall it is the regime of interest in most applications. It illustrates the success of trace-norm methods in multi-task learning settings. It indeed leads to a $\sqrt{\frac{r}{m}}$ estimation improvement with respect to the single-task baseline, which proceeds to T independent linear regressions.

We believe that the extra $\sqrt{\frac{d}{m}}$ factor in the second term is only an artefact of the analysis as explained in Section 4.1. It is confirmed empirically. In this case, leveraging the low rank structure of the parameter matrix through the trace norm regularization would lead to a $\sqrt{\frac{T}{d}}$ improvement over single-task learning and the trace norm regularized estimator would be comparable to the baseline oracle that knows beforehand the r-dimensional subspace induced by the parameters.

4.1. Restricted strong convexity

To define the restricted strong convexity condition, we first need to define matrices $U \in \mathbb{R}^{d\times d}$ and $V \in \mathbb{R}^{T\times T}$, such that the SVD of M^* reads $M^* = U\Sigma^* V^\top$, where $\Sigma^* \in \mathbb{R}^{d\times T}$ has only its first r diagonal elements that are non-zero. We now define the following cone of matrices, which is key to the analysis

$$\mathcal{C} = \left\{\Delta \in \mathbb{R}^{d\times T} \mid \|\Delta_{22}\|_* \leq 3\left\|\begin{pmatrix}\Delta_{11} & \Delta_{12} \\ \Delta_{21} & 0\end{pmatrix}\right\|_* \text{ where } \Delta = U \begin{array}{c|c|c} & \overbrace{\quad}^{r} & \overbrace{\quad}^{T-r} \\ r\{ & \Delta_{11} & \Delta_{12} \\ \hline d-r\{ & \Delta_{21} & \Delta_{22} \end{array} V^\top\right\}. \tag{5}$$

Note that matrices of the form $U\begin{pmatrix}\cdot & \cdot \\ \cdot & 0\end{pmatrix}V^\top$ are of rank at most $2r$ with the block dimensions given in Equation (5). Any matrix in $\mathcal{C}$ is thus close to low-rank, since a submatrix of rank $2r$ counts for a significant amount of its nuclear norm. We now define the linear operator $\mathcal{L} : \mathbb{R}^{d\times T} \to \mathbb{R}^{m\times T}$

$$\mathcal{L} : M \mapsto \frac{1}{\sqrt{mT}}(\langle x_i^t, M^{(t)}\rangle)_{\substack{1\leq i\leq m,\\ 1\leq t\leq T}}$$

which is lower bounded in norm over $\mathcal{C}$ with high probability by Lemma 1 below.

Lemma 1 (Restricted strong convexity). *With probability larger than $1 - 2e^{-cr(d+T)} - 2Te^{-d}$, the operator $\mathcal{L}$ satisfies*

$$\|\mathcal{L}(\Delta)\|_F^2 \geq \frac{c_0}{T}\|\Delta\|_F^2 - \frac{c_1 rd(d+T)}{m^2 T}\max_{t\in[T]}\|\Delta^{(t)}\|_2^2 \ln\left(\frac{dT}{m}\right) \qquad \textit{for all } \Delta \in \mathcal{C}, \tag{6}$$

where c, c_0 and c_1 are positive universal constants.

This condition generalizes the restricted eigenvalue condition, which is used to prove estimation guarantees of the Lasso or Dantzig estimator in the problem of sparse linear regression (van de Geer and Bühlmann, 2009). It is weaker than the restricted isometry property, which does not hold in the multi-task setting (Rohde and Tsybakov, 2011).

The proof of Lemma 1 is deferred to Appendix C.1. Lemma 11 by Tripuraneni et al. (2021) states a similar condition on the subset of matrices of rank at most $2r$. Besides correcting minor errors in its proof, we extend this condition to the set $\mathcal{C}$, which is much larger: its ε-covering number scales exponentially with $\frac{1}{\varepsilon^2}$. As a consequence, our bound includes an additional $\frac{d}{m}$ factor in the last term. Although we believe this $\frac{d}{m}$ factor to be an artefact of the analysis, covering arguments might not yield better bounds, since the considered subset for the restricted strong convexity condition is much larger here. We let the investigation of more advanced techniques as future work.

4.2. Effective noise level

The next lemma bounds the effect of label noise on the prediction.

Lemma 2 (Effective noise level). *If $T = \Omega(d)$, then for universal positive constants c_0, c_1, c_2, with probability at least $1 - (2T + c_0)e^{-c_1 d}$:*

$$\left| \frac{1}{mT} \sum_{(i,t)\in[m]\times[T]} \varepsilon_i^t \langle x_i^t, M^{(t)} \rangle \right| \leq \tau \|M\|_* \qquad \textit{uniformly for all } M \in \mathbb{R}^{d\times T}, \tag{7}$$

where $\tau = \frac{c_2 \sigma}{\sqrt{T}} \sqrt{\frac{T + d^2/m}{mT}}$.

The proof is given in Appendix C.2 and uses concentration results on heavy tailed distributions (Bakhshizadeh et al., 2020) to bound the operator norm of the matrix $\frac{1}{mT} \sum_{(i,t)\in[m]\times[T]} \varepsilon_i^t e_t (x_i^t)^\top$.

For "nice" fixed features x_i^t, the best possible bound on τ is of order $\frac{\sigma}{\sqrt{T}} \sqrt{\frac{T+d}{mT}}$ using classical bounds on the spectral norm of Gaussian matrices. Our bound is thus tight, up to the $\frac{d^2}{m}$ term, which is actually due to the randomness of the features x_i^t. Because of the random design, we cannot directly use Bernstein inequality but instead use concentration on heavy tailed distributions. In any case, when the number of tasks is large enough, the T term prevails over $\frac{d^2}{m}$ and the bound becomes similar to the easier setting of fixed features.

4.3. Proof of Theorem 2

The proof assumes that Equations (6) and (7) both hold, which happens with high probability thanks to Lemmas 1 and 2. By definition, the estimator $\hat{M}$ minimizes the objective function in $\mathcal{W}$, to which M^* belongs, thanks to Assumption 2. In particular:

$$\frac{1}{mT} \sum_{(i,t)\in[m]\times[T]} (y_i^t - \langle x_i^t, \hat{M}^{(t)} \rangle)^2 + \lambda \|\hat{M}\|_* \leq \frac{1}{mT} \sum_{(i,t)\in[m]\times[T]} (y_i^t - \langle x_i^t, M^{*(t)} \rangle)^2 + \lambda \|M^*\|_*. \tag{8}$$

Using simple manipulations, this inequality is equivalent for $\hat{\Delta} = \hat{M} - M$ to

$$\|\mathcal{L}(\hat{\Delta})\|^2 \leq \frac{2}{mT} \sum_{(i,t)\in[m]\times[T]} \varepsilon_i^t \langle x_i^t, \hat{\Delta}^{(t)} \rangle + \lambda \left(\|M^*\|_* - \|\hat{M}\|_* \right)$$

Thanks to Equation (7), this becomes $\|\mathcal{L}(\hat{\Delta})\|^2 \leq 2\tau\|\hat{\Delta}\|_* + \lambda\left(\|M^*\|_* - \|\hat{M}\|_*\right)$.

With the SVD decomposition $M^* = U\Sigma^* V^\top$, we can decompose $\hat{\Delta} = \hat{\Delta}_1 + \hat{\Delta}_2$ where

$$\hat{\Delta} = U\begin{pmatrix}\hat{\Delta}_{11} & \hat{\Delta}_{12}\\ \hat{\Delta}_{21} & \hat{\Delta}_{22}\end{pmatrix}V^\top; \quad \hat{\Delta}_1 = U\begin{pmatrix}\hat{\Delta}_{11} & \hat{\Delta}_{12}\\ \hat{\Delta}_{21} & 0\end{pmatrix}V^\top \quad \text{and} \quad \hat{\Delta}_2 = U\begin{pmatrix}0 & 0\\ 0 & \hat{\Delta}_{22}\end{pmatrix}V^\top.$$

Using the triangle inequality, $\|\hat{M}\|_* \geq \|M^* + \hat{\Delta}_2\|_* - \|\hat{\Delta}_1\|_*$. Moreover, as M^* and $\hat{\Delta}_2$ have orthogonal column and rowspaces, $\|M^* + \hat{\Delta}_2\|_* = \|M^*\|_* + \|\hat{\Delta}_2\|_*$. Equation (8) then leads with the choice $\lambda = 4\tau$ to

$$\|\mathcal{L}(\hat{\Delta})\|^2 \leq 6\tau\|\hat{\Delta}_1\|_* - 2\tau\|\hat{\Delta}_2\|_*.$$

From there, we can first observe that $3\|\hat{\Delta}_1\|_* \geq \|\hat{\Delta}_2\|_*$, i.e. $\hat{\Delta} \in \mathcal{C}$. Moreover, since $\hat{\Delta}_1$ is of rank at most $2r$, $\|\hat{\Delta}_1\|_* \leq \sqrt{2r}\|\hat{\Delta}_1\|_F \leq \sqrt{2r}\|\hat{\Delta}\|_F$, i.e.

$$\|\mathcal{L}(\hat{\Delta})\|^2 \leq 6\tau\sqrt{2r}\|\hat{\Delta}\|_F. \tag{9}$$

Now since $\hat{\Delta} \in \mathcal{C}$, Equation (6) directly gives

$$\|\mathcal{L}(\hat{\Delta})\|^2 \geq \frac{c_0}{T}\|\hat{\Delta}\|_F^2 - \frac{c_1 r d(d+T)}{m^2 T}\max_{t\in[T]}\|\hat{\Delta}^{(t)}\|_2^2 \ln\left(\frac{dT}{m}\right). \tag{10}$$

Moreover, as $\hat{M}$ and M^* are both in $\mathcal{W}$, $\max_t \|\hat{\Delta}^{(t)}\|_2^2 \leq 4C$. Combining Equations (9) and (10), this yields

$$\frac{c_0}{T}\|\hat{\Delta}\|_F^2 - 6\tau\sqrt{2r}\|\hat{\Delta}\|_F - \frac{4Cc_1 r d(d+T)}{m^2 T}\ln\left(\frac{dT}{m}\right) \leq 0.$$

Note that the left expression is a 2-degree polynomial in $\|\hat{\Delta}\|_F$. To be non-positive, it requires Equation (4) to hold, which concludes the proof.

5. Meta-learning: transfer on a new task

Meta-learning is of significant interest in modern applications, where the knowledge acquired on previous tasks is transferred to a single new task. The objective of the meta-learning setting is to estimate the parameters of a new single task, based on the regression obtained on the previous T tasks. This section provides estimation error guarantees for this setting.

In the following, we use the decomposition $M^* = B\alpha$ where $B \in \mathbb{R}^{d\times r}$ and $\alpha \in \mathbb{R}^{r\times T}$ such that $B^\top B = I_r$[2]. Now consider the matrix $\tilde{M}$, defined as the rank r matrix which is the closest to $\hat{M}$:

$$\tilde{M} \in \operatorname*{argmin}_{\substack{L\in\mathbb{R}^{d\times T}\\ \text{rank(L)}\leq r}} \|L - \hat{M}\|_F. \tag{11}$$

$\tilde{M}$ can be computed from the SVD of $\hat{M}$ by keeping its r largest singular values. Now decompose $\tilde{M}$ as $\tilde{M} = \tilde{B}\tilde{\alpha}$ where $\tilde{B} \in \mathbb{R}^{d\times r}$, $\tilde{\alpha} \in \mathbb{R}^{r\times T}$ and $\tilde{B}^\top\tilde{B} = I_r$.

The meta-learning setting considers a $T+1$-st task, with m observations (x_i^{T+1}, y_i^{T+1}) generated as described in Section 2. Using the estimated subspace matrix $\tilde{B}$, we compute the least squares estimate

$$\tilde{\alpha}_{T+1} \in \operatorname*{argmin}_{\theta\in\mathbb{R}^r}\sum_{i=1}^m\left(y_i^{T+1} - \langle\tilde{B}\theta, x_i^{T+1}\rangle\right)^2.$$

[2] B corresponds to the r first columns of U defined in Section 4.1, up to any rotation of $\mathbb{R}^r$.

The idea behind meta-learning is that even for a small number of observations m on this single task, the parameters vector $M^{*\,(T+1)}$ can still be well estimated using the T previous tasks.

Before bounding the error $\|\tilde{B}\tilde{\alpha}_{T+1} - M^{*\,(T+1)}\|_2$, Lemma 3 bounds the angles between the r-dimensional subspaces corresponding to B and $\tilde{B}$. It is adapted from Tripuraneni et al. (2021, Lemma 16) to take into account the additional error that might appear from the low-rank projection step given in Equation (11). Its proof is given in Appendix C.4.

Lemma 3. *For B and $\tilde{B}$ defined as above:*

$$\sin^2\theta\left(B,\tilde{B}\right) \le \frac{4r\|\hat{M}-M^*\|_F^2}{T\nu},$$

where $\nu = r\lambda_r\left(\frac{M^*M^{*\top}}{T}\right)$ and $\theta\left(B,\tilde{B}\right)$ is the principal angle between the subspaces corresponding to the orthogonal projections B and $\tilde{B}$.

The variable ν in Lemma 3 is introduced for clarity. Note that if the tasks are *rich*[3], then $\nu = \Omega(1)$. Lemma 3 states that the r-dimensional feature subspace is well estimated if the parameter matrix is well estimated. From there, Theorem 4 by Tripuraneni et al. (2021) allows to bound the error of the estimation of the parameter vector for the $T+1$-th task. Corollary 1 below is stated in the balanced case, where the new task also has m samples. It can easily be extended to the unbalanced case.

Corollary 1. *If Equation (4) holds and $m = \Omega(r\log(r))$, then for $\tilde{B}$ and $\tilde{\alpha}_{T+1}$ defined above, with probabilty at least $1-\frac{c}{m^{100}}$:*

$$\|\tilde{B}\tilde{\alpha}_{T+1} - M^{*\,(T+1)}\|_2^2 \le c'C\sigma^2\frac{r^2}{m\nu}\left(\frac{d^2}{mT}+\log(m)\right) + c'C^2\frac{r^2d}{m^2\nu}\left(\frac{d}{T}+1\right),$$

where c and c' are universal constants.

The squared estimation error on a new task is roughly bounded by $\frac{r^2d}{m^2}$. In contrast, it is known that a (single task) linear regression on this task would lead to an estimation error of order $\frac{d}{m}$ (Hsu et al., 2012), i.e., leveraging the low rank structure of the parameters and the past observations leads at least to an improvement $\frac{r^2}{m}$ for the trace norm regularized estimator.

As a comparison, the best known bounds for Burer-Monteiro factorization (Thekumparampil et al., 2021a) and Method of Moments (Tripuraneni et al., 2021) on a new task are of order $\sigma^2\frac{r}{m}$ for a large number of tasks, thus being comparable to the oracle baseline. These approaches yield better meta-learning bounds, but require respectively $m = \Omega(\log(T))$ or a spherically symmetric feature distribution.

The discrepancy between these bounds is due to different analyses. Our meta-learning bound directly derives from the multi-task bound using Lemma 3, while the tight analyses of Burer-Monteiro (with alternate minimization) and Method of Moments directly bound the principal angle between the estimated subspaces. Lemma 3 indeed considers the unrealistic worst case, where the entirety of the estimation error of the matrix M^* is due to the error in subspace estimation. As a consequence, Lemma 3 leads to a loose bound on the subspace angle of $\mathcal{O}\left(\frac{r^2d}{m^2}\right)$, while this angle clearly goes to 0 when the number of tasks grows to infinity in the simulations of Section 6. Showing it actually converges to 0 for a large number of tasks is left for future work and should be done by directly

[3] For example if $\|M^*\|_F^2 = \Omega(T)$ and $\frac{\lambda_1(M^*M^{*\,T})}{\lambda_r(M^*M^{*\,T})} = \mathcal{O}(1)$

bounding the subspace angle. Such a result would then lead to an optimal meta-learning bound for trace norm based methods, without any further requirement on the number of observations per task or on the feature distribution.

6. Experiments

This section compares empirically different multi-task methods on synthetic datasets and discusses the practical aspects of their implementation.

6.1. Simulations

With the exception of Figure 4, our experimental setup follows that of Tripuraneni et al. (2021); Thekumparampil et al. (2021b). More specifically, we set $d = 100, r = 5$ and sample $x_i^t \overset{\text{iid}}{\sim} \mathcal{N}(0, I_d)$, $\varepsilon_i^t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2)$. The following experiments study how the normalized Frobenius distance $\|\hat{M} - M^*\|_F / \sqrt{T}$ and the angle between the subspaces $\sin\theta\left(B, \tilde{B}\right)$ behave when varying the number of tasks T, the size of each task m and the label noise level σ. In all experiments the markers show the average and the shaded area shows the standard deviation over 12 independent runs. The code is available in github.com/MichaelKonobeev/meta and additional experimental details are given in Appendix B.

In this section, altmin corresponds to the alternating minimization algorithm (Thekumparampil et al., 2021a); bm and mom respectively correspond to Burer-Monteiro factorization and the Method of Moments (Tripuraneni et al., 2021); nuc corresponds to the nuclear norm regularized estimator. Additionally, we implement two other baselines labeled single, which implements independent least squares regression of the d-dimensional parameter vectors (the columns of M) for each task; and oracle which knows the ground-truth subspace B and only estimates the matrix α by performing independent least squares regression for each task in the projected r-dimensional space. Some algorithms perform very poorly for certain values of T, m, σ in Frobenius distance. In such cases, we omit displaying these points in the figures for the sake of readability.

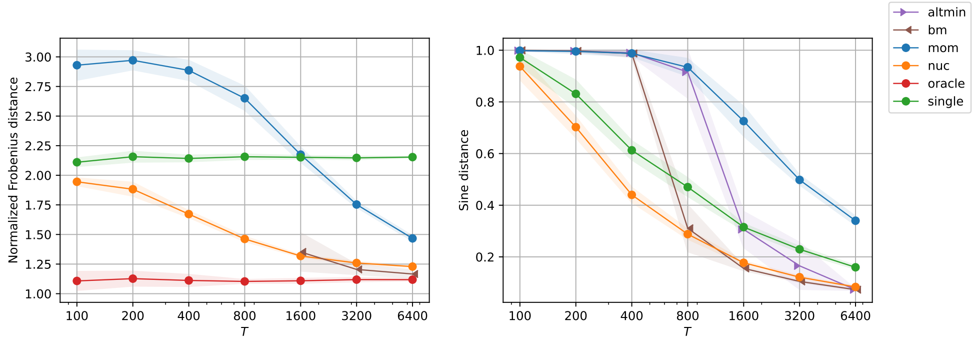

Figure 1: Evolution of the estimation error with the number of tasks T for $m = 10$.

Figure 1 displays the normalized Frobenius distance and the angle between the subspaces when varying the number of tasks T for a small number of observations per task ($m = 10$). We observe that nuc outperforms the other algorithms in both distances. When the number of tasks becomes large, altmin and bm nevertheless become comparable in sine distance (even in Frobenius distance for bm).

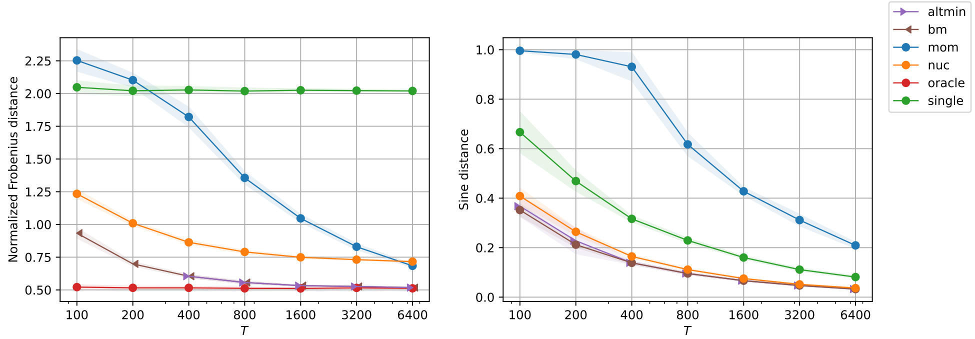

Figure 2: Evolution of the estimation error with the number of tasks T for $m = 25$.

On the other hand, `altmin` and `bm` outperform `nuc` with larger m ($m = 25$) as shown in Figure 2. The regime of interest for `nuc` thus seems to be for small values of m, as can be seen in Figure 3 below. These empirical results confirm the different known theoretical bounds.

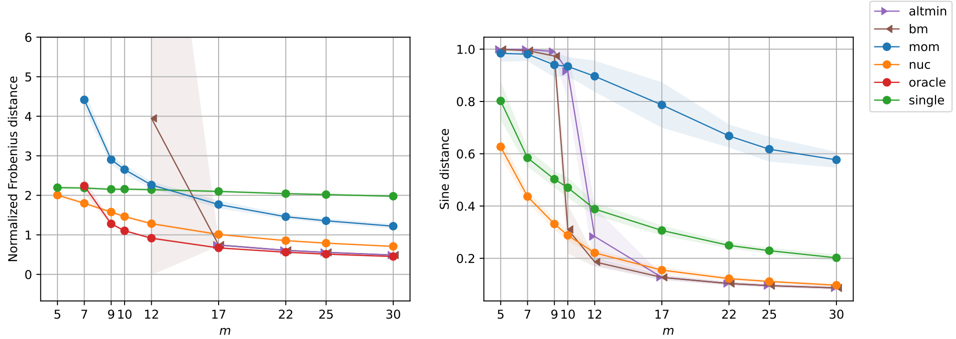

Figure 3: Evolution of the estimation error with the task size m for $T = 800$.

Method of Moments is largely outperformed by the other algorithms, unless the number of tasks is very large. Furthermore, as highlighted in Figure 4 below, it might fail for non-spherically symmetric distributions[4]. In particular, the largest principal angle between its estimated subspace and the real one remains equal to $\frac{\pi}{2}$. Method of Moments yet outperforms `single` in Frobenius distance, because it still manages to learn some of the directions of the r-dimensional subspace.

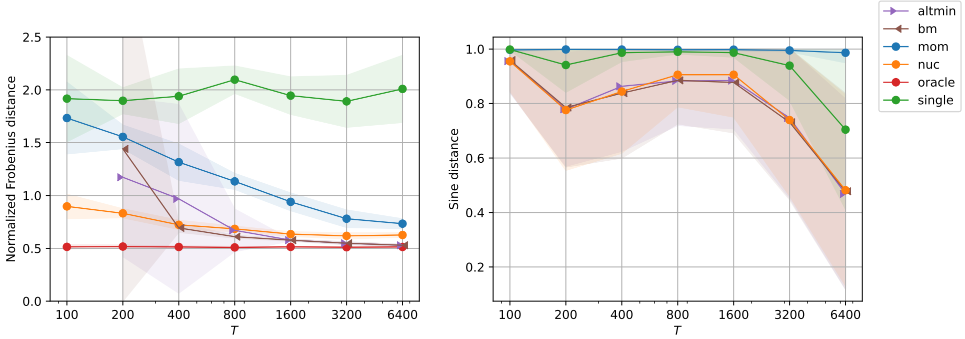

Figure 4: Evolution of the estimation error with the number of tasks T for $m = 25$ and a non-spherically symmetric feature distribution.

[4] Further details on the chosen feature and parameters distribution are given in Appendix B.

Finally, Figure 5 studies the evolution of the estimation error with the level noise σ. Algorithms `altmin` and `bm` do not perform well in Frobenius norm here and are not displayed for a clearer figure. Although `nuc` scales better with σ than `mom`, both methods seem to not recover the exact parameters matrix in the noiseless setting, confirming the $(1+\sigma)$ dependence in Theorem 1 for the former.

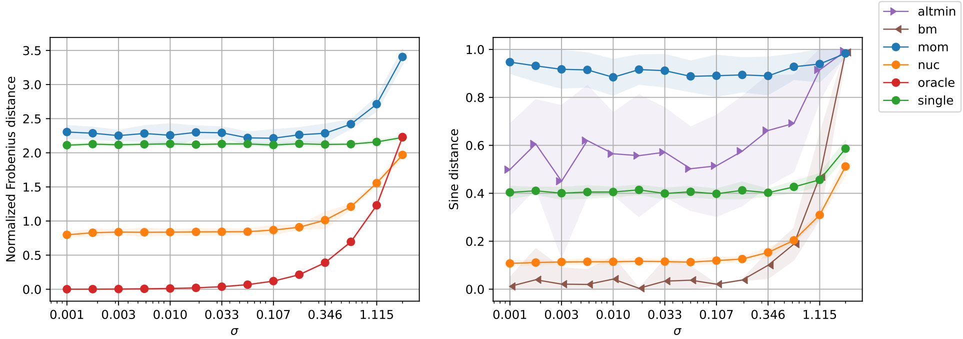

Figure 5: Evolution of the estimation error with the noise level σ for $T = 800$ and $m = 10$.

As explained in Section 5, `nuc` is comparable to `altmin` in subspace estimation in the experiments. This suggests that the meta-learning bound of Corollary 1 is not tight and it needs further investigation.

6.2. Numerical complexity

A local minimum of the Burer-Monteiro factorization can be approximated up to ε using first order methods such as gradient descent (Jin et al., 2017) in $\mathcal{O}\left(\frac{1}{\varepsilon^2}\right)$ steps. Computing the gradient of the objective function here requires at each step a multiplication between $d \times r$ and $r \times T$ matrices, leading to a total complexity of $\mathcal{O}\left(\frac{rdT}{\varepsilon^2}\right)$.

For trace norm regularization, algorithms approximating a solution of Equation (3) up to ε in $\mathcal{O}\left(\frac{1}{\sqrt{\varepsilon}}\right)$ steps exist (Ji and Ye, 2009; Toh and Yun, 2010), but require an SVD computation (of complexity dT^2) at each step, thus leading to a considerable numerical complexity for a large number of tasks. Instead, Jaggi and Sulovskỳ (2010) propose an algorithm approximating the solution of the equivalent constrained problem with a numerical complexity $\mathcal{O}\left(\frac{dT}{\varepsilon^2}\right)$. In practice, it might thus be preferable to consider the constrained problem, as it is cheaper in computation and the considered optimization programs are equivalent for properly tuned regularization/constraint parameters.

On the other hand, Method of Moments is much better in terms of complexity as it only computes a single truncated SVD and then proceeds to T linear regressions with r parameters. In total, its complexity is thus of order $\mathcal{O}\left(rT(d+m^2)\right)$, largely outperforming the other methods in terms of numerical complexity. The Method of Moments is thus computationally cheaper than both Burer-Monteiro factorization and trace norm constrained minimization, which end up being similar in computational cost.

Parameter tuning. Another important practical aspect of these algorithms is parameter tuning. Although trace norm methods require tuning the regularization parameter, Theorem 2 needs this parameter to depend on known variables except from the noise level σ. On the other hand, Burer-Monteiro factorization and Method of Moments both require the knowledge of the rank of the parameters matrix, which is unknown. The parameter tuning is thus easier in practice for trace norm based methods, being one of their main practical advantages in multi-task learning.

7. Conclusion

This work proposes the first multi-task estimation error bound when the number of samples per task is very limited. It illustrates the interest of nuclear norm regularization for low rank matrix estimation in this intricate regime and confirms empirically its good performance on synthetic datasets, with respect to the other known methods. It confirms that learning shared representation is possible with a limited number of samples per task, in the sample linear representation model, grasping insights of the empirical success on learning non-linear representation.

In light of the experiments, the proposed bounds might be improvable, especially in the subspace estimation. Such an improvement is left open for future work and would require more refined analytical tools.

Acknowledgements

We warmly thank Nilesh Tripuraneni and Alexandre Tsybakov for their time and precious feedback.

References

Yonatan Amit, Michael Fink, Nathan Srebro, and Shimon Ullman. Uncovering shared structures in multiclass classification. In *Proceedings of the 24th international conference on Machine learning*, pages 17–24, 2007.

Rie Kubota Ando, Tong Zhang, and Peter Bartlett. A framework for learning predictive structures from multiple tasks and unlabeled data. *Journal of Machine Learning Research*, 6(61):1817–1853, 2005.

Andreas Argyriou, Theodoros Evgeniou, and Massimiliano Pontil. Convex multi-task feature learning. *Machine learning*, 73(3):243–272, 2008.

Milad Bakhshizadeh, Arian Maleki, and Victor H de la Pena. Sharp concentration results for heavy-tailed distributions. *arXiv preprint arXiv:2003.13819*, 2020.

Florentina Bunea, Yiyuan She, and Marten H Wegkamp. Optimal selection of reduced rank estimators of high-dimensional matrices. *The Annals of Statistics*, 39(2):1282–1309, 2011.

Emmanuel J Candes and Yaniv Plan. Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. *IEEE Transactions on Information Theory*, 57 (4):2342–2359, 2011.

Rich Caruana. Multitask learning. *Machine learning*, 28(1):41–75, 1997.

Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *Journal of Machine Learning Research*, 11:2901–2934, 2010.

Nicolò Cesa-Bianchi, Pierre Laforgue, Andrea Paudice, and Massimiliano Pontil. Multitask online mirror descent. *arXiv preprint arXiv:2106.02393*, 2021.

Bin Cheng, Guangcan Liu, Jingdong Wang, Zhongyang Huang, and Shuicheng Yan. Multi-task low-rank affinity pursuit for image segmentation. In *2011 International Conference on Computer Vision*, pages 2439–2446. IEEE, 2011.

Giulia Denevi, Carlo Ciliberto, Riccardo Grazzi, and Massimiliano Pontil. Learning-to-learn stochastic gradient descent with biased regularization. In *International Conference on Machine Learning*, pages 1566–1575. PMLR, 2019.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pages 1126–1135. PMLR, 2017.

Zaid Harchaoui, Matthijs Douze, Mattis Paulin, Miroslav Dudik, and Jérôme Malick. Large-scale image classification with trace-norm regularization. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 3386–3393. IEEE, 2012.

Daniel Hsu, Sham M Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on learning theory*, pages 9–1. JMLR Workshop and Conference Proceedings, 2012.

Alan Julian Izenman. Reduced-rank regression for the multivariate linear model. *Journal of multivariate analysis*, 5(2):248–264, 1975.

Martin Jaggi and Marek Sulovskỳ. A simple algorithm for nuclear norm regularized problems. In *Proceedings of the 27th International Conference on International Conference on Machine Learning*, pages 471–478, 2010.

Shuiwang Ji and Jieping Ye. An accelerated gradient method for trace norm minimization. In *Proceedings of the 26th International Conference on Machine Learning*, pages 457–464, 2009.

Chi Jin, Rong Ge, Praneeth Netrapalli, Sham M Kakade, and Michael I Jordan. How to escape saddle points efficiently. In *International Conference on Machine Learning*, pages 1724–1732. PMLR, 2017.

Karim Lounici, Massimiliano Pontil, Alexandre Tsybakov, and Sara van de Geer. Taking advantage of sparsity in multi-task learning. In *Conference on Learning Theory*, 2009.

Andreas Maurer. Bounds for linear multi-task learning. *Journal of Machine Learning Research*, 7: 117–139, 2006.

Andreas Maurer, Massimiliano Pontil, and Bernardino Romera-Paredes. The benefit of multitask representation learning. *Journal of Machine Learning Research*, 17(81):1–32, 2016.

Sahand N Negahban, Pradeep Ravikumar, Martin J Wainwright, and Bin Yu. A unified framework for high-dimensional analysis of m-estimators with decomposable regularizers. *Statistical science*, 27(4):538–557, 2012.

Guillaume Obozinski, Martin J Wainwright, and Michael I Jordan. Union support recovery in high-dimensional multivariate regression. In *2008 46th Annual Allerton Conference on Communication, Control, and Computing*, pages 21–26. IEEE, 2008.

Massimiliano Pontil and Andreas Maurer. Excess risk bounds for multitask learning with trace norm regularization. In *Conference on Learning Theory*, pages 55–76. PMLR, 2013.

Angelika Rohde and Alexandre Tsybakov. Estimation of high-dimensional low-rank matrices. *The Annals of Statistics*, 39(2):887–930, 2011.

Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Sample efficient linear meta-learning by alternating minimization. *arXiv preprint arXiv:2105.08306*, 2021a.

Kiran Koshy Thekumparampil, Prateek Jain, Praneeth Netrapalli, and Sewoong Oh. Statistically and computationally efficient linear meta-representation learning. *Advances in Neural Information Processing Systems*, 34, 2021b.

Kim-Chuan Toh and Sangwoon Yun. An accelerated proximal gradient algorithm for nuclear norm regularized linear least squares problems. *Pacific Journal of optimization*, 6(615-640):15, 2010.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pages 10434–10443. PMLR, 2021.

Sara van de Geer and Peter Bühlmann. On the conditions used to prove oracle results for the lasso. *Electronic Journal of Statistics*, 3:1360–1392, 2009.

Roman Vershynin. *High-Dimensional Probability: An Introduction With Applications In Data Science*, volume 47. Cambridge university press, 2018.

Oriol Vinyals, Charles Blundell, Timothy Lillicrap, Daan Wierstra, et al. Matching networks for one shot learning. *Advances in neural information processing systems*, 29:3630–3638, 2016.

Yu-Xiong Wang, Deva Ramanan, and Martial Hebert. Learning to model the tail. *Advances in Neural Information Processing Systems*, 30, 2017.

Niloofar Yousefi, Yunwen Lei, Marius Kloft, Mansooreh Mollaghasemi, and Georgios C Anagnostopoulos. Local rademacher complexity-based learning guarantees for multi-task learning. *Journal of Machine Learning Research*, 19(1):1385–1431, 2018.

Appendix A. Multi-task analysis of Method of Moments

This section provides a multi-task analysis of the Method of Moments (MoM) algorithm introduced in (Tripuraneni et al., 2021). Tripuraneni et al. (2021) only provided a meta-learning type bound (similar to Corollary 1) for this algorithm, in the particular case of a Gaussian feature distribution. This section adapts this result to derive a multi-task bound for Algorithm 1, with any spherically symmetric feature distribution[5].

Algorithm 1: Method of Moments

```
Input: $(x_i^t, y_i^t)_{i,t} \in \left(\mathbb{R}^d \times \mathbb{R}\right)^{mT}$
$\hat{B}_1 \hat{D}_1 \hat{B}_1^\top \leftarrow$ top $r$-SVD of $\sum_{i=1}^m \sum_{t=\lceil \frac{T}{2} \rceil+1}^{T} y_i^t x_i^t (x_i^t)^\top$          // $\hat{B}_1 \in \mathbb{R}^{d \times r}$
$\hat{\alpha}_1 \leftarrow \mathrm{argmin}_{\alpha \in \mathbb{R}^{r \times \lceil \frac{T}{2} \rceil}} \sum_{i=1}^m \sum_{t=1}^{\lceil \frac{T}{2} \rceil} \left(y_i^t - \langle \hat{B}_1 \alpha^{(t)}, x_i^t \rangle\right)^2$
$\hat{B}_2 \hat{D}_2 \hat{B}_2^\top \leftarrow$ top $r$-SVD of $\sum_{i=1}^m \sum_{t=1}^{\lceil \frac{T}{2} \rceil} y_i^t x_i^t (x_i^t)^\top$
$\hat{\alpha}_2 \leftarrow \mathrm{argmin}_{\alpha \in \mathbb{R}^{r \times T - \lceil \frac{T}{2} \rceil}} \sum_{i=1}^m \sum_{t=\lceil \frac{T}{2} \rceil+1}^{T} \left(y_i^t - \langle \hat{B}_2 \alpha^{(t)}, x_i^t \rangle\right)^2$
return $(\hat{B}_1 \hat{\alpha}_1, \hat{B}_2 \hat{\alpha}_2)$
```

This algorithm directly estimates the low dimensional subspace by aggregating all the observations (for different tasks) together. This estimation relies on the fourth moments of the features, hence its name. Note that the considered MoM algorithm is slightly modified with respect to the original one: we split the tasks in two batches and estimate a batch parameters using the estimated subspace on the other batch. This trick is used to get independence between the estimated subspace and the estimated parameters, which allows to apply Theorem 4 by Tripuraneni et al. (2021). This result could not be directly used without splitting the data.

Theorem 3 bounds the estimation error of MoM. Its proof is deferred to Appendix C.5.

Theorem 3. *Consider the setting defined in Section 2. Denote $M_1^* \in \mathbb{R}^{d \times \lceil \frac{T}{2} \rceil}$ the matrix corresponding to the first $\lceil \frac{T}{2} \rceil$ columns of M^* and M_2^* the matrix corresponding to the remaining columns. Additionally, if the feature distribution is spherically symmetric and $m \geq cr\log(r)$ and $mT \geq c\frac{\log^6(dmT)\left(\sigma^4+C^2\right)r^2}{\kappa^2\nu^2}d$, then with probability at least $1 - \frac{2}{m^2T^2}$, the estimator returned by Algorithm 1 verifies:*

$$\|\hat{M} - M^*\|_F^2 \leq c'\sigma^2 \frac{r}{m} T \log(mT) + c' C \log^6(dmT) r^2 \frac{\sigma^4 + C^2}{\kappa^2 \bar{\nu}^2} \frac{d}{m}, \tag{12}$$

where $\bar{\nu} = r\min\left(\lambda_r\left(\frac{M_1^* M_1^{*\top}}{T}\right), \lambda_r\left(\frac{M_2^* M_2^{*\top}}{T}\right)\right)$, $\kappa = \mathbb{E}[\langle e_1, x\rangle^4] - \mathbb{E}[\langle e_1, x\rangle^2 \langle e_2, x\rangle^2] > 0$, c and c' are universal positive constants.

Similarly to the term ν in Section 5, $\bar{\nu} = \Omega(1)$ for rich tasks. For large T, the first term then prevails and Theorem 3 recovers (up to a logarithmic term) the estimation error of a linear regression on an r-dimensional subspace. The method of moments algorithm however requires the feature distribution to be spherically symmetric and might fail otherwise as illustrated by Figure 4 in Section 6.

[5] A distribution is spherically symmetric if it is invariant under any orthogonal transformation.

Moreover, the method of moments is significantly worse than Burer-Monteiro factorization and trace norm regularization in practice, hence the need for a better understanding of these methods.

Appendix B. Experimental details

In Figures 1 to 4, we choose $\sigma = 1$. Except for Figure 4, the ground-truth matrix M^* is generated as follows. We first take B as the top r left singular vectors of a $d \times d$ random matrix whose entries are i.i.d. and drawn from $\mathcal{N}(0,1)$. Next, we sample $\alpha \in \mathbb{R}^{r\times T}$ with $\alpha_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and compute the resulting matrix $M^* = B\alpha$.

In all experiments except the one in Figure 5 the value of the regularization coefficient for nuc is chosen as $\lambda = \frac{\sigma}{\sqrt{T}}\sqrt{\frac{T+d^2/m}{mT}}$. In the experiment with varying σ in Figure 5, we found that for small values of σ the regularization coefficient computed using the above formula becomes too small. We thus use validation sets of size $0.2m$ for each task to find the value of the regularization coefficient via grid search.

For the setting in Figure 4 of Section 6, the feature distribution samples independently each coordinate. The k-th coordinate is given by $\langle x_i^t, e_k\rangle = \cos(\frac{k}{d}\frac{\pi}{2})\xi_{i,k}^t + \sin(\frac{k}{d}\frac{\pi}{2})\eta_{i,k}^t$, where $\xi_{i,k}^t \overset{\text{iid}}{\sim} \mathcal{U}[-\sqrt{3},\sqrt{3}]$ and $\eta_{i,k}^t \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$. The r-dimensional parameters are generated by first sampling $r \times r$ matrix α' with each element $(\alpha')_{i,j} \overset{\text{iid}}{\sim} \mathcal{N}(0,1)$ and next completing this to a $T \times r$ matrix by sampling $T-r$ vectors uniformly at random among the set of columns of the matrix α'.

This choice of feature distribution is to ensure different fourth moments of the features along different directions, while this choice of parameters avoids that the average parameters matrix is too well behaved.

Appendix C. Proofs

This section provides all the proofs deferred from the main text.

C.1. Proof of Lemma 1

In the whole section, we assume $m = \mathcal{O}(d)$. Adapting Lemma 4 below, we can actually show if $m \gtrsim d$ that with probability at least $1 - 2T\exp(-d)$:

$$\frac{c}{\sqrt{T}} \leq \min_{\Delta\in\mathbb{R}^{d\times T}} \frac{\|\mathcal{L}(\Delta)\|_F}{\|\Delta\|_F} \leq \max_{\Delta\in\mathbb{R}^{d\times T}} \frac{\|\mathcal{L}(\Delta)\|_F}{\|\Delta\|_F} \leq \frac{c'}{\sqrt{T}},$$

thus leading to a restricted isometry condition that also implies Lemma 1.

The proof first bounds the operator norm of $\mathcal{L}$.

Lemma 4. *With probability at least* $1 - 2T\exp(-d)$*:*

$$\max_{\Delta\in\mathbb{R}^{d\times T}} \frac{\|\mathcal{L}(\Delta)\|_F}{\|\Delta\|_F} \leq c\frac{\sqrt{d}+\sqrt{m}}{\sqrt{mT}},$$

for some universal constant c.

Proof. Note that by definition of $\mathcal{L}$, for any $\Delta \in \mathbb{R}^{d\times T}$:

$$\|\mathcal{L}(\Delta)\|_2^F = \frac{1}{mT}\sum_{t=1}^{T}\|X_t\Delta^{(t)}\|_2^2, \tag{13}$$

where $\Delta^{(t)}$ is the t-th column of Δ and $X_t = [x_1^t, \dots, x_m^t]^\top \in \mathbb{R}^{m\times d}$.

By design of the setting, X_t is a matrix with independent, mean zero, sub-gaussian isotropic random vectors in $\mathbb{R}^d$. Theorem 4.6.1 by Vershynin (2018) directly yields for a universal constant c that

$$\|X_t\|_2 \leq (1+c)\sqrt{d} + c\sqrt{m} \qquad \text{with probability at least } 1 - 2e^{-d}.$$

Taking a union bound over all tasks, with probability at least $1-2Te^{-d}$, $\|X_t\|_2 \leq (1+c)\sqrt{d}+c\sqrt{m}$ for all tasks $t \in [T]$. This finally leads to Lemma 4 using Equation (13). ■

We then show an RSC condition on the set of low rank matrices similar to Lemma 11 by Tripuraneni et al. (2021).

Lemma 5. *For any $r' \in \mathbb{N}$, with probability at least $1 - \exp\left(c_1 r'(d+T)\right)$:*

$$\|\mathcal{L}(\Delta)\|_F^2 \geq 3\frac{\|\Delta\|_F^2}{8T} - c'\frac{r'(d+T)}{mT}\max_{t\in[T]}\|\Delta^{(t)}\|^2 \ln\left(\frac{dT}{m}\right) \qquad \textit{uniformly over all matrices of rank at most } r',$$

where c, c', c_1, c_2 are positive universal constants.

Proof. By homogeneity, it suffices to show this for any matrix in $\Gamma_{r'}$ where

$$\Gamma_{r'} = \{M \in \mathbb{R}^{d\times T} \mid \operatorname{rank}(M) \leq r' \text{ and } \|M\|_F = 1\}.$$

For any $\varepsilon \in (0,1)$, we know there exists an ε-covering of $\Gamma_{r'}$ of cardinality smaller than $\left(\frac{9}{\varepsilon}\right)^{r'(d+T+1)}$ (Candes and Plan, 2011, Lemma 3.1). For $\varepsilon = \min\left\{\frac{c}{4}\frac{\sqrt{m}}{\sqrt{d}+\sqrt{m}}, \frac{1}{\sqrt{T}}\right\}$ where c is the constant in Lemma 4, let $\mathcal{N}$ be an ε-covering of $\Gamma_{r'}$ of minimal size. By union bound, taking $\log\frac{1}{\delta}$ of order $r'(d+T)\ln\left(\frac{dT}{m}\right)$ in Lemma 6 of Appendix C.3 then states that with probability at least $1-\exp\left(-c_1 r'(T+d)\right)$, for all $M \in \mathcal{N}$:

$$\|\mathcal{L}(M)\|^2 \geq \frac{1}{T} - \frac{c_1}{4}\frac{\max_t\|M^{(t)}\|_2}{\sqrt{m}T}\sqrt{r'(d+T)\ln\left(\frac{dT}{m}\right)} - c_1^2\frac{\max_t\|M^{(t)}\|^2}{mT}r'(d+T)\ln\left(\frac{dT}{m}\right). \tag{14}$$

Also, Lemma 4 states that with probability at least $1-2T\exp(-d)$,

$$\|\mathcal{L}\|_2^2 \leq c\frac{d+m}{mT}. \tag{15}$$

Assume in the following that Equations (14) and (15) both hold. Consider $\Delta \in \Gamma_{r'}$ and decompose as $\Delta = M + A$ where $M \in \mathcal{N}$ and $\|A\|_F \leq \varepsilon$.

Note that Equation (14) leads to the following inequality, using $a^2 - \frac{ab}{4} - b^2 \geq \frac{7}{8}a^2 - \frac{9}{8}b^2$:

$$\|\mathcal{L}(M)\|^2 \geq \frac{7}{8T} - \frac{9c_1^2}{8}\frac{\max_t\|M^{(t)}\|^2}{mT}r'(d+T)\ln\left(\frac{dT}{m}\right).$$

Moreover, $\max_t\|M^{(t)}\|^2 \leq 2\max_t\|\Delta^{(t)}\|^2 + 2\max_t\|A^{(t)}\|^2$. As $\varepsilon \leq \frac{1}{\sqrt{T}} \leq \max_t\|\Delta^{(t)}\|$, this last inequality yields that $\max_t\|M^{(t)}\|^2 \leq 4\max_t\|\Delta^{(t)}\|^2$, and we thus have:

$$\|\mathcal{L}(M)\|^2 \geq \frac{7}{8T} - \frac{9c_1^2}{2}\frac{\max_t\|\Delta^{(t)}\|^2}{mT}r'(d+T)\ln\left(\frac{dT}{m}\right). \tag{16}$$

Thanks to Equation (15) and the choice of ε, we also have

$$\|\mathcal{L}(A)\|^2 \leq \frac{1}{16T}. \tag{17}$$

Finally, this leads to

$$\begin{aligned} \|\mathcal{L}(\Delta)\|^2 &\geq (\|\mathcal{L}(M)\| - \|\mathcal{L}(A)\|)^2 && \text{by triangle inequality} \\ &\geq \frac{\|\mathcal{L}(M)\|^2}{2} - \|\mathcal{L}(A)\|^2 && \text{using } (a-b)^2 \geq \frac{a^2}{2} - b^2 \\ &\geq \frac{3}{8T} - \frac{9c_1^2}{8}\frac{\max_t \|\Delta^{(t)}\|^2}{mT} r'(d+T)\ln\left(\frac{dT}{m}\right) && \text{from Equations (16) and (17).} \end{aligned}$$

■

We can now prove Lemma 1. By homogeinity, it suffices to show it for any $\Delta \in \mathcal{C}$ of Frobenius norm 1. Lemma 5 yields that with high probability:

$$\|\mathcal{L}(\Delta)\|_F^2 \geq \frac{3\|\Delta\|_F^2}{8T} - c\frac{r(d+T)}{mT\varepsilon^2}\max_{t\in[T]}\|\Delta^{(t)}\|^2 \ln\left(\frac{dT}{m}\right) \qquad \text{for all matrices of rank at most } \frac{32}{\varepsilon^2}r. \tag{18}$$

Recall that Lemma 4 states that with high probability,

$$\|\mathcal{L}\|_2^2 \leq c\frac{d+m}{mT}. \tag{19}$$

Assume in the following of the proof that both Equations (18) and (19) hold for $\varepsilon = \frac{1}{4\sqrt{c}}\frac{\sqrt{m}}{\sqrt{d}+\sqrt{m}}$.

Let $\Delta = \tilde{U}\tilde{\Sigma}\tilde{V}^\top$ be the SVD decomposition of Δ, i.e. $\tilde{U}^\top\tilde{U} = I_d, \tilde{V}^\top\tilde{V} = I_T$ and $\tilde{\Sigma} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{\min(d,T)})$, where the sequence (σ_i) is non-increasing.

Denote in the following $\tilde{\Sigma}_{r'} = \operatorname{diag}(\sigma_1, \ldots, \sigma_{r'}, 0, \ldots, 0)$ and $\Delta_{r'} = \tilde{U}\tilde{\Sigma}_{r'}\tilde{V}^\top$. Note that by definition of $\mathcal{C}$, $\|\Delta\|_* \leq 4\sqrt{2r}\|\Delta\|_F$, i.e.

$$\sum_{k=1}^{\min(d,T)} \sigma_k \leq 4\sqrt{2r}.$$

By monotonicity of the sequence, this implies the following decrease rate for the singular values σ_k:

$$\sigma_k \leq \frac{4\sqrt{2r}}{k} \qquad \text{for any } k \leq \min(d,T). \tag{20}$$

From this, it follows:

$$\begin{aligned} \|\Delta - \Delta_{r'}\|_F^2 &= \sum_{k=r'+1}^{\min(d,T)} \sigma_k^2 \\ &\leq 32r \sum_{k=r'+1}^{\min(d,T)} \frac{1}{k^2} && \text{using Equation (20)} \\ &\leq \frac{32r}{r'} && \text{by integral comparison.} \end{aligned}$$

Fix in the following $r' = \frac{32r}{\varepsilon^2}$ and decompose $\Delta = \Delta_{r'} + (\Delta - \Delta_{r'})$. Note that $\|\Delta - \Delta_{r'}\|_F \leq \varepsilon$ and $\Delta_{r'}$ is of rank at most $\frac{32}{\varepsilon^2}r$. Moreover, the columnspaces of $\Delta_{r'}$ and $\Delta - \Delta_{r'}$ are orthogonal, thanks to the used decomposition. Thanks to that, it follows that $\max_{t\in[T]}\|\Delta_{r'}^{(t)}\|_2 \leq \max_{t\in[T]}\|\Delta^{(t)}\|_2$. From Equations (18) and (19), we then have for $\varepsilon = \frac{1}{4\sqrt{c}}\frac{\sqrt{m}}{\sqrt{d}+\sqrt{m}}$.

$$\|\mathcal{L}(\Delta_{r'})\|_F^2 \geq \frac{3\|\Delta_{r'}\|_F^2}{8T} - 2\frac{rd(d+T)}{m^2T}\max_{t\in[T]}\|\Delta^{(t)}\|^2 \ln\left(\frac{dT}{m}\right)$$

$$\|\mathcal{L}(\Delta-\Delta_{r'})\|_F^2 \leq \frac{1}{16T}.$$

As $\|\Delta_{r'}\|_F^2 \geq 1-\frac{1}{16}$, we can show similarly to the proof of Lemma 5 that these two inequalities yield

$$\|\mathcal{L}(\Delta)\|^2 \geq \frac{1}{10T} - c_2\frac{rd(d+T)}{m^2T}\max_{t}\|\Delta^{t)}\|^2 \ln\left(\frac{dT}{m}\right)$$

for universal positive constants c_1, c_2. Moreover, this inequality holds uniformly over all $\Delta \in \mathcal{C}$ of norm 1, as soon as the Equations (18) and (19) simultaneously hold. Lemmas 4 and 5 allow to conclude.

C.2. Proof of Lemma 2

By trace duality property we have

$$\left|\frac{1}{mT}\sum_{(i,t)\in[m]\times[T]}\varepsilon_i^t\langle x_i^t, M^{(t)}\rangle\right| = \left|\frac{1}{mT}\sum_{(i,t)\in[m]\times[T]}\varepsilon_i^t\mathrm{tr}(e_t(x_i^t)^\top M)\right| \leq \|M\|_*\|F\|,$$

where $(e_t)_j = \mathbb{I}\{j=t\}$ for $1\leq j\leq T$ and $F := \frac{1}{mT}\sum_{t=1}^N\sum_{i=1}^m\varepsilon_i^t e_t(x_i^t)^\top$.

The remaining of the proof aims at bounding the spectral norm of $F\in\mathbb{R}^{T\times d}$. Equivalently, we consider $A = \frac{T\sqrt{m}}{\sigma}F$. Note that the t-th row of A is

$$A_t = \frac{1}{\sqrt{m}\sigma}\sum_{i=1}^m\varepsilon_i^t x_i^t.$$

In particular, the rows of A_t are i.i.d. and isotropic, i.e. $\mathbb{E}[A_i^\top A_i] = I_d$. We can then show Lemma 7 in Appendix C.3, which shows for a fixed $y\in\mathbb{S}^{d-1}$ and for any $\varepsilon\geq\varepsilon_0$:

$$\mathbb{P}\left(\|Ay\|_2^2 \geq (1+\varepsilon)T\right) \leq (2T+c)\exp\left(-c'\min\{T\varepsilon, \sqrt{mT\varepsilon}\}\right), \tag{21}$$

where c, c' and ε_0 are universal constants.

Let $\mathcal{N}$ be a $\frac{1}{4}$-net covering of $\mathbb{S}^{d-1}$ of cardinality 9^d. Using Equation (21) and a union bound argument leads, for any $\varepsilon\geq\varepsilon_0$ to

$$\mathbb{P}\left(\max_{y\in\mathcal{N}}\|Ay\|_2^2 \geq (1+\varepsilon)T\right) \leq (2T+c)\exp\left(d\ln(9) - c'\min\{T\varepsilon, \sqrt{mT\varepsilon}\}\right)$$

As we assumed that $T=\Omega(d)$, taking $\varepsilon = c_2\left(1+\frac{d^2}{mT}\right)$ for some positive constant c_2 gives

$$\mathbb{P}\left(\max_{y\in\mathcal{N}}\|Ay\|_2^2 \geq c_1(T+\frac{d^2}{m})\right) \leq (2T+c)e^{-c_0 d}.$$

Thus, with probability at least $1-(2T+c)e^{-c_0 d}$, $\max_{y\in\mathcal{N}}\|Ay\|_2 \leq \sqrt{c_1}\sqrt{T+\frac{d^2}{m}}$. Using a classical covering argument (e.g. Vershynin, 2018, Lemma 4.4.1), this implies that

$$\|A\|_2 \leq \frac{4}{3}\sqrt{c_1}\sqrt{T+\frac{d^2}{m}},$$

which leads to Lemma 2, since $\|A\|_2 = \frac{T\sqrt{m}}{\sigma}\|F\|_2$.

C.3. Auxiliary lemmas

Lemma 6. *For any matrix $M \in \mathbb{R}^{d\times T}$ and $\delta > 0$, with probability at least $1-\delta$:*

$$\left| \|\mathcal{L}(M)\|^2 - \frac{\|M\|_F^2}{T} \right| \leq \frac{c}{4} \frac{\max_t \|M^{(t)}\|_2 \|M\|_F}{\sqrt{m}T} \sqrt{\log\frac{1}{\delta}} + c^2 \frac{\max_t \|M^{(t)}\|^2}{mT} \log\frac{1}{\delta},$$

where c is a universal constant.

Proof. Recall that

$$\|\mathcal{L}(M)\|^2 = \frac{1}{mT} \sum_{(i,t)\in[m]\times[T]} \langle M^{(t)}, x_i^t\rangle^2.$$

As a consequence, $\mathbb{E}[\|\mathcal{L}(M)\|^2] = \frac{\|M\|_F^2}{T}$. Moreover, the random variables $\langle M^{(t)}, x_i^t\rangle^2$ are independent and $c'\|M^{(t)}\|^2$-sub-exponential for some constant c'. Bernstein inequality then yields for some universal constant c_1 (Vershynin, 2018, Theorem 2.8.1):

$$\begin{aligned}\mathbb{P}\left(\left| \|\mathcal{L}(M)\|^2 - \frac{\|M\|_F^2}{T} \right| \geq \frac{\varepsilon}{mT}\right) &\leq 2\exp\left(-c_1 \min\left\{\frac{\varepsilon^2}{c'^2 m \sum_{t=1}^T \|M^{(t)}\|^4}, \frac{\varepsilon}{c' \max_t \|M^{(t)}\|^2}\right\}\right)\\ &\leq 2\exp\left(-c_1 \min\left\{\frac{\varepsilon^2}{c'^2 m \|M\|_F^2 \max_t \|M^{(t)}\|^2}, \frac{\varepsilon}{c' \max_t \|M^{(t)}\|^2}\right\}\right).\end{aligned}$$

The second inequality comes from the inequality $\sum_{t=1}^T \|M^{(t)}\|^4 \leq \|M\|_F^2 \max_t \|M^{(t)}\|^2$.

Taking $\varepsilon = c\max_t \|M^{(t)}\|_2 \|M\|_F \sqrt{m\log\frac{1}{\delta}} + c^2 \max_t \|M^{(t)}\|^2 \log\frac{1}{\delta}$ for a large enough constant c leads to Lemma 6. ■

Lemma 7. *Let $A = \frac{1}{\sigma\sqrt{m}} \sum_{t=1}^N \sum_{i=1}^m \varepsilon_i^t e_t (x_i^t)^\top$. For a fixed $y \in \mathbb{S}^{d-1}$ and any $\varepsilon \geq \varepsilon_0$:*

$$\mathbb{P}\left(\|Ay\|_2^2 \geq (1+\varepsilon)T\right) \leq (2T+2)\exp\left(-c' \min\{T\varepsilon, \sqrt{mT\varepsilon}\}\right),$$

where ε_0 and c' are positive universal constants.

Proof. By definition of A and A_t:

$$\begin{aligned}\|Ay\|_2^2 &= \sum_{t=1}^T \langle A_t, y\rangle^2\\ &= \sum_{t=1}^T \frac{1}{m}\left(\sum_{i=1}^m \frac{\varepsilon_i^t}{\sigma}\langle x_i^t, y\rangle\right)^2.\end{aligned}$$

Define for the remaining of the proof $X_i^t = \langle x_i^t, y\rangle$, $Y_i^t = \frac{\varepsilon_i^t}{\sigma}$ and $Z_t = \frac{1}{m}\left(\sum_{i=1}^m X_i^t Y_i^t\right)^2$. By design of the problem, X_i^t and Y_i^t are independent, 0-mean, 1-subgaussian variables. The Bernstein inequality (e.g. Vershynin, 2018, Theorem 2.8.1.) then gives for a universal positive constant c':

$$\mathbb{P}\left(\left|\sum_{i=1}^m X_i^t Y_i^t\right| \geq \varepsilon\right) \leq 2\exp\left(-c\min\{\frac{\varepsilon^2}{m}, \varepsilon\}\right).$$

And so

$$\begin{aligned}\mathbb{P}\left(Z_t \geq \varepsilon\right) &= \mathbb{P}\left(\frac{1}{m}\left(\sum_{i=1}^{m} X_i^t Y_i^t\right)^2 \geq \varepsilon\right) \\ &\leq 2\exp\left(-c\min\{\varepsilon, \sqrt{m\varepsilon}\}\right).\end{aligned}$$

Moreover, note that the Z_t are independent, positive and $\mathbb{E}[Z_t] = 1$. We now want to use the following concentration result on heavy tailed distributions:

Lemma 8 (Bakhshizadeh et al. 2020, Theorem 1). *Let $Z^L = Z\mathbb{I}\{Z \leq L\}$ and c be any constant such that $\alpha_L \leq c$ for any $L \in \mathbb{R}$ where*

$$\alpha_L = \mathbb{E}\left[\left(Z^L - 1\right)\mathbb{I}\{Z \leq 1\} + \left(Z^L - 1\right)^2 \exp\left\{\frac{c' \min(\varepsilon, \sqrt{m\varepsilon}) - \ln(2)}{2L}(Z^L - 1)\right\}\mathbb{I}\{Z^L > 1\}\right]$$

and

$$\varepsilon_{\max} = \sup\left\{\varepsilon \geq 0 \mid \varepsilon \leq \frac{c}{2}\frac{c' \min(\varepsilon, \sqrt{m\varepsilon}) - \ln(2)}{T\varepsilon}\right\},$$

then for any $\varepsilon \geq \varepsilon_{\max}$:

$$\mathbb{P}\left(\sum_{t=1}^{T} Z_t - T > T\varepsilon\right) \leq 2\exp\left(-\frac{c' \min(T\varepsilon, \sqrt{Tm\varepsilon})}{4}\right) + 2T\exp\left(-c' \min(T\varepsilon, \sqrt{Tm\varepsilon})\right).$$

Using Lemma 2 from (Bakhshizadeh et al., 2020), we can bound α_L as follows:

$$\alpha_L \leq 1 + 2\int_0^\infty \exp\left(-\frac{c'}{2}\min(u+1, \sqrt{m(u+1)})\right)(2u + \frac{c't}{2}\min(u, \sqrt{mu}))\mathrm{d}u.$$

A simple calculation allows to bound the integral by some constant value that does not depend on m, i.e. we can use Lemma 8 where c is a universal constant. $\varepsilon_{\max}$ is then smaller than some universal constant ε_0 and thus, for $\varepsilon \geq \varepsilon_0$:

$$\mathbb{P}\left(\sum_{t=1}^{T} Z_t > (1+\varepsilon)T\right) \leq (2T+2)\exp\left(-\frac{c'}{4}\min\{T\varepsilon, \sqrt{mT\varepsilon}\}\right).$$

This leads to Lemma 7 as $\sum_{t=1}^{T} Z_t = \|Ay\|_2^2$. ■

C.4. Proof of Lemma 3

As M^* is of rank r, the definition of $\tilde{M}$ in Equation (11) implies

$$\|\hat{M} - \tilde{M}\|_F \leq \|\hat{M} - M^*\|_F.$$

By triangle inequality, $\|\tilde{M} - M^*\|_F \leq 2\|\hat{M} - M^*\|_F$. A direct application of Lemma 16 by Tripuraneni et al. (2021), which we recall below, then allows to conclude.

Lemma 9 (Tripuraneni et al. 2021, Lemma 16). *Suppose we have matrices $\tilde{B}, B \in \mathbb{R}^{d\times r}$ and $\tilde{\alpha}, \alpha \in \mathbb{R}^{r\times T}$ such that $\tilde{B}^\top\tilde{B} = I_r = B^\top B$ and $\|\tilde{B}\tilde{\alpha} - B\alpha\|_F^2 \leq \varepsilon$, then*

$$\sin^2\theta\left(B, \tilde{B}\right) \leq \frac{\varepsilon}{\lambda_r\left(\alpha\alpha^\top\right)}.$$

C.5. Proof of Theorem 3

Similarly to Theorem 7 by Tripuraneni et al. (2021), we can show the following lemma. Its proof is omitted as it follows the exact same lines as the proof from (Tripuraneni et al., 2021).

Lemma 10. *For any set $S \subset [T]$ such that $\operatorname{card}(S) \geq c$, then with probability at least $1 - \frac{1}{m^2T^2}$:*

$$\left\| \frac{1}{m \operatorname{card}(S)} \sum_{i=1}^{m} \sum_{t \in S} \left((y_i^t)^2 x_i^t \left(x_i^t\right)^\top - \mathbb{E}\left[(y_i^t)^2 x_i^t \left(x_i^t\right)^\top \right] \right) \right\| \leq c' \log^3(dmT) \left(\sigma^2 + C\right) \left(\sqrt{\frac{d}{mT}} + \frac{d}{mT} \right),$$

for universal constants c and c'.

Lemma 11 below explicits the expected term in Lemma 10. It generalizes the Lemma 2 by Tripuraneni et al. (2021) to the case of any spherically symmetric distribution.

Lemma 11. *If x is spherically symmetric and $y = \langle u, x \rangle + \varepsilon$, where ε is independent from x, centered and σ-subgaussian, then*

$$\mathbb{E}\left[(y^t)^2 x^t \left(x^t\right)^\top \right] = \kappa u u^\top + \left(\operatorname{Var}(\varepsilon) + \gamma \|u\|^2 \right) I_d,$$

where $\gamma = \mathbb{E}[\langle e_1, x\rangle^2 \langle e_2, x\rangle^2]$ and $\kappa = \mathbb{E}[\langle e_1, x \rangle^4] - \gamma$.

Note that in the Gaussian case $\gamma = 1$ and $\kappa = 2$, hence recovering the previous result by Tripuraneni et al. (2021).

Proof. Using the relation between y and x, it comes:

$$\mathbb{E}\left[(y^t)^2 x^t \left(x^t\right)^\top \right] = \operatorname{Var}(\varepsilon) I_d + \mathbb{E}[x^\top u u^\top x x x^\top].$$

By rescaling and invariance under orthogonal transformations, it actually suffices to show Lemma 11. We then have, still by invariance under orthogonal transformation

$$\begin{aligned}
\left(\mathbb{E}[x^\top e_1 e_1^\top x x x^\top] \right)_{i,j} &= \mathbb{E}[\langle x, e_1\rangle^2 \langle x, e_i \rangle \langle x, e_j \rangle] \\
&= \begin{cases} \mathbb{E}[\langle x, e_1\rangle^4] & \text{if } i = j = 1, \\ \mathbb{E}[\langle x, e_1\rangle^2 \langle x, e_2\rangle^2] & \text{if } i = j \neq 1, \\ \mathbb{E}[\langle x, e_1\rangle^2 \langle x, e_2\rangle \langle x, e_3 \rangle] & \text{if } i \neq j \text{ and } i \neq 1 \text{ and } i \neq j, \\ \mathbb{E}[\langle x, e_1\rangle^3 \langle x, e_2\rangle] & \text{if } 1 = i \neq j \text{ or } 1 = j \neq i. \end{cases}
\end{aligned}$$

By considering the orthogonal transformation $e_2 \mapsto -e_2$, the term is 0 in the last two cases. This finally gives

$$\mathbb{E}[x^\top e_1 e_1^\top x x x^\top] = \kappa e_1 e_1^\top + \left(\operatorname{Var}(\varepsilon) + \gamma \|e_1\|^2 \right) I_d,$$

which leads to Lemma 11 by rescaling and rotating e_1 to $\frac{u}{\|u\|}$. ■

We can now bound the subspace estimation error due to the estimators $\hat{B}_1$ and $\hat{B}_2$ in Algorithm 1.

Lemma 12. *Consider $\hat{B}_1$, $\hat{B}_2$ defined in Algorithm 1 and B defined in Section 5. Then if $mT \geq c \frac{\log^6(dmT)\left(\sigma^4 + C^2\right) r^2}{\kappa^2 \nu^2} d$, with probability at least $1 - \frac{2}{m^2T^2}$:*

$$\sin\theta\left(\hat{B}_i, B \right) \leq c' \log^3(dmT) r \frac{\sigma^2 + C}{\kappa \bar{\nu}} \sqrt{\frac{d}{mT}}, \quad \textit{for } i = 1, 2,$$

where $\theta(B_i, B)$ is the principal angle between the subspaces induced by B_i and B, c and c' are universal constants.

Proof. The proof assumes that the concentration bound given by Lemma 10 holds and uses the Davis-Kahan theorem on it. Note that $\hat{B}_1$ derives from the top-r SVD of $A = \frac{1}{m\lfloor\frac{T}{2}\rfloor}\sum_{i=1}^m \sum_{t=\lceil\frac{T}{2}\rceil+1} (y_i^t)^2 x_i^t \left(x_i^t\right)^\top$.

Thanks to Lemmas 10 and 11, it holds with probability at least $1 - \frac{1}{m^2T^2}$ that

$$A = \kappa\bar{\Gamma} + \big(\mathrm{Var}(\varepsilon) + \gamma\mathrm{tr}(\bar{\Gamma})I_d\big) + E,$$

$$\text{with } \|E\| \leq c' \log^3(dmT)(\sigma^2 + C)\left(\sqrt{\frac{d}{mT}} + \frac{d}{mT}\right).$$

For T large enough, as given in Lemma 12, we have $\|E\| \leq \frac{\kappa\bar{\nu}}{2r}$. Now note that $\lambda_r(A-E) - \lambda_{r+1}(A-E) \geq \kappa\frac{\bar{\nu}}{r}$ and B corresponds to the top-r left singular vectors of the $A - E$.

Davis-Kahan theorem then yields that $\sin\theta\left(\hat{B}_1, B\right) \leq \frac{2r\|E\|}{\kappa\bar{\nu}}$. The same argument for $\hat{B}_2$ finally yields to Lemma 12. ■

Theorem 3 then follows from Lemma 12 and Theorem 4 from Tripuraneni et al. (2021), using the independence between the features used for the estimator $\hat{B}_i$ and the ones used for the estimator $\hat{\alpha}_i$.

It now just remains to show that $\kappa \geq 0$. We actually have the following series of (in)equalities:

$$\begin{aligned}\mathbb{E}[\langle x, e_1\rangle^2\langle x, e_2\rangle^2] &\leq \mathbb{E}[\langle x, e_1\rangle^2]\mathbb{E}[\langle x, e_2\rangle^2] \\ &= \mathbb{E}[\langle x, e_1\rangle^2]^2 \\ &\leq \mathbb{E}[\langle x, e_1\rangle^4].\end{aligned}$$

The first inequality derives from Cauchy-Schwarz inequality. The equality is by invariance over orthogonal transformation, while the last inequality comes from Jensen inequality. This directly implies that $\kappa \geq 0$.

Now show that the case of equality is impossible. In particular, it would imply from the first inequality that $\langle x, e_1\rangle^2 = \langle x, e_2\rangle^2$ almost surely. By invariance under rotation, we can even show that for any u of norm 1: $\langle x, e_1\rangle^2 = \langle x, u\rangle^2$, which implies that $x = 0$ almost surely, hence leading to a contradiction. We thus have $\kappa > 0$.