Algorithmic Regularization in Over-parameterized Matrix Sensing and Neural Networks with Quadratic Activations

Yuanzhi Li   YUANZHIL@CS.PRINCETON.EDU
Princeton University

Tengyu Ma   TENGYUMA@STANFORD.EDU
Facebook AI Research

Hongyang Zhang   HONGYANG@CS.STANFORD.EDU
Stanford University

Editors: Sébastien Bubeck, Vianney Perchet and Philippe Rigollet

Abstract

We show that the gradient descent algorithm provides an implicit regularization effect in the learning of over-parameterized matrix factorization models and one-hidden-layer neural networks with quadratic activations.

Concretely, we show that given $\tilde{O}(dr^2)$ random linear measurements of a rank r positive semidefinite matrix $X^\star$, we can recover $X^\star$ by parameterizing it by $UU^\top$ with $U \in \mathbb{R}^{d \times d}$ and minimizing the squared loss, even if $r \ll d$. We prove that starting from a small initialization, gradient descent recovers $X^\star$ in $\tilde{O}(\sqrt{r})$ iterations approximately. The results solve the conjecture of Gunasekar et al. Gunasekar et al. (2017) under the restricted isometry property.

The technique can be applied to analyzing neural networks with one-hidden-layer quadratic activations with some technical modifications.

Keywords: Generalization theory; Implicit regularization; Matrix factorization; Neural networks.

1. Introduction

Over-parameterized models are crucial in deep learning, but their workings are far from understood. Over-parameterization — the technique of using more parameters than statistically necessary — apparently improves the training: theoretical and empirical results have suggested that it can enhance the geometric properties of the optimization landscape in simplified settings Livni et al. (2014); Hardt et al. (2016); Hardt and Ma (2017); Soudry and Carmon (2016) and thus make it easier to train over-parameterized models.

On the other hand, over-parameterization often doesn't hurt the test performance, even if the number of parameters is much larger than the number of examples. Large neural networks used in practice have enough expressiveness to fit any labels of the training datasets Zhang et al. (2016); Hardt and Ma (2017). The training objective function may have multiple global minima with almost zero training error, some of which generalize better than the others Keskar et al. (2016); Dinh et al. (2017). However, local improvement algorithms such as stochastic gradient descent, starting with proper initialization, may prefer some generalizable local minima to the others and thus provide an implicit effect of regularization Srebro et al. (2011); Neyshabur et al. (2014); Hardt et al. (2015); Neyshabur et al. (2017b); Wilson et al. (2017). Such regularization seems to depend on the algorithmic choice, the initialization scheme, and certain intrinsic properties of the data.

The phenomenon and intuition above can be theoretically fleshed out in the context of linear models Soudry et al. (2017), whereas less is known for non-linear models whose training objectives are usually non-convex. The very important work of Gunasekar et al. Gunasekar et al. (2017) initiates the study of low-rank matrix factorization models with over-parameterization and conjectures that gradient descent prefers small trace norm solution in over-parameterized models with thorough empirical evidences.

This paper resolves the conjecture for the matrix sensing problem — recovering a low-rank matrix from linear measurements — under the restricted isometry property (RIP). We show that with a full-rank factorized parameterization, gradient descent on the squared loss with finite step size, starting with a small initialization, converges to the true low-rank matrix (which is also the minimum trace norm solution.) One advantage of the over-parameterized approach is that without knowing/guessing the correct rank, the algorithms can automatically pick up the minimum rank or trace norm solution that fits the data.

The analysis can be extended to learning one-hidden-layer neural networks with quadratic activations. We hope such theoretical analysis of algorithmic regularization in the non-convex setting may shed light on other more complicated models where over-parameterization is crucial (if not necessary) for efficient training.

1.1. Setup and Main Results

Let $X^\star$ be an unknown rank-r symmetric positive semidefinite (PSD) matrix in $\mathbb{R}^{d \times d}$ that we aim to recover. Let $A_1, \cdots, A_m \in \mathbb{R}^{d \times d}$ be m given symmetric measurement matrices.[1] We assume that the label vector $y \in \mathbb{R}^m$ is generated by linear measurements

$$y_i = \langle A_i, X^\star \rangle.$$

Here $\langle A, B \rangle = \text{tr}(A^\top B)$ denotes the inner product of two matrices. Our goal is to recover the matrix $X^\star$.[2]

Without loss of generality, we assume that $X^\star$ has spectral norm 1. Let $\sigma_r(X)$ denote the $r-th$ singular value of a matrix X, and let $\kappa = 1/\sigma_r(X^\star)$ be the condition number of $X^\star$. We focus on the regime where $r \ll d$ and $m \approx d \cdot \text{poly}(r \log d) \ll d^2$.

Let $U \in \mathbb{R}^{d \times d}$ be a matrix variable. We consider the following mean squared loss objective function with over-parameterization:

$$\min_{U \in \mathbb{R}^{d \times d}} f(U) = \frac{1}{2m} \sum_{i=1}^m \left(y_i - \langle A_i, UU^\top \rangle \right)^2 \tag{1}$$

Since the label is generated by $y_i = \langle A_i, X^\star \rangle$, any matrix U satisfying $UU^\top = X^\star$ is a local minimum of f with zero training error. These are the ideal local minima that we are shooting for. However, because the number of parameters d^2 is much larger than the number of observation m, there exist other choices of U satisfying $f(U) = 0$ but $UU^\top \neq X^\star$.

---

1. Given that the matrix $X^\star$ is symmetric, we can assume that A_i's are symmetric without loss of generality: Because $\langle A_i, X^\star \rangle = \langle \frac{1}{2}(A_i + A_i^\top), X^\star \rangle$ for any symmetric matrix $X^\star$, we can always replace A_i by $\frac{1}{2}(A_i + A_i^\top)$.
2. Our analysis can naturally handle a small amount of Gaussian noise in the label vector y, but for simplicity we only work with the noiseless case.

A priori, such over-parameterization will cause over-fitting. However, we will show that the following gradient descent algorithm with small initialization converges to a desired local minimum, instead of other non-generalizable local minima:

$$U_0 = \alpha B, \text{ where } B \in \mathbb{R}^{d \times d} \text{ is any orthonormal matrix}$$
$$U_{t+1} = U_t - \eta \nabla f(U_t) \tag{2}$$

The following theorem assumes the measurements matrices $A_1, \ldots, A_m$ satisfy restricted isometry property (RIP), which is formally defined in Section 2. Casual readers may simply assume that the entries of A_i's are drawn i.i.d from standard normal distribution[3] and the number of observations $m \lesssim dr^2 \log^3 d$: it's known Recht et al. (2010) that in this case $A_1, \ldots, A_m$ meet the requirement of the following theorem, that is, they satisfy $(4r, \delta)$-RIP with $\delta \lesssim 1/(\sqrt{r} \log d)$ with high probability. [4]

Theorem 1.1 *Let c be a sufficiently small absolute constant. Assume that the set of measurement matrices $(A_1, \ldots, A_m)$ satisfies $(4r, \delta)$-restricted isometry property (defined in Section 2 formally) with $\delta \leq c/(\kappa^3 \sqrt{r} \log^2 d)$. Suppose the initialization and learning rate satisfy $0 < \alpha \leq c \min\{\delta \sqrt{r}\kappa, 1/d\}$ and $\eta \leq c\delta$. Then for every $(\kappa \log(\frac{d}{\alpha}))/\eta \lesssim T \lesssim 1/(\eta \sqrt{d}\kappa\alpha)$, we have*

$$\|U_T U_T^\top - X^\star\|_F^2 \lesssim \alpha\sqrt{d}/\kappa^2.$$

Note that the recovery error $\|U_T U_T^\top - X^\star\|_F^2$ can be viewed as the test error (defined in Equation (8) formally) — it's the expectation of the test error on a fresh measurement A_j drawn from the standard normal distribution. The theorem above shows that gradient descent can provide an algorithmic regularization so that the generalization error depends on the size of the initialization α, instead of the number of parameters. Because the convergence is not very sensitive to the initialization, we can choose small enough α (e.g., $1/d^5$) to get approximately zero generalization error. Moreover, when α is small, gradient descent can run for a long period of time without overfitting the data. We show in Section B that empirically indeed the generalization error depends on the size of the initialization and gradient descent is indeed very stable.

The analysis also applies to *stochastic gradient descent*, as long as each batch of the measurement matrices satisfies RIP.[5] We also remark that our theory suggests that early stopping for this problem is not necessary when the initialization is small enough — the generalization error bounds apply until $1/(\eta\sqrt{d}\kappa\alpha)$ iterations. We corroborate this with empirical results in Section B.

We remark that we achieve a good iteration complexity bound $1/\eta \approx 1/\delta \approx \sqrt{r}$ for the gradient descent algorithm, which was not known in previous work even for low-rank parameterization, nor for the case with infinite samples (which is the PCA problem). Part of the technical challenges is to allow finite step size η and inverse-poly initialization α (instead of exponentially small initialization). The dependency of δ on κ and r in the theorem is possibly not tight. We conjecture that δ only needs to be smaller than an absolute constant, which is left for future work.

---

3. Or equivalently, as discussed in the previous footnote, causal readers may assume $A_i = \frac{1}{2}(Q_i + Q_i^\top)$ where Q_i is from standard normal distribution. Such symmetrization doesn't change the model since $\langle Q_i, X^\star \rangle = \langle \frac{1}{2}(Q_i + Q_i^\top), X^\star \rangle$

4. Technically, to get such RIP parameters that depends on r, one would need to slightly modify the proof of (Recht et al., 2010, Theorem 4.2) at the end to get the dependency of m on δ.

5. Smaller batch size should also work when the learning rate is sufficiently small, although its analysis seems to require more involved techniques and is left for future work.

Insights of the analysis: Interestingly, our analysis ideas seem to be different from other previous work in a conceptual sense. The analysis of the logistic regression case Soudry et al. (2017) relies on that the iterate eventually moves to infinity. The folklore analysis of the algorithmic regularization of SGD for least squares and the analysis in Gunasekar et al. (2017) for the matrix regression with commutable measurements both follow the two-step plan: a) the iterates always stays on a low-rank manifold that only depends on the inputs (the measurement matrices) but not on the label vector y; b) generalization follows from the low complexity of the low-rank manifold. Such input-dependent but label-independent manifold doesn't seem to exist in the setting when A_i's are random.

Instead, we show that the iterates stay in the set of matrices with approximate rank smaller or equal to the minimal possible rank that can fit the data, which is a set that depends on the labels y but not on the inputs A_i's. We implicitly exploit the fact that gradient descent on the population risk with small initialization only searches through the space of solutions with a *lower* rank than that of the true matrix $X^\star$. The population risk is close to the empirical risk on matrices with rank smaller than or equal to the true rank. Hence, we can expect the learning dynamics of the empirical risk to be similar to that of the population risk, and therefore the iterates of GD on the empirical risk remain approximately low-rank as well. Generalization then follows straightforwardly from the low-rankness of the iterates. See Section 3 for more high-level discussions.

We note that the factorized parameterization also plays an important role here. The intuition above would still apply if we replace $UU^\top$ with a single variable X and run gradient descent in the space of X with small enough initialization. However, it will converge to a solution that *doesn't* generalize. The discrepancy comes from another crucial property of the factorized parameterization: it provides certain denoising effect that encourages the empirical gradient to have a smaller eigenvalue tail. This ensures the eigenvalues tails of the iterates to grow sufficiently slowly. This point will be more precise in Section 3 once we apply the RIP property. In section B, we also empirically demonstrate that GD in the original space of $X^\star$ with projection to the PSD cone doesn't provide as good generalization performance as GD in the factorized space.

Finally, we remark that the cases with rank $r > 1$ are technically much more challenging than the rank-1 case. For the rank-1 case, we show that the spectrum of U_t remains small in a fixed rank-$(d-1)$ subspace, which is exactly the complement of the column span of $X^\star$. Hence the iterates are approximately rank one. By contrast, for the rank-r case, a direct extension of this proof strategy only gives a much weaker result compared to Theorem 1.1. Instead, we identify an *adaptive* rank-$(d-r)$ subspace in which U_t remains small. Clearly, the best choice of this adaptive subspace is the subspace of the least $(d-r)$ left singular vectors of U_t. However, we use a more convenient surrogate. We refer the reader to Section 4 for detailed descriptions.

1.2. Extensions to Neural Networks with Quadratic Activations

Our results can be applied to learning one-hidden-layer neural networks with quadratic activations. We setup the notations and state results below and defer the details to Section A.

Let $x \in \mathbb{R}^d$ be the input and $U^\star \in \mathbb{R}^{d \times r}$ be the first layer weight matrix. We assume that the weight on the second layer is simply the all one's vector $\mathbf{1} \in \mathbb{R}^r$. Formally, the label y is assumed to be generated by

$$y = \mathbf{1}^\top q(U^{\star\top} x) \tag{3}$$

where $q(\cdot)$ is the element-wise quadratic function. For simplicity, we assume that x comes from standard normal distribution $\mathcal{N}(0, \mathrm{Id}_{d \times d})$. It's not hard to see that the representational power of the

hypothesis class with $r = d$ is the same as those with $r > d$. Thus we only focus on the case when $r \leq d$. For the purpose of this paper, the most interesting regime is the scenario when $r \ll d$.

We use an over-parameterized model with a variable $U \in \mathbb{R}^{d \times d}$. The prediction $\hat{y}$ is parameterized by $\hat{y} = \mathbf{1}^\top q(U^\top x)$, and we use the mean squared error $(y - \hat{y})^2$ as the loss function. We use a variant of stochastic gradient descent (or gradient descent) on the mean squared loss.

The following theorem shows that the learned model will generalize with $\tilde{O}(dr^5\kappa^6)$ examples, despite that the number of parameters d^2 can be much larger than the number of samples (when $d \gg r$ or r is considered as a constant).[6] We will start with an initialization U_0 in the same way as in equation (2), and denote $U_1, \ldots, U_T$ as the iterates. Let κ be the condition number of $U^\star U^{\star\top}$.

Theorem 1.2 *Given $\tilde{O}(dr^5\kappa^6)$ examples, a variant of gradient descent (Algorithm 1 in Section A) with initialization $\alpha \lesssim \min\{1/d, 1/(r^2\kappa^4\log^2 d)\}$ and learning rate $\eta \lesssim \frac{1}{\kappa^3 r^{1.5}\log^2 d}$ returns a solution with generalization error at most $O(d\kappa\alpha)$ at any iteration t such that $(\kappa\log(d/\alpha))/\eta \lesssim t \lesssim 1/(\eta\sqrt{d\kappa\alpha})$.*

The same analysis also applies to stochastic gradient descent as long as the batch size is at least $\gtrsim dr^5\kappa^6$. The analysis exploits the connection (pointed out by Soltanolkotabi et al. (2017)) between neural networks with quadratic activations and matrix sensing with rank-1 measurements Kueng et al. (2017); Zhong et al. (2015); Chen et al. (2015): one can view $xx^\top$ as the measurement matrix in matrix sensing. However, these measurements don't satisfy the RIP property. We will modify the learning algorithm slightly to cope with it. See Section A for details.

Organization: The rest of this paper is organized as follows: In Section 2, we define notations and present a review of the restricted isometry property. In Section 3, we lay out the key theoretical insights towards proving Theorem 1.1 and give the analysis for the rank-1 case as a warm-up. In Section 4, we outline the main steps for proving Theorem 1.1 and Section D completes the proofs of these steps. Section A and Section E give the proof of Theorem 1.2. Section B contains numeric simulations. Finally, Section F provide the proofs of concentration properties we have used.

Notations: Let Id_U denotes the projection to the column span of U, and let Id denotes the identity matrix. Let U^+ denote the Moore-Penrose pseudo-inverse of the matrix U. Let $\|\cdot\|$ denotes the Euclidean norm of a vector and spectral norm of a matrix. Let $\|\cdot\|_F$ denote the Frobenius norm of a matrix. Suppose $A \in \mathbb{R}^{m \times n}$, then $\sigma_{\max}(A)$ denote its largest singular value and $\sigma_{\min}(A)$ denotes its $\min\{m, n\}$-th largest singular value. Alternatively, we have $\sigma_{\min}(A) = \min_{x:\|x\|=1} \|Ax\|$. Let $\langle A, B \rangle = \mathrm{tr}(A^\top B)$ denote the inner product of two matrices. We use $\sin(A, B)$ to denote the sine of the principal angles between the columns spaces of A and B.

Unless explicitly stated otherwise, $O(\cdot)$-notation hides absolute multiplicative constants. Concretely, every occurrence of $O(x)$ is a placeholder for some function $f(x)$ that satisfies $\forall x \in \mathbb{R}, |f(x)| \leq C|x|$ for some absolute constant $C > 0$. Similarly, $a \lesssim b$ means that there exists an absolute constant $C > 0$ such that $a \leq Cb$. We use the notation $\mathrm{poly}(n)$ as an abbreviation for $n^{O(1)}$.

---

6. The dependency on r here is likely to be loose. Although we note that this is the first bound of this kind for this problem that shows over-parameterized models can still generalize.

2. Preliminaries and Related Work

Recall that we assume $X^\star$ is rank-r and positive semidefinite. Let $X^\star = U^\star \Sigma^\star U^{\star\top}$ be the eigen-decomposition of $X^\star$, where $U^\star \in \mathbb{R}^{d \times r}$ is an orthonormal matrix and $\Sigma^\star \in \mathbb{R}^{r \times r}$ is a diagonal matrix. The assumptions that $\|X^\star\| = 1$ and $\sigma_r(X^\star) = 1/\kappa$ translate to that $\forall i \in [r], 1/\kappa \leq \Sigma_{ii}^\star \leq 1$. Under the above notations, we see that the target solution for the variable U is equal to $U = U^\star \Sigma^{\star 1/2} R$ where R can be arbitrary orthonormal matrix. For convenience, we define the matrix M_t as

$$M_t = \frac{1}{m} \sum_{i=1}^m \langle A_i, U_t U_t^\top - X^\star \rangle A_i \qquad (4)$$

Then the update rule can be rewritten as

$$U_{t+1} = (\mathrm{Id} - \eta M_t) U_t \qquad (5)$$

where Id is the identity matrix. One of the key challenges is to understand how the matrix $\mathrm{Id} - \eta M_t$ transforms U_t, so that U_0 converges the target solution $U^\star \Sigma^{\star 1/2} R$ quickly.

Suppose that $A_1, \ldots, A_m$ are drawn from Gaussian distribution and optimistically suppose that they are *independent* with U_t. Then, we have that $M_t \approx U_t U_t^\top - X^\star$, since the expectation of M_t with respect to the randomness of A_i's is equal to $U_t U_t^\top - X^\star$. However, they are two fundamental issues with this wishful thinking: a) obviously U_t depends on A_i's heavily for $t > 1$, since in every update step A_i's are used; b) even if A_i's are independently with U_t, there are not enough A_i's to guarantee M_t concentrates around its mean $U_t U_t^\top - X^\star$ in Euclidean norm. To have such concentration, we need $m > d^2$, whereas we only have $m = d \times \mathrm{poly}(r \log d)$ samples.

Restricted isometry propety: The restricted isometry property (RIP) allows us to partially circumvent both the technical issues a) and b) above. It says that using the set of linear measurement matrices $A_1, \ldots, A_m$, we can preserve the Frobenius norm of any rank-r matrices approximately.

Definition 2.1 *(Restricted isometry property Recht et al. (2010)) A set of linear measurement matrices $A_1, \ldots, A_m$ in $\mathbb{R}^{d \times d}$ satisfies (r, δ)-restricted isometry property (RIP) if for any $d \times d$ matrix X with rank at most r, we have*

$$(1 - \delta)\|X\|_F^2 \leq \frac{1}{m} \sum_{i=1}^m \langle A_i, X \rangle^2 \leq (1 + \delta)\|X\|_F^2. \qquad (6)$$

The crucial consequence of RIP that we exploit in this paper is the meta statement as follows:

$$\mathcal{M}(Q) := \frac{1}{m} \sum_{i=1}^m \langle A_i, Q \rangle A_i \text{ behaves like } Q \text{ for approximately low-rank } Q \qquad (7)$$

We will state several lemmas below that reflect the principle above. The following lemma says that $\langle \mathcal{M}(X), Y \rangle$ behaves like $\langle X, Y \rangle$ for low-rank matrices X and Y.

Lemma 2.2 *(Candes, 2008, Lemma 2.1) Let $\{A_i\}_{i=1}^m$ be a family of matrices in $\mathbb{R}^{d \times d}$ that satisfy (r, δ)-restricted isometry property. Then for any matrices $X, Y \in \mathbb{R}^{d \times d}$ with rank at most r, we have:*

$$\left| \frac{1}{m} \sum_{i=1}^m \langle A_i, X \rangle \langle A_i, Y \rangle - \langle X, Y \rangle \right| \leq \delta \|X\|_F \|Y\|_F$$

The following lemma says that $\mathcal{M}(X)$ behaves like X when multiplied by a matrix R with small operator norm.

Lemma 2.3 *Let $\{A_i\}_{i=1}^m$ be a family of matrices in $\mathbb{R}^{d\times d}$ that satisfy (r, δ)-restricted isometry property. Then for any matrix $X \in \mathbb{R}^{d\times d}$ of rank at most r, and any matrix $R \in \mathbb{R}^{d\times d'}$, where d' can be any positive integer, we have:*

$$\left\| \frac{1}{m} \sum_{i=1}^m \langle A_i, X \rangle A_i R - XR \right\| \leq \delta \|X\|_F \cdot \|R\|.$$

Lemma 2.3 is proved in Section F[7]. We can also extend Lemma 2.3 to the cases when X has a higher rank (see Lemma F.1 and Lemma F.2). The bounds are not as strong as above (which is inevitable because we only have m measurements), but are useful when X itself is relatively small.

2.1. Related Work

Generalization theory beyond uniform convergence: This work builds upon the remarkable work of Gunasekar et al. Gunasekar et al. (2017), which raises the conjecture of the implicit regularization in matrix factorization models and provides theoretical evidence for the simplified setting where the measurements matrices are commutable. Implicit regularization of gradient descent is studied in the logistic regression setting by Soudry et al. Soudry et al. (2017).

Recently, the work of Hardt et al. Hardt et al. (2015) studies the implicit regularization provided by stochastic gradient descent through uniform stability Bousquet and Elisseeff (2002); Mukherjee et al. (2006); Shalev-Shwartz et al. (2010). Since the analysis therein is independent of the training labels and therefore it may give pessimistic bounds Zhang et al. (2017). Brutzkus et al. Brutzkus et al. (2017) use a compression bound to show network-size independent generalization bounds of one-hidden-layer neural networks on linearly separable data.

Bartlett et al. Bartlett et al. (2017), Neyshabur et al. Neyshabur et al. (2017a), and Cisse et al. Cisse et al. (2017) recently prove spectrally-normalized margin-based generalization bounds for neural networks. Dziugaite and Roy Dziugaite and Roy (2017) provide non-vacuous generalization bounds for neural networks from PCA-Bayes bounds. As pointed out by Bartlett et al. Bartlett et al. (2017), it's still unclear why SGD without explicit regularization can return a large margin solution. This paper makes progress on explaining the regularization power of gradient descent, though on much simpler non-linear models.

Matrix factorization problems: Early works on matrix sensing and matrix factorization problems use convex relaxation (nuclear norm minimization) approaches and obtain tight sample complexity bounds Recht et al. (2010); Srebro and Shraibman (2005); Candès and Recht (2009); Recht (2011); Candès et al. (2011). Tu et al. Tu et al. (2015) and Zheng and Lafferty Zheng and Lafferty (2016) analyze the convergence of non-convex optimization algorithms from spectral initialization. The recent work of Ge et al. Ge et al. (2016) and Bhojanapalli et al. Bhojanapalli et al. (2016) shows that the non-convex objectives on matrix completion and matrix sensing with low-rank parameterization don't have any spurious local minima, and stochastic gradient descent algorithm on them converges to the global minimum. Such a phenomenon was already known for the PCA problem and recently shown for phase retrieval, robust PCA, and random tensor decomposition as well

---

7. We suspect that Lemma 2.3 is already known, however we haven't been able to find a reference.

(e.g., see Srebro and Jaakkola (2013); Ge et al. (2016); Bhojanapalli et al. (2016); Ge et al. (2017); Ge and Ma (2017); Sun et al. (2016) and references therein). Soltanolkotabi et al. Soltanolkotabi et al. (2017) analyzes the optimization landscape of over-parameterized one-hidden-layer neural networks with quadratic activations. Empirically, Jose et al. Jose et al. (2017) show that factorized parameterizations of recurrent neural networks provide additional regularization effect.

3. Proof Overview and Rank-1 Case

In this section, we demonstrate the key ideas of our proofs and give an analysis of the rank-1 case as a warm-up. The main intuition is that the iterate U_t stays approximately low-rank in the sense that:

(a) The $(r+1)$-th singular value $\sigma_{r+1}(U_t)$ remains small for any $t \geq 0$;

(b) The top r singular vectors and singular values of $U_t U_t^\top$ converge to those of $X^\star$ in logarithmic number of iterations.

Propositions (a) and (b) can be clearly seen when the number of observations m approaches infinity and $A_1, \ldots, A_m$ are Gaussian measurements. Let's define the population risk $\bar{f}$ as

$$\bar{f}(U_t) = \mathbb{E}_{(A_i)_{k\ell} \sim N(0,1)} [f(U_t)] = \|U_t U_t^\top - X^\star\|_F^2 \tag{8}$$

In this case, the matrix M_t (defined in (4)) corresponds to $U_t U_t^\top - X^\star$, and therefore the update rule for U_t can be simply rewritten as

$$U_{t+1} = U_t - \eta \nabla \bar{f}(U_t) = U_t - \eta (U_t U_t^\top - X^\star) U_t$$
$$= U_t(\operatorname{Id} - \eta U_t^\top U_t) + \eta X^\star U_t$$

Observe that the term $\eta X^\star U_t$ encourages the column span of U_{t+1} to move towards the column span of $X^\star$, which causes the phenomenon in Proposition (b). On the other hand, the term $U_t(\operatorname{Id} - \eta U_t U_t^\top)$ is performing a contraction of all the singular values of U_t, and therefore encourages them to remain small. As a result, U_t decreases in those directions that are orthogonal to the span of $X^\star$, because there is no positive force to push up those directions.

So far, we have described intuitively that the iterates of GD on the population risk remains approximately low-rank. Recall that the difficulty was that the empirical risk f doesn't uniformly concentrate well around the population risk $\bar{f}$.[8] However, the uniform convergence can occur, at least to some extent, in the restricted set of approximately low-rank matrices! In other words, since the gradient descent algorithm only searches a limited part of the whole space, we only require restricted uniform convergence theorem such as restricted isometry property. Motivated by the observations above, a natural meta proof plan is that:

1. The trajectory of the iterates of GD on the population risk stays in the set of approximately low-rank matrices.

2. The trajectory of the empirical risk behaves similarly to that of the population risk in the set of approximately low-rank matrices.

---

8. Namely, we don't have uniform convergence results in the sense that $|f(U) - \bar{f}(U)|$ is small for all matrices U. (For examples, for many matrices we can have $f(U) = 0$ but $\bar{f}(U) \gg 0$ because we have more variables than observations.)

It turns out that implementing the plan above quantitatively can be technically challenging: the distance from the iterate to the set of low-rank matrices can accumulate linearly in the number of steps. Therefore we have to augment the plan with a strong result about the rate of convergence:

3. The iterates converge to the true solution $X^\star$ fast enough before its effective rank increases.

For the rest of this section, we demonstrate a short proof of the rank-1 case to implement the intuitions described above. We note that the results of the rank-r case in Section 4 is significantly tighter than the results presented in this section. The analysis involves more delicate techniques to control the growth of the top r eigenvectors, and requires a much stronger convergence analysis.

3.1. Warm-up: Rank-1 Case

In this subsection, we assume that $X^\star = u^\star u^{\star\top}$ for $u^\star \in \mathbb{R}^{d \times 1}$ and that $\|u^\star\| = 1$. We decompose the iterates U_t into the subspace of $u^\star$ and its complement:

$$U_t = \mathrm{Id}_{u^\star} U_t + (\mathrm{Id} - \mathrm{Id}_{u^\star})U_t$$
$$:= u^\star r_t^\top + E_t \tag{9}$$

where we denote by $r_t := U_t^\top u^\star$ and $E_t := (\mathrm{Id} - \mathrm{Id}_{u^\star})U_t$. [9]

In light of the meta proof plan discussed above, we will show that the spectral norm and Frobenius norm of the "error term" E_t remains small throughout the iterations, whereas the "signal" term $u^\star r_t^\top$ grows exponentially fast (in the sense that the norm of r_t grows to 1.) Note that any solution with $\|r_t\| = 1$ and $E_t = 0$ will give exact recovery, and for the purpose of this section we will show that $\|r_t\|$ will converges approximately to 1 and E_t stays small.

Under the representation (9), from the original update rule (5), we derive the update for E_t:

$$E_{t+1} = (\mathrm{Id} - \mathrm{Id}_{u^\star}) \cdot (\mathrm{Id} - \eta M_t)U_t$$
$$= E_t - \eta \cdot (\mathrm{Id} - \mathrm{Id}_{u^\star})M_t U_t \tag{10}$$

Throughout this section, we assume that the set of measurement matrices $(A_1, \ldots, A_m)$ satisfies $(4, \delta)$-RIP with $\delta \le c$ where c is a sufficiently small absolute constant (e.g., $c = 0.01$ suffices).

Theorem 3.1 *In the setting of this subsection, suppose $\alpha \le \delta\sqrt{\frac{1}{d}\log\frac{1}{\delta}}$ and $\eta \lesssim c\delta^2 \log^{-1}(\frac{1}{\delta\alpha})$. Then after $T = \Theta(\log\frac{1}{\alpha\delta}/\eta)$ iterations, we have:*

$$\|U_T U_T^\top - X^\star\|_F \le O(\delta\log\frac{1}{\delta})$$

As we already mentioned, Theorem 3.1 is weaker than Theorem 1.1 even for the case with $r = 1$. In Theorem 1.1 (or Theorem 4.1), the final error depends linearly on the initialization, whereas the error here depends on the RIP parameter. Improving Theorem 3.1 would involve finer inductions, and we refer the readers to Section 4 for the stronger results.

The following Proposition gives the growth rate of E_t in spectral norm and Euclidean norm, in a typical situation when E_t and r_t are bounded above in Euclidean norm. The proof is presented in Section C.

---

9. Observe that we have restricted the column subspace of the signal term $R_t = u^\star r_t^\top$, so that $R_t R_t^\top$ is always a multiple of $X^\star$. In section 4, we will introduce an adaptive subspace instead to decompose U_t into the signal and the error terms.

Proposition 3.2 (Error dynamics) *In the setting of Theorem 3.1. Suppose that $\|E_t\|_F \leq 1/2$ and $\|r_t\|^2 \leq 3/2$. Then E_{t+1} can be bounded by*

$$\|E_{t+1}\|_F^2 \leq \|E_t\|_F^2 + 2\eta\delta\|E_t U_t^\top\| + 9\eta^2. \tag{11}$$
$$\|E_{t+1}\| \leq (1 + 2\eta\delta)\|E_t\| + 2\eta\delta\|r_t\|.$$

A recurring technique in this section, as alluded before, is to establish the approximation

$$U_{t+1} = U_t - \eta M_t U_t \approx U_t - \eta(U_t U_t - X^\star)U_t$$

As we discussed in Section 2, if $U_t U_t - X^\star$ is low-rank, then the approximation above can be established by Lemma 2.2 and Lemma 2.3. However, $U_t U_t - X^\star$ is only approximately low-rank, and we therefore we will decompose it into

$$U_t U_t^\top - X^\star = \underbrace{(U_t U_t^\top - X^\star - E_t E_t^\top)}_{\text{rank} \leq 4} + \underbrace{E_t E_t^\top}_{\text{second-order in } E_t} \tag{12}$$

Note that $U_t U_t^\top - X^\star - E_t E_t^\top = \|r_t\|^2 u^\star u^{\star\top} + u^\star r_t^\top E_t^\top + E_t r_t u^{\star\top} - X^\star$ has rank at most 4, and therefore we can apply Lemma 2.2 and Lemma 2.3. For the term $E_t E_t^\top$, we can afford to use other looser bounds (Lemma F.1 and F.2) because E_t itself is small.

The next Proposition shows that the signal term grows very fast, when the signal itself is not close to norm 1 and the error term E_t is small.

Proposition 3.3 (Signal dynamics) *In the same setting of Proposition 3.2, we have,*

$$\left\|r_{t+1} - (1 + \eta(1 - \|r_t\|^2))r_t\right\| \leq \eta\|E_t\|^2\|r_t\| + 2\eta\delta(\|E_t\| + \|r_t\|). \tag{13}$$

The following proposition shows that $\|r_t\|$ converges to 1 approximately and E_t remains small by inductively using the two propositions above.

Proposition 3.4 (Control r_t and E_t by induction) *In the setting of Theorem 3.1, when the number of iterations is $T = \Theta(\log(\frac{1}{\alpha\delta}))/\eta)$,*

$$\|r_T\| = 1 \pm O(\delta) \tag{14}$$
$$\|E_T\|_F^2 \lesssim \delta^2 \log(1/\delta) \tag{15}$$

The proofs of the above Propositions are in Section C. Theorem 3.1 follows from Proposition 3.4 straightforwardly.

Proof [Proof of Theorem 3.1] Using the conclusions of Proposition 3.4, we have

$$\left\|U_T U_T^\top - X^\star\right\|_F^2 = (1 - \|r_T\|^2)^2 + 2\|E_T r_T\|^2 + \|E_T E_T^\top\|_F^2$$
$$\leq (1 - \|r_T\|^2)^2 + 2\|E_T\|^2\|r_T\|^2 + \|E_T\|_F^4$$
$$\leq O(\delta^2) + O(\delta^2 \log^2 \frac{1}{\delta}) + O(\delta^4 \log^2(\frac{1}{\delta})) = O(\delta^2 \log^2 \frac{1}{\delta})$$

$\blacksquare$

4. Proof Outline of Rank-r Case

In this section we outline the proof of Theorem 1.1. The proof is significantly more sophisticated than the rank-1 case (Theorem 3.1), because the top r eigenvalues of the iterates grow at different speed. Hence we need to align the signal and error term in the right way so that the signal term grows monotonically. Concretely, we will decompose the iterates into a signal and an error term according to a dynamic subspace, as we outline below. Moreover, the generalization error analysis here is also tighter than Theorem 3.1. We first state a slightly stronger version of Theorem 1.1:

Theorem 4.1 *There exists a sufficiently small absolute constant $c > 0$ such that the following is true. For every $\alpha \in (0, c/d)$, assume that the set of measurement matrices $(A_1, \ldots, A_m)$ satisfies (r, δ)-restricted isometry property with $\delta \leq c/(\kappa^3 \sqrt{r} \log^2 \frac{d}{\alpha})$, $\eta \leq c\delta$, and let T_0 be equal to $\max \left\{ \frac{\kappa \log(d/\alpha)}{\eta}, \frac{1}{\eta \sqrt{d\kappa\alpha}} \right\}$. For every $t \lesssim T_0$,*

$$\left\| U_t U_t^\top - X^\star \right\|_F^2 \leq (1 - \eta/(8\kappa))^{t-T_0} + O(\alpha\sqrt{d}/\kappa^2).$$

As a consequence, for $T_1 = \Theta((\kappa \log(\frac{d}{\alpha}))/\eta)$, we already have

$$\left\| U_{T_1} U_{T_1}^\top - X^\star \right\|_F^2 \lesssim \alpha\sqrt{d}/\kappa^2.$$

When the condition number κ and rank r are both constant, this theorem says that if we shoot for a final error ε, then we should pick our initialization $U_0 = \alpha B$ with $\alpha = O(\varepsilon/d)$. As long as the RIP-parameter $\delta = O(\frac{1}{\log \frac{d}{\varepsilon}})$, after $O(\log \frac{d}{\varepsilon})$ iterations we will have that $\left\| U_t U_t^\top - X^\star \right\|_F^2 \leq \varepsilon$.

Towards proving the theorem above, we suppose the eigen-decomposition of $X^\star$ can be written as $X^\star = U^\star \Sigma^\star U^{\star\top}$ where $U^\star \in \mathbb{R}^{d \times r}$ is an orthonormal matrix $\Sigma^\star \in \mathbb{R}^{r \times r}$ is a diagonal matrix. We maintain the following decomposition of U_t throughout the iterations:

$$U_t = \underbrace{\mathrm{Id}_{S_t} U_t}_{:=Z_t} + \underbrace{(\mathrm{Id} - \mathrm{Id}_{S_t}) U_t}_{:=E_t} \tag{16}$$

Here S_t is r dimensional subspace that is recursively defined by

$$S_0 = \mathrm{span}(U^\star) \tag{17}$$
$$S_t = (\mathrm{Id} - \eta M_{t-1}) \cdot S_{t-1}, \ \forall t \geq 1. \tag{18}$$

Here $(\mathrm{Id} - \eta M_t) \cdot S_{t-1}$ denotes the subspace $\{(\mathrm{Id} - \eta M_t) v : v \in S_{t-1}\}$. Note that $\mathrm{rank}(S_0) = \mathrm{rank}(U^\star) = r$, and thus by induction we will have that for every $t \geq 0$,

$$\mathrm{span}(Z_t) \subset \mathrm{span}(S_t),$$
$$\mathrm{rank}(Z_t) \leq \mathrm{rank}(S_t) \leq r.$$

Note that by comparison, in the analysis of rank-1 case, the subspace S_t is chosen to be $\mathrm{span}(U^\star)$ for every t, but here it starts off as $\mathrm{span}(U^\star)$ but changes throughout the iterations. We will show that S_t stays close to $\mathrm{span}(U^\star)$. Moreover, we will show that the error term E_t — though growing exponentially fast — always remains much smaller than the signal term Z_t, which grows exponentially with a faster rate. Recall that $\sin(A, B)$ denotes the principal angles between the column span of matrices A, B. We summarize the intuitions above in the following theorem.

11

Theorem 4.2 *There exists a sufficiently small absolute constant $c > 0$ such that the following is true: For every $\alpha \in (0, c/d)$, assume that the set of measurement matrices $(A_1, \ldots, A_m)$ satisfies (r, δ)-RIP with $\delta \leq c^4/(\kappa^3 \sqrt{r} \log^2 \frac{d}{\alpha})$. $\rho = O(\frac{\sqrt{r}\delta}{\kappa \log(\frac{d}{\alpha})})$, $\eta \leq c\delta$. Then for $t \leq T_1 = \Theta(\frac{\kappa}{\eta} \log(\frac{d}{\alpha}))$, we have that*

$$\sin(Z_t, U^\star) \lesssim \eta\rho t \tag{19}$$

$$\|E_t\| \leq (1 + O(\eta^2 \rho t))^t \|E_0\| \leq 4\|E_0\| \leq 1/d \tag{20}$$

$$\sigma_{\min}(U^{\star\top} Z_t) \geq \|E_t\| \tag{21}$$

$$\sigma_{\min}(U^{\star\top} Z_t) \geq \min\left\{ \left(1 + \frac{\eta}{8\kappa}\right)^t \sigma_{\min}(U^{\star\top} Z_0), \frac{1}{2\sqrt{\kappa}} \right\} \tag{22}$$

$$\|Z_t\| \leq 5 \tag{23}$$

It follows from equation (22) *that after $\Theta(\frac{\kappa}{\eta} \log(\frac{d}{\alpha}))$ steps, we have $\sigma_{\min}(Z_t) \geq \frac{1}{2\sqrt{\kappa}}$.*

Note that the theorem above only shows that the least singular value of Z_t goes above $1/(2\sqrt{\kappa})$. The following proposition completes the story by showing that once the signal is large enough, $U_t U_t^\top$ converges with a linear rate to the desired solution $X^\star$ (up to some small error.)

Proposition 4.3 *In the setting of Theorem 4.2, suppose $\|Z_t\| \leq 5$, $\sin(Z_t, U^\star) \leq 1/3$, and $\sigma_{\min}(Z_t) \geq \frac{1}{2\sqrt{\kappa}}$, then we have:*

$$\left\|U_{t+1} U_{t+1}^\top - X^\star\right\|_F^2 \leq \left(1 - \frac{\eta}{8\kappa}\right) \left\|U_t U_t^\top - X^\star\right\|_F^2 + O\left(\eta\sqrt{dr}\, \|E_t\|\right).$$

We defer the proof of Proposition 4.3 to Section D.5, which leverages the fact that function f satisfies the Polyak-Lojasiewicz condition Polyak (1963) when $U_t U_t^\top$ is well-conditioned. The proofs of Theorem 4.1 and 4.2 are presented in Section D.

5. Conclusions

The generalization performance of over-parameterized non-linear models, especially neural networks, has been a very intriguing research topic. This paper theoretically explains the regularization power of the optimization algorithms in learning matrix factorization models and one-hidden-layer neural nets with quadratic activations. In these cases, the gradient descent algorithm prioritizes to searching through the low complexity models.

It's an very interesting open question to establish similar results for deeper neural networks with other activations (e.g., ReLU) and loss functions (e.g., logistic loss). We remark that likely such a result will require not only a better understanding of statistics, but also a deep grasp of the behavior of the optimization algorithms for non-linear models, which in turns is another fascinating open question.

References

Radosław Adamczak, Alexander Litvak, Alain Pajor, and Nicole Tomczak-Jaegermann. Quantitative estimates of the convergence of the empirical covariance matrix in log-concave ensembles. *Journal of the American Mathematical Society*, 23(2):535–561, 2010.

Peter Bartlett, Dylan J Foster, and Matus Telgarsky. Spectrally-normalized margin bounds for neural networks. Technical report, Technical Report Preprint, 2017.

S. Bhojanapalli, B. Neyshabur, and N. Srebro. Global Optimality of Local Search for Low Rank Matrix Recovery. *ArXiv e-prints*, May 2016.

Olivier Bousquet and André Elisseeff. Stability and generalization. *Journal of Machine Learning Research*, 2(Mar):499–526, 2002.

Alon Brutzkus, Amir Globerson, Eran Malach, and Shai Shalev-Shwartz. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*, 2017.

Emmanuel J Candes. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9-10):589–592, 2008.

Emmanuel J Candès and Benjamin Recht. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717–772, 2009.

Emmanuel J Candès, Xiaodong Li, Yi Ma, and John Wright. Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11, 2011.

Yuxin Chen, Yuejie Chi, and Andrea J Goldsmith. Exact and stable covariance estimation from quadratic sampling via convex programming. *IEEE Transactions on Information Theory*, 61(7): 4034–4059, 2015.

Moustapha Cisse, Piotr Bojanowski, Edouard Grave, Yann Dauphin, and Nicolas Usunier. Parseval networks: Improving robustness to adversarial examples. In *International Conference on Machine Learning*, pages 854–863, 2017.

Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio. Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017.

Gintare Karolina Dziugaite and Daniel M Roy. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008*, 2017.

R. Ge and T. Ma. On the Optimization Landscape of Tensor Decompositions. *ArXiv e-prints*, June 2017.

Rong Ge, Jason D. Lee, and Tengyu Ma. Matrix completion has no spurious local minimum. *Advances in Neural Information Processing Systems (NIPS)*, 2016. URL http://arxiv.org/abs/1605.07272.

Rong Ge, Chi Jin, and Yi Zheng. No spurious local minima in nonconvex low rank problems: A unified geometric analysis. *arXiv preprint arXiv:1704.00708*, 2017.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. Implicit regularization in matrix factorization. *arXiv preprint arXiv:1705.09280*, 2017.

Moritz Hardt and Tengyu Ma. Identity matters in deep learning. In *5th International Conference on Learning Representations (ICLR 2017)*, 2017.

Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240*, 2015.

Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *CoRR*, abs/1609.05191, 2016. URL http://arxiv.org/abs/1609.05191.

Cijo Jose, Moustpaha Cisse, and Francois Fleuret. Kronecker recurrent units. 2017.

Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.

Anders Krogh and John A Hertz. A simple weight decay can improve generalization. In *Advances in neural information processing systems*, pages 950–957, 1992.

Richard Kueng, Holger Rauhut, and Ulrich Terstiege. Low rank matrix recovery from rank one measurements. *Applied and Computational Harmonic Analysis*, 42(1):88–116, 2017.

Roi Livni, Shai Shalev-Shwartz, and Ohad Shamir. On the computational efficiency of training neural networks. In *Advances in Neural Information Processing Systems*, pages 855–863, 2014.

Sayan Mukherjee, Partha Niyogi, Tomaso Poggio, and Ryan Rifkin. Learning theory: stability is sufficient for generalization and necessary and sufficient for consistency of empirical risk minimization. *Advances in Computational Mathematics*, 25(1):161–193, 2006.

Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nathan Srebro. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564*, 2017a.

Behnam Neyshabur, Srinadh Bhojanapalli, and Nati Srebro. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems*, pages 5943–5952, 2017b.

Boris Teodorovich Polyak. Gradient methods for minimizing functionals. *Zhurnal Vychislitel'noi Matematiki i Matematicheskoi Fiziki*, 3(4):643–653, 1963.

Benjamin Recht. A simpler approach to matrix completion. *The Journal of Machine Learning Research*, 12:3413–3430, 2011.

Benjamin Recht, Maryam Fazel, and Pablo A Parrilo. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Learnability, stability and uniform convergence. *Journal of Machine Learning Research*, 11(Oct):2635–2670, 2010.

M. Soltanolkotabi, A. Javanmard, and J. D. Lee. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *ArXiv e-prints*, July 2017.

Daniel Soudry and Yair Carmon. No bad local minima: Data independent training error guarantees for multilayer neural networks. *arXiv preprint arXiv:1605.08361*, 2016.

Daniel Soudry, Elad Hoffer, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

Nathan Srebro and Tommi Jaakkola. Weighted low-rank approximations. In *ICML*, 2013.

Nathan Srebro and Adi Shraibman. Rank, trace-norm and max-norm. In *International Conference on Computational Learning Theory*, pages 545–560. Springer, 2005.

Nati Srebro, Karthik Sridharan, and Ambuj Tewari. On the universality of online mirror descent. In *Advances in neural information processing systems*, pages 2645–2653, 2011.

Ju Sun, Qing Qu, and John Wright. A geometric analysis of phase retrieval. *Forthcoming*, 2016.

Stephen Tu, Ross Boczar, Mahdi Soltanolkotabi, and Benjamin Recht. Low-rank solutions of linear matrix equations via Procrustes flow. *arXiv preprint arXiv:1507.03566*, 2015.

Ashia C Wilson, Rebecca Roelofs, Mitchell Stern, Nathan Srebro, and Benjamin Recht. The marginal value of adaptive gradient methods in machine learning. *arXiv preprint arXiv:1705.08292*, 2017.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.

Yuchen Zhang, Jason Lee, Martin Wainwright, and Michael Jordan. On the learnability of fully-connected neural networks. In *Artificial Intelligence and Statistics*, pages 83–91, 2017.

Qinqing Zheng and John Lafferty. Convergence analysis for rectangular matrix completion using burer-monteiro factorization and gradient descent. *arXiv preprint arXiv:1605.07051*, 2016.

Kai Zhong, Prateek Jain, and Inderjit S Dhillon. Efficient matrix sensing using rank-1 gaussian measurements. In *International Conference on Algorithmic Learning Theory*, pages 3–18. Springer, 2015.

Appendix A. Neural networks with Quadratic Activations

In this section, we state the algorithms and generalization bounds for learning over-parameterized neural nets with quadratic activations, and give the key lemma for the analysis.

Let $(x_1, y_1), \ldots, (x_m, y_m)$ be n examples where x_i's are from distribution $\mathcal{N}(0, \mathrm{Id}_{d \times d})$ and y_i's are generated according to equation (3). Let $\hat{y} = \mathbf{1}^\top q(U^\top x)$ be our prediction. We use mean squared loss as the empirical loss. For technical reasons, we will optimize a truncated version of the empirical risk as

$$\tilde{f}(U) = \frac{1}{m} \sum_{i=1}^{n} (\hat{y}_i - y_i)^2 \mathbf{1}_{\|U^\top x\|^2 \leq R}$$

for some parameter R that will be logarithmic in dimension later. We design a variant of gradient descent as stated in Algorithm 1. We remark mostly driven by the analysis, our algorithm has an explicit re-scaling step. It resembles the technique of weight decay Krogh and Hertz (1992), which has similar effect to that of an ℓ_2 regularization. In the noiseless setting, the issue with vanilla weight decay or ℓ_2 regularizer is that the recovery guarantees will depend on the strength of the regularizer and thus cannot achieve zero. An alternative is to use a truncated ℓ_2 regularizer that only penalizes when U_t has norm bigger than a threshold. Our scaling that we are using is dynamically decided, and in contrast to this truncated regularizer, it scales down the iterate when the norm of U_t is small, and it scales up the iterate when U_t has norm bigger than the norm of $U^\star$.[10] Analyzing standard gradient descent is left for future work.

We note that one caveat here is that for technical reason, we assume that we know the Frobenius norm of the true parameter $U^\star$. It can be estimated by taking the average of the prediction $\frac{1}{m} \sum_{i=1}^{m} y_i$ since $\mathbb{E}[y] = \|U^\star\|_F^2$, and the algorithm is likely to be robust to the estimation error of $\|U^\star\|_F^2$. However, for simplicity, we leave such a robustness analysis for future work.

Algorithm 1 Algorithm for neural networks with quadratic activations

Inputs: n examples $(x_1, y_1), \ldots, (x_m, y_m)$ where x_i's are from distribution $\mathcal{N}(0, \mathrm{Id}_{d \times d})$ and y_i's are generated according to equation (3). Let $\tau = \|U^\star\|_F^2$.

Initialize U_0 as in equation (2)

For $t = 1$ to T:

$$\tilde{U}_t = U_t - \eta \nabla \tilde{f}(U_t)$$
$$U_{t+1} = \frac{1}{1 - \eta(\|U_t\|_F^2 - \tau)} \tilde{U}_t$$

As alluded before, one-hidden-layer neural nets with quadratic activation closely connects to matrix sensing because we can treat write the neural network prediction by:

$$\mathbf{1}^\top q(U^{\star \top} x) = \langle xx^\top, U^\star U^{\star \top} \rangle$$

Therefore, the i-th example x_i corresponds to the i-th measurement matrix in the matrix sensing via $A_i = x_i x_i^\top$. Assume $\{x_1 x_1^\top, \ldots, x_m x_m^\top\}$ satisfies RIP, then we can re-use all the proofs for

---

10. But note that such scaling up is unlikely to occur because the iterate stays low-rank

matrix sensing. However, this set of rank-1 measurement matrices *doesn't* satisfy RIP with high probability. The key observation is that if we truncated the observations properly, then we can make the truncated set of these rank-1 measurements satisfy RIP property again. Mathematically, we prove the following Lemma.

Lemma A.1 *Let* $(A_1, \ldots, A_m) = \{x_1 x_1^\top, \ldots, x_m x_m^\top\}$ *where* x_i*'s are i.i.d. from* $\sim \mathcal{N}(0, \mathrm{Id})$. *Let* $R = \log\left(\frac{1}{\delta}\right)$. *Then, for every* $q, \delta \in [0, 0.01]$ *and* $m \gtrsim d \log^4 \frac{d}{q\delta}/\delta^2$, *we have: with probability at least* $1 - q$, *for every symmetric matrix* X,

$$\left\| \frac{1}{m} \sum_{i=1}^m \langle A_i, X \rangle A_i 1_{|\langle A_i, X \rangle| \leq R} - 2X - \mathrm{tr}(X) \, \mathrm{Id} \right\| \leq \delta \|X\|_\star$$

Suppose X has rank at most r matrices and spectral norm at most 1, we have,

$$\left\| \frac{1}{m} \sum_{i=1}^m \langle A_i, X \rangle A_i 1_{|\langle A_i, X \rangle| \leq R} - 2X - \mathrm{tr}(X) \, \mathrm{Id} \right\| \leq r\delta \tag{24}$$

We can see that equation (24) implies Lemma 2.3 with a simple change of parameters (by setting δ to be a factor of r smaller). The proof uses standard technique from supreme of random process, and is deferred to Section E.

Theorem 1.2 follows straightforwardly from replacing the RIP property by the Lemma above. We provide a proof sketch below.

Proof [Proof Sketch of Theorem 1.2] The basic ideas is to re-use the proof of Theorem 4.1 at every iteration. First of all, we will replace all the RIP properties [11] in (38), (40), (46), (42), (59), (60), (72) and (76) by Lemma A.1. The only difference is that we will let the δ when applying Lemma A.1 to be $1/r$ smaller than the δ in Theorem 4.1.

We note that in Lemma A.1, there is an additional scaling of the identity term compared to Lemma 2.3. This is the reason why we have to change our update rule. We note that the update rule in Algorithm 1 undo the effect of this identity term and is identical to the update rule for matrix sensing problem: Let $c_t = \mathrm{tr}(U_t U_t^\top) - \mathrm{tr}(X^\star)$. Denote by $M_t' = M_t - c_t \, \mathrm{Id}$. The update in Algorithm 1 can be re-written as:

$$\tilde{U}_{t+1} = (\mathrm{Id} - \eta M_t' - \eta c_t \, \mathrm{Id}) U_t \tag{25}$$

$$= (1 - \eta c_t) \left(\mathrm{Id} - \frac{\eta}{1 - \eta c_t} M_t' \right) U_t$$

Hence we still have $U_{t+1} = \frac{1}{1-\eta c_t} \tilde{U}_{t+1} = \left(\mathrm{Id} - \frac{\eta}{1-\eta c_t} M_t' \right) U_t$. Thus the update rule here corresponds to the update rule for the matrix sensing case, and the rest of the proof follows from the proof of Theorem 4.1.

∎

---

11. We only require RIP on symmetric matrices.

Appendix B. Simulations

In this section, we present simulations to complement our theoretical results. In the first experiment, we show that the generalization performance of gradient descent depends on the choice of initialization. In particular, smaller initializations enjoy better generalization performance than larger initializations. In the second experiment, we demonstrate that gradient descent can run for a large number of iterations and the test error keeps decreasing, which suggests early stopping is not necessary. In the third experiment, we show that a natural projected gradient descent procedure works poorly compared to gradient descent on the factorized model, which suggests the power of using a factorized model. In the last experiment, we report results for running stochastic gradient descent on the quadratic neural network setting, starting from a large initialization.

We generate the true matrix by sampling each entry of $U^\star$ independently from a standard Gaussian distribution and let $X^\star = U^\star U^{\star\top}$. Each column of $U^\star$ is normalized to have unit norm, so that the spectral norm of $X^\star$ is close to one. For every sensing matrix A_i, for $i = 1, \ldots, m$, we sample the entries of A_i independently from a standard Gaussian distribution. Then we observe $b_i = \langle A_i, X^\star \rangle$. When an algorithm returns a solution $\hat{X}$, we measure training error by:

$$\sqrt{\frac{\sum_{i=1}^m (\langle A_i, \hat{X} \rangle - b_i)^2}{\sum_{i=1}^m b_i^2}}.$$

We measure test error by:

$$\frac{\|\hat{X} - X^\star\|_F}{\|X^\star\|_F}.$$

For the same $X^\star$, we repeat the experiment three times, by resampling the set of sensing matrices $\{A_i\}_{i=1}^m$. We report the mean and the error bar.

Choice of initialization. Let $U_0 = \alpha \operatorname{Id}$. We use $m = 5dr$ samples, where rank $r = 5$. We plot the training and test error for different values of α. Figure 1 shows that the gap between the training and test error narrows down as α decreases. We use step size 0.0025 and run gradient descent for 10^4 iterations.
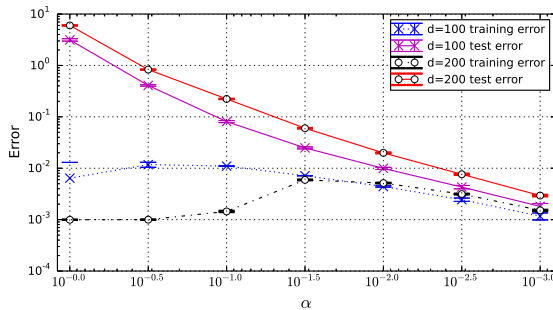

Figure 1: Generalization performance depends on the choice of initialization: the gap between training and test error decreases as α decreases. Here the number of samples is $5dr$, where rank $r = 5$. We initialize with $\alpha \operatorname{Id}$, and run 10^4 iterations with step size 0.0025.

In Figure 2, we run for longer iterations to further compare the generalization performance of initialization $U_0 = \alpha \operatorname{Id}$ for $\alpha = 1.0, 10^{-3}$. We report the mean values at each iteration over three runs. When $\alpha = 1.0$, despite the training error decreases below 10^{-4}, the test error remains to be on the order of 10^{-1}.
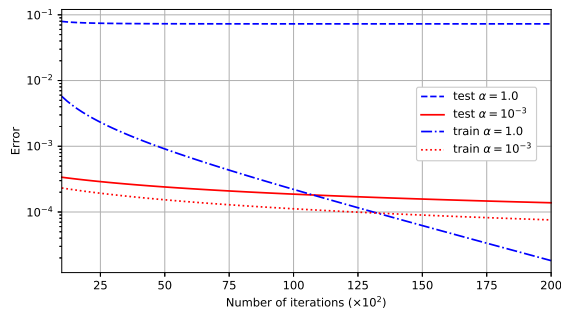

Figure 2: Further comparison between the generalization performance of large versus small initializations. We plot the data points from iteration 500 onwards to simplify the scale of the y-axis. The step size is 0.0025.

Accuracy. In this experiment, we fix the initialization to be $U_0 = 0.01 \operatorname{Id}$, and apply the same set of parameters as the first experiment. We keep gradient descent running for 10^5 iterations, to see if test error keeps decreasing or diverges at some point. Figure 3 confirms that test error goes down gradually.
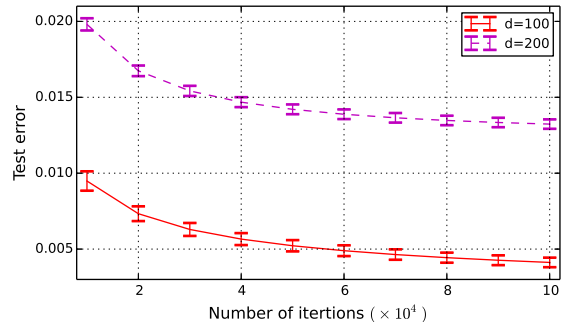

Figure 3: Test error keeps decreasing as the number of iterations goes to 10^5. Here the number of samples is $m = 5dr$, where rank $r = 5$. Note that the initial test error is approximately 1.

Projected gradient descent. In this experiment, we consider the following natural projected gradient descent (PDG) procedure. Let $f(X) = \frac{1}{m} \sum_{i=1}^{m} \left(\langle A_i, X \rangle - b_i \right)^2$. At every iteration, we first take a gradient step over $f(X)$, then project back to the PSD cone. We consider the sample complexity of PGD by varying the number of sensing matrices m from $5d$ to $35d$. Here the rank of $X^\star$ is 1. We found that the performance of PGD is much worse compared to gradient descent on

the factorized model. For both procedures, we use step size equal to 0.0025. We stop when the training error is less than 0.001, or when the number of iterations reaches 10^4. Figure 4 shows that gradient descent on the factorized model consistently recovers $X^\star$ accurately. On the other hand, the performance of projected gradient descent gets even worse as d increases from 100 to 150.
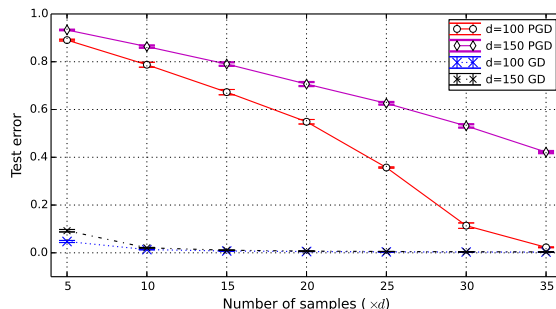

Figure 4: Projected gradient descent (PGD) requires more samples to recover $X^\star$ accurately, than gradient descent on the factorized model. Moreover, the performance of PGD gets worse as d increases.

Stochastic gradient descent. In this experiment, we complement our theoretical results by running stochastic gradient descent from large initializations. We generate $m = 5dr$ random samples and compute their true labels as the training dataset. For stochastic gradient descent, at every iteration we pick a training data point uniformly at random from the training dataset. We run one gradient descent step using the training data point. We initialize with $U_0 = \mathrm{Id}$ and use step size 8×10^{-5}. Figure 5(a) shows that despite the training error decreases to 10^{-7}, the test error remains large. We also report the results of running gradient descent on the same instance for comparison. As we have already seen, Figure 5(b) shows that gradient descent also gets stuck at a point with large test error, despite the training error being less than 10^{-7}.
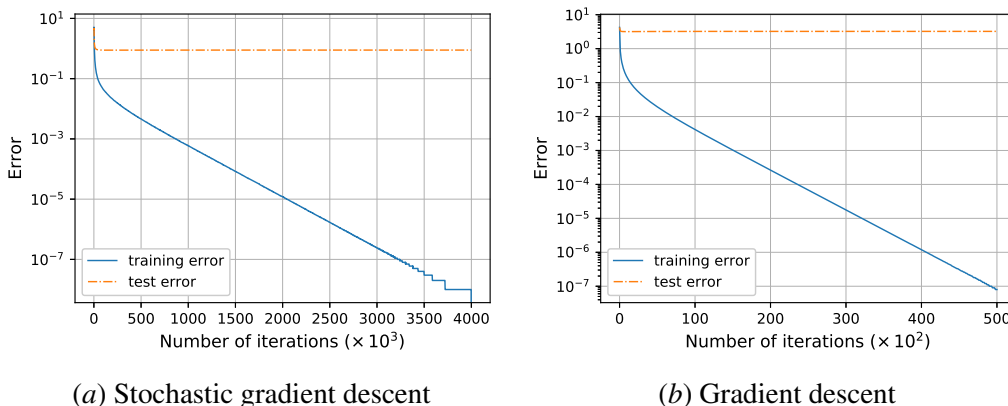

(a) Stochastic gradient descent

(b) Gradient descent

Figure 5: Stochastic gradient descent, when initialized with the identity matrix, does not generalize to test data. Here $d = 100$ and $r = 5$.

Appendix C. Proofs For Rank-1 Case

In this section, we present the missing proofs from Section 3.

Proof [Proof Sketch of Proposition 3.2] Using the update rule (10) for E_t, we have that

$$\|E_{t+1}\|_F^2 = \|E_t\|_F^2 - 2\eta \cdot \langle E_t, (\mathrm{Id} - \mathrm{Id}_{u^\star})M_t U_t\rangle + \eta^2 \|(\mathrm{Id} - \mathrm{Id}_{u^\star})M_t U_t\|_F^2 \qquad (26)$$

When η is sufficiently small and $\|M_t\|_F, \|U_t\|$ are bounded from above, the third term on the RHS is negligible compared to the second term. Therefore, we focus on the second term that is linear in η.

$$\langle E_t, (\mathrm{Id} - \mathrm{Id}_{u^\star})M_t U_t\rangle$$
$$= \frac{1}{m}\sum_{i=1}^m \langle A_i, U_t U_t^\top - X^\star\rangle\langle A_i, (\mathrm{Id} - \mathrm{Id}_{u^\star})E_t U_t^\top\rangle \qquad (27)$$

where in the last line we rearrange the terms and use the fact that $(\mathrm{Id} - \mathrm{Id}_{u^\star})$ is symmetric. Now we use Lemma 2.2 to show that equation (27) is close to $\langle U_t U_t^\top - X^\star, (\mathrm{Id} - \mathrm{Id}_{u^\star})E_t U_t^\top\rangle$, which is its expectation w.r.t the randomness of A_i's if A_i's were chosen from spherical Gaussian distribution. If $U_t U_t^\top - X^\star$ was a rank-1 matrix, then this would follow from Lemma 2.2 directly. However, $U_t U_t^\top$ is approximately low-rank. Thus, we decompose it into a low-rank part and an error part with small trace norm as in equation (12). Since $U_t U_t^\top - X^\star - E_t E_t^\top$ has rank at most 4, we can apply Lemma 2.2 to control the effect of A_i's,

$$\frac{1}{m}\sum_{i=1}^m \langle A_i, (U_t U_t^\top - X^\star - E_t E_t^\top)\rangle\langle A_i, E_t U_t^\top\rangle$$
$$\geq \langle U_t U_t^\top - X^\star - E_t E_t^\top, E_t U_t^\top\rangle - \delta\|U_t U_t^\top - X^\star - E_t E_t^\top\|_F\|E_t U_t^\top\|$$
$$\geq \langle U_t U_t^\top - X^\star - E_t E_t^\top, E_t U_t^\top\rangle - 1.5\delta\|E_t U_t^\top\| \qquad (28)$$

where the last inequality uses that $\|U_t U_t^\top - X^\star - E_t E_t^\top\|_F^2 = (1 - \|r_t\|^2)^2 + 2\|E_t r_t\|^2 \leq 1 + \|E_t\|^2\|r_t\|^2 \leq 11/8$.

For the $E_t E_t^\top$ term in the decomposition (12), by Lemma F.1, we have that

$$\frac{1}{m}\sum_{i=1}^m \langle A_i, E_t E_t^\top\rangle\langle A_i, E_t U_t^\top\rangle \geq \langle E_t E_t^\top, E_t U_t^\top\rangle - \delta\|E_t E_t^\top\|_\star\|E_t U_t^\top\|$$
$$\geq \langle E_t E_t^\top, E_t U_t^\top\rangle - 0.5\delta\|E_t U_t^\top\| \qquad (29)$$

Combining equation (27), (28) and (29), we conclude that

$$\langle E_t, (\mathrm{Id} - \mathrm{Id}_{u^\star})M_t U_t\rangle \geq \langle U_t U_t^\top - X^\star, E_t U_t^\top\rangle - 2\delta\|E_t U_t^\top\| \qquad (30)$$

Note that $u^{\star\top}E_t = 0$, which implies that $X^\star E_t = 0$ and $U_t^\top E_t = E_t^\top E_t$. Therefore,

$$\langle U_t U_t^\top - X^\star, E_t U_t^\top\rangle = \langle U_t U_t^\top, E_t U_t^\top\rangle$$
$$= \langle U_t^\top, U_T^\top E_t U_t^\top\rangle = \langle U_t^\top, E_T^\top E_t U_t^\top\rangle$$
$$= \langle E_t U_t^\top, E_t U_t^\top\rangle = \|E_t U_t^\top\|_F^2 \geq \|E_t U_t^\top\|^2$$

We can also control the third term in RHS of equation (26) by $\eta^2\|(\mathrm{Id} - \mathrm{Id}_{u^\star})M_tU_t\|_F^2 \leq 9\eta^2$. Since the bound here is less important (because one can always choose small enough η to make this term dominated by the first order term), we left the details to the reader. Combining the equation above with (30) and (26), we conclude the proof of equation (11). Towards bounding the spectral norm of E_{t+1}, we use a similar technique to control the difference between $(\mathrm{Id} - \mathrm{Id}_{u^\star})M_tU_t$ and $(\mathrm{Id} - \mathrm{Id}_{u^\star})(U_tU_t^\top - X^\star)U_t$ in spectral norm. By decomposing $U_tU_t^\top - X^\star$ as in equation (12) and applying Lemma 2.3 and Lemma F.2 respectively, we obtain that

$$\|(\mathrm{Id} - \mathrm{Id}_{u^\star})M_tU_t - (\mathrm{Id} - \mathrm{Id}_{u^\star})(U_tU_t^\top - X^\star)U_t\|$$
$$\leq 4\delta\left(\|U_tU_t^\top - X^\star - E_tE_t^\top\|_F + \|E_tE_t^\top\|_\star\right)\|U_t^\top\|$$
$$\leq 8\delta\|U_t\| \leq 8\delta(\|r_t\| + \|E_t\|) \qquad \text{(by the assumptions that } \|E_t\|_F \leq 1/2, \|r_t\| \leq 3/2\text{)}$$

Observing that $(\mathrm{Id} - \mathrm{Id}_{u^\star})(U_tU_t^\top - X^\star)U_t = E_tU_t^\top U_t$. Plugging the equation above into equation (10), we conclude that

$$\|E_{t+1}\| \leq \|E_t(\mathrm{Id} - \eta U_t^\top U_t)\| + 2\eta\delta(\|r_t\| + \|E_t\|)$$
$$\leq \|E_t\| + 2\eta\delta(\|r_t\| + \|E_t\|)$$

∎

Proof [Proof of Proposition 3.3] By the update rule (5), we have that

$$r_{t+1} = U_{t+1}^\top u^\star = U_t^\top(\mathrm{Id} - \eta M_t^\top)u^\star$$
$$= r_t - \eta U_t^\top M_t^\top u^\star.$$

By decomposing $M_t = \frac{1}{m}\sum_{i=1}^m \langle A_i, U_tU_t^\top - X^\star\rangle A_i$ as in equation (12), and then Lemma 2.3 and F.2, we obtain that

$$\left\|r_{t+1} - (r_t - \eta U_t^\top(U_tU_t^\top - X^\star)u^\star)\right\| \leq \delta\left(\left\|U_tU_t^\top - X^\star - E_tE_t^\top\right\|_F + \|E_tE_t^\top\|_\star\right)\|U_t\|.$$

Observe that $U_t^\top(U_tU_t^\top - X^\star)u^\star = U_t^\top U_t r_t - r_t = (r_tr_t^\top + E_t^\top E_t)r_t - r_t = (\|r_t\|^2 - 1)r_t - E_t^\top E_t r_t$. Also note that $\|U_tU_t^\top - X^\star - E_tE_t^\top\|_F^2 \leq 11/8$ and $\|E_t\|_F^2 \leq 1/4$, we have that

$$\|r_{t+1} - (1 + \eta(1 - \|r_t\|^2))r_t\| \leq \eta\|E_t^\top E_t r_t\| + 2\eta\delta\|U_t\|$$

Since $\|U_t\| \leq \|r_t\| + \|E_t\|$, we obtain the conclusion. ∎

Proof [Proof Sketch of Proposition 3.4] We will analyze the dynamics of r_t and E_t in two stages. The first stage consists of all the steps such that $\|r_t\| \leq 1/2$, and the second stage consists of the rest of steps. We will show that

a) Stage 1 has at most $O(\log(\frac{1}{\alpha})/\eta)$ steps. Throughout Stage 1, we have

$$\|E_t\| \leq 9\delta \tag{31}$$
$$\|r_{t+1}\| \geq (1 + \eta/3)\|r_t\| \tag{32}$$

b) In Stage 2, we have that

$$\|E_t\|_F^2 \lesssim \delta^2 \log \frac{1}{\delta}$$
$$\|r_t\| \leq 1 + O(\delta).$$

And after at most $O(\log(\frac{1}{\delta})/\eta)$ steps in Stage 2, we have $\|r_t\| \geq 1 - O(\delta)$.

We use induction with Proposition 3.2 and 3.3. For $t = 0$, we have that $\|E_0\| = \|r_0\| = \alpha$ because $U_0 = \alpha B$, where B is an orthonormal matrix. Suppose equation (31) is true for some t, then we can prove equation (32) holds by Proposition 3.3:

$$\|r_{t+1}\| \geq (1 + \eta(1 - \|r_t\|^2 - 2\delta - O(\delta^2)))\|r_t\|$$
$$\geq (1 + \eta/3)\|r_t\| \qquad \text{(by } \delta \leq 0.01 \text{ and } \|r_t\| \leq 1/2)$$

Suppose equation (32) holds, we can prove equation (31) holds by Proposition 3.2. We first observe that $t \leq O(\log \frac{1}{\alpha}/\eta)$, since r_t grows by a rate of $1 + \frac{\eta}{3}$. Denote by $\lambda = 1 + 2\eta\delta$. We have

$$\|E_{t+1}\| \leq \lambda\|E_t\| + 2\eta\delta\|r_t\| \Rightarrow \frac{\|E_{t+1}\|}{\lambda^{t+1}} \leq \frac{\|E_t\|}{\lambda^t} + 2\eta\delta \times \left(\frac{\|r_t\|}{\lambda^{t+1}}\right)$$

$$\Rightarrow \|E_{t+1}\| \leq \lambda^{t+1} \times \alpha + 2\eta\delta \times \sum_{i=0}^{t} \frac{\lambda^i\|r_t\|}{(1 + \frac{\eta}{3})^i} \leq 9\delta, \qquad \text{(by } \|r_{t+1}\| \geq \|r_t\|(1 + \frac{\eta}{3}).)$$

where the last inequality uses that

$$\lambda^{t+1}\alpha \leq \alpha \times \exp(2\eta\delta \times O(\log(\frac{1}{\alpha}))/\eta) = \alpha^{1-O(\delta)} \leq o(\delta), \text{ and}$$

$$\sum_{i=1}^{t} \frac{\lambda^i\|r_t\|}{(1 + \frac{\eta}{3})^i} \leq \frac{1 + \frac{\eta}{3}}{2(\frac{\eta}{3} - 2\eta\delta)}$$

For claim b), we first apply the bound obtained in claim a) on $\|E_t\|$ and the fact that $\|U_t\| \leq 3/2$ (this follows trivially from our induction, so we omit the details). We have already proved that $\|E_t\| \leq 9\delta$ in the first stage. For the second stage, as long as the number of iterations is less than $O(\log(\frac{1}{\delta})/\eta)$, we can still obtain via Proposition 3.2 that:

$$\|E_t\| \leq 9\delta + 4\eta\delta \times O(\log(\frac{1}{\delta}/\eta)) = O(\delta \log(\frac{1}{\delta})) \qquad \text{(since } \|E_t\| \leq 1/2 \text{ and } \|r_t\| \leq 3/2)$$

To summarize, we bound $\|E_t\|_F$ as follows:

$$\|E_t\|_F^2 \leq \|E_0\|_F^2 + 9\eta^2 t + O(\eta t \delta^2 \log(\frac{1}{\delta})) \lesssim \delta^2 \log(\frac{1}{\delta}),$$

because $\|E_0\|_F^2 = \alpha^2 d \leq \delta^2 \log\frac{1}{\delta}$, $\eta t \leq O(1)$, and $\eta^2 t \leq O(\eta) \leq O(\delta^2 \log(\frac{1}{\delta}))$. For the bound on $\|r_t\|$, we note that since $\|E_t\| \leq O(\delta \log(\frac{1}{\delta}))$, we can simplify Proposition 3.3 by

$$\left\|r_{t+1} - (1 + \eta(1 - \|r_t\|^2))r_t\right\| \leq 2\eta\delta\|r_t\| + O(\eta\delta^2 \log^2(\frac{1}{\delta})).$$

The proof that $\|r_t\| = 1 \pm O(\delta)$ after at most $O(\log(\frac{1}{\delta})/\eta)$ steps follows similarly by induction. The details are left to the readers. ∎

Appendix D. Proof of Theorem 4.1

We first lay out the proof outline proof outline of Theorem 4.2. We decompose it into the following propositions, from which Theorem 4.2 follows by induction.

The following proposition gives the base case for the induction, which straightforwardly follows from the definition $U_0 = \alpha B$ where B is an orthonormal matrix.

Proposition D.1 (Base Case) *In the setting of Theorem 4.2, we have*

$$\|E_0\| = \alpha = \sigma_{\min}(Z_0) \le 1/d, \quad U^{\star\top} E_0 = 0, \quad and \quad \sin(Z_0, U^\star) = 0.$$

The following Proposition bounds the growth rate of the spectral norm of the error E_t.

Proposition D.2 *In the setting of Theorem 4.2, suppose $\|E_t\| \le 1/d$. Then,*

$$\|E_{t+1}\| \le \left(1 + \eta O\left(\delta\sqrt{r} + \sin(Z_t, U^\star)\right)\right) \times \|E_t\|.$$

When $\|U_t U_t^\top - X^\star\|_F$ is small, the growth of E_t becomes slower:

$$\|E_{t+1}\| \le \left(1 + \eta O\left(\|U_t U_t^\top - X^\star\|_F + \sqrt{r}\,\|E_t\|\right)\right) \times \|E_t\|.$$

The following Proposition shows that effectively we can almost pretend that Z_{t+1} is obtained from applying one gradient descent step to Z_t, up to some some error terms.

Proposition D.3 *In the setting of Theorem 4.2, suppose for some t we have $\|E_t\| \le 1/d$ and $\|Z_t\| \le 5$. Then,*

$$\left\| Z_{t+1} - \left(Z_t - \eta\nabla f(Z_t) - \eta E_t Z_t^\top Z_t - 2\eta\,\mathrm{Id}_{S_t} M_t E_t\right)\right\| \le \eta\tau_t, \tag{33}$$

where $\tau_t \lesssim \delta\sqrt{r}\,\|E_t\|$.

The following proposition shows that the angle between the span of Z_t to the span of $U^\star$ is growing at mostly linearly in the number of steps.

Proposition D.4 *In the setting of Theorem 4.2, assuming equation (33) holds for some t with $\tau_t \le \rho\sigma_r(U_t)$, $\|Z_t\| \le 5$ and $\sigma_{\min}(Z_t) \ge \|E_t\|/2$. Then, as long as $\sin(Z_t, U^\star) \le \sqrt{\rho}$ we have:*

$$\sin(Z_{t+1}, U^\star) \le \sin(Z_t, U^\star) + O(\eta\rho + \eta\,\|E_t\|) \quad and \quad \|Z_{t+1}\| \le 5. \tag{34}$$

Then we show that the projection of the signal term Z_t to the subspace of $U^\star$ increases at an exponential rate (until it goes above $1/(2\sqrt{k})$). Note that $U^\star$ is sufficiently close to the span of Z_t and therefore it implies that the least singular value of Z_t also grows.

Proposition D.5 *In the setting of Theorem 4.2, suppose equation (34) holds for some t, and $\|Z_t\| \le 5$ and $\sigma_{\min}(Z_t) \ge \|E_t\|/2$, we have that:*

$$\sigma_{\min}(U^{\star\top} Z_{t+1}) \ge \min\{(1 + \frac{\eta}{8\kappa})\sigma_{\min}(U^{\star\top} Z_t), \frac{1}{2\sqrt{\kappa}}\}.$$

The proofs of the above propositions are deferred to Section D. Now we are ready to prove Theorem 4.2.

Proof [Proof of Theorem 4.2] When $t = 0$, the base case follows from Proposition D.1. Assume that equations (19), (20), (21), (22), and (23) are true before or at step t, we prove the conclusion for step $t + 1$. By Proposition D.3 we have that equation 33 are true with $\tau_t \lesssim \delta\sqrt{r}$. We have that $\sigma_r(U_t) \geq \sigma_r(Z_t)$, because the column subspace of Z_t and E_t are orthogonal. By induction hypothesis, $\sigma_r(Z_t) \geq \|E_t\|/2$, and $\sigma_r(U_t) \geq \|E_t\|/2$. Set the ρ in Proposition D.4 to $O(\frac{\delta\sqrt{r}}{\kappa\log(\frac{d}{\alpha})})$. When $t \leq T_1 = O(\frac{\kappa}{\rho}\log(\frac{d}{\alpha}))$, we have that $\sin(Z_t, U^\star) \leq \eta\rho t \leq \sqrt{\rho}$. Hence the assumptions of Proposition D.4 are verified. Therefore,

$$\sin(Z_{t+1}, U^\star) \lesssim \eta\rho t + O(\eta\rho) + O(\eta\|E_t\|) \lesssim \eta\rho(t+1)$$

because $\|E_t\| \leq 1/d \ll \rho$, and $\|Z_t\| \leq 5$. Next,

$$\begin{aligned}
\|E_{t+1}\| &\leq \left(1 + \eta O\left(\delta\sqrt{r} + \sin(Z_t, U^\star)\right)\right) \cdot \|E_t\| \\
&\leq (1 + O(\eta^2\rho T_1))^{T_1}\|E_0\| &&\text{(by equations (19), (20), and } \delta\sqrt{r} \leq O(\eta\rho T_1)) \\
&\leq 4\|E_0\| \leq 1/d &&\text{(since } \eta^2\rho T_1^2 \leq O(\delta\sqrt{r}\kappa\log(\frac{d}{\alpha})) \leq O(1))
\end{aligned}$$

Therefore we can apply Proposition D.5 to obtain that $\sigma_{\min}(U^{\star\top}Z_t)$ grows by a rate of at least $1 + \frac{\eta}{8\kappa}$. On the other hand by Proposition D.2, $\|E_t\|$ grows by a rate of at most $1 + \eta O(\delta\sqrt{r} + \sqrt{\rho}) \leq 1 + \frac{\eta}{8\kappa}$. Hence we obtain Equation 21. ■

Finally we prove Theorem 4.1.

Proof First of all, for $t \leq T_1 = \Theta(\kappa\log\frac{d}{\alpha}/\eta)$, using Theorem 4.2, we know that the requirements in Proposition 4.3 is satisfied. Applying Proposition 4.3, we prove the theorem for $t \leq T_1$.

Then, we inductively show that the error is bounded by $O(\alpha\sqrt{d}/\kappa^2)$ from the T_1-th iteration until the T_0-th iteration. Suppose at iteration t, we have $\|U_tU_t^\top - X^\star\|_F^2 \lesssim \alpha\sqrt{d}/\kappa^2$. Thus, using Proposition D.2, we know that $\|E_t\|$ grows by a rate of at most $1 + \eta O(\sqrt{\alpha}d^{1/4}/\kappa)$ for this t. This implies that for every $t \in [T_1, T_0]$, we have

$$\begin{aligned}
\|E_{t+1}\| &\leq \left(1 + \eta O(\sqrt{\alpha}d^{1/4}/\kappa)\right)\|E_t\| \\
&\leq \|E_{T_1}\|\left(1 + \eta O(\sqrt{\alpha}d^{1/4}/\kappa)\right)^{T_0} \leq 4\|E_{T_1}\| \leq 16\|E_0\|.
\end{aligned}$$

By inductive hypothesis we recall $\|U_tU_t^\top - X^\star\|_F^2 \lesssim \kappa\sqrt{d}/\kappa^2$, which implies by elementary calculation that $\sigma_{\min}(Z_t) \geq \frac{1}{2\sqrt{\kappa}}, \|Z_t\| \leq 5$ and $\sin(Z_t, U^\star) \leq \frac{1}{3}$ (by using $U_t = Z_t + E_t$ and $\|E_t\| \leq 16\|E_0\|$). Thus, the requirements in Proposition 4.3 hold, and applying it we obtain that $\left\|U_{t+1}U_{t+1}^\top - X^\star\right\|_F^2 \lesssim \alpha\sqrt{d}/\kappa^2$. This completes the induction. ■

D.1. Proof of Proposition D.2

We start off with a straightforward triangle inequality for bounding E_{t+1} given E_t.

Lemma D.6 *By update rule (equation ([5])) and the definition of E_t (equation ([16])), we have that*

$$E_{t+1} = (\mathrm{Id} - \mathrm{Id}_{S_{t+1}})(\mathrm{Id} - \eta M_t) E_t .$$

It follows that

$$\|E_{t+1}\| \le \|(\mathrm{Id} - \eta M_t) E_t\| \le \|E_t\| + \eta \|M_t E_t\| . \tag{35}$$

Therefore, next we will bound the norm of $M_t E_t$. The key idea is to use the restricted isometry property to control the effect of M_t (see the meta claim equation ([7]) in Section [2] for more intuitions.)

Lemma D.7 *In the setting of Proposition [D.2], we have that*

$$\|M_t E_t\| \le \|E_t\| \left(\delta \|Z_t Z_t^\top - X^\star\|_F + 2\delta \|Z_t E_t^\top\|_F + \delta \|E_t E_t^\top\|_\star + \|Z_t E_t^\top\| + \|X^\star(\mathrm{Id} - \mathrm{Id}_{S_t})\| \right) \tag{36}$$

As a direct consequence, using the assumption $\|E_t\| \le 1/d$,

$$\|E_{t+1}\| \le \|E_t\| \left(1 + O\left(\eta\delta\sqrt{r} + \eta \|X^\star(\mathrm{Id} - \mathrm{Id}_{S_t})\| \right) \right) \tag{37}$$

Proof We first note that $E_t = (\mathrm{Id} - \mathrm{Id}_{S_t}) E_t$ by the update rule and definition [16]. It follows that $M_t E_t = M_t(\mathrm{Id} - \mathrm{Id}_{S_t}) E_t$. Next, since $U_t U_t^\top - X^\star - E_t E_t^\top$ has rank at most $4r$, by Lemma [2.3] we have that

$$\left\| \frac{1}{m} \sum_{i=1}^m \langle A_i, U_t U_t^\top - X^\star - E_t E_t^\top \rangle A_i (\mathrm{Id} - \mathrm{Id}_{S_t}) E_t - (U_t U_t^\top - X^\star - E_t E_t^\top)(\mathrm{Id} - \mathrm{Id}_{S_t}) E_t \right\|$$
$$\le \delta \times \|U_t U_t^\top - X^\star - E_t E_t^\top\|_F \|(\mathrm{Id} - \mathrm{Id}_{S_t}) E_t\| \tag{38}$$
$$= \delta \times \left(\|Z_t Z_t^\top - X^\star\|_F + 2\|Z_t E_t^\top\|_F \right) \|E_t\|$$

Note that $(U_t U_t^\top - X^\star - E_t E_t^\top)(\mathrm{Id} - \mathrm{Id}_{S_t}) E_t = Z_t E_t^\top E_t - X^\star(\mathrm{Id} - \mathrm{Id}_{S_t}) E_t$. It follows that

$$\left\| \frac{1}{m} \sum_{i=1}^m \langle A_i, U_t U_t^\top - X^\star - E_t E_t^\top \rangle A_i (\mathrm{Id} - \mathrm{Id}_{S_t}) E_t \right\|$$
$$\le \|X^\star(\mathrm{Id} - \mathrm{Id}_{S_t}) E_t\| + \left\| Z_t E_t^\top E_t \right\| + \delta \left(\|Z_t Z_t^\top - X^\star\|_F + 2\|Z_t E_t^\top\|_F \right) \|E_t\|$$
$$\le \|E_t\| \left(\delta \|Z_t Z_t^\top - X^\star\|_F + 2\delta \|Z_t E_t^\top\|_F + \|Z_t E_t^\top\| + \|X^\star(\mathrm{Id} - \mathrm{Id}_{S_t})\| \right) \tag{39}$$

By Lemma [F.2], we have that

$$\left\| \frac{1}{m} \sum_{i=1}^m \langle A_i, E_t E_t^\top \rangle A_i (\mathrm{Id} - \mathrm{Id}_{S_t}) E_t - E_t E_t^\top (\mathrm{Id} - \mathrm{Id}_{S_t}) E_t \right\| \le \delta \|E_t E_t^\top\|_\star \|E_t\| \tag{40}$$

Combining equation ([40]) and ([39]) we complete the proof of equation ([36]). To prove equation ([37]), we will use equation ([35]) and that $\|Z_t Z_t^\top - X^\star\|_F \lesssim \sqrt{r}$ (Corollary [D.16]), $\|Z_t\|_F^2 \le \sqrt{r} \times \|Z_t Z_t^\top\|_F \lesssim r$. ∎

When $\|U_t U_t^\top - X^\star\|$ is small, the growth of $\|E_t\|$ becomes slower.

Lemma D.8 *In the setting of Proposition D.2, we have that*

$$\|M_t\| \leq \|U_t U_t^\top - X^\star\| + \delta \|U_t U_t^\top - X^\star\|_F + 3\sqrt{r}\|E_t\|.$$

As a direct consequence, $\|M_t\| \leq O(1)$. And

$$\|M_t E_t\| \lesssim \|E_t\| \left(\|U_t U_t^\top - X^\star\|_F + \sqrt{r}\,\|E_t\| \right) \tag{41}$$

Proof By the definition of M_t, we have that

$$
\begin{aligned}
\|M_t\| &= \left\| \frac{1}{m} \sum_{i=1}^m \langle A_i, U_t U_t^\top - X^\star \rangle A_i \right\| \\
&= \left\| \frac{1}{m} \sum_{i=1}^m \langle A_i, U_t U_t^\top - X^\star - E_t E_t^\top \rangle A_i + \frac{1}{m} \sum_{i=1}^m \langle A_i, E_t E_t^\top \rangle A_i \right\| \\
&\leq \|U_t U_t^\top - X^\star\| + \delta \times \left(\|U_t U_t^\top - X^\star - E_t E_t^\top\|_F + \|E_t E_t^\top\|_* \right) \tag{42} \\
&\leq (1+\delta)\|U_t U_t^\top - X^\star\|_F + 3\sqrt{r}\|E_t\|
\end{aligned}
$$

where the second to last line is because of Lemma 2.3 and Lemma F.2. For the last line, we use

$$\|U_t U_t^\top - X^\star - E_t E_t^\top\|_F = \|Z_t Z_t^\top - X^\star\|_F + 2\|Z_t E_t^\top\|_F \lesssim \|Z_t Z_t^\top - X^\star\|_F + \sqrt{r}\|E_t\|,$$

because of Corollary D.16. And $\|Z_t Z_t^\top - X^\star\|_F \leq \|U_t U_t^\top - X^\star\|_F + O(\sqrt{r}\,\|E_t\|)$, because

$$
\begin{aligned}
\left\| U_t U_t^\top - X^\star \right\|_F &= \left\| (Z_t + E_t)(Z_t + E_t)^\top - X^\star \right\|_F \\
&\geq \left\| Z_t Z_t^\top - X^\star \right\|_F - 2\,\|Z_t\|_F\,\|E_t\| - \|E_t\|\,\|E_t\|_F
\end{aligned}
$$

■

Finally we complete the proof of Proposition D.2.

Proof [Proof of Proposition D.2] Using the fact that $X^\star$ has spectral norm less than 1, wee can bound the term $\|X^\star(\mathrm{Id} - \mathrm{Id}_{S_t})\|$ in equation (37) by

$$\|X^\star(\mathrm{Id} - \mathrm{Id}_{S_t})\| \leq \left\| U^{\star\top}(\mathrm{Id} - \mathrm{Id}_{S_t}) \right\| = \|(\mathrm{Id} - \mathrm{Id}_{S_t})U^\star\| = \sin(S_t, U^\star).$$

Since S_t is the column span of Z_t, using the equation above and equation (37) we conclude the proof. ■

D.2. Proof of Proposition D.3

We first present a simpler helper lemma that will be used in the proof. One can view X in the following lemma as a perturbation. The lemma bounds the effect of the perturbation to the left hand side of equation (43).

Lemma D.9 *Let $S \in \mathbb{R}^{d \times r}$ be a column orthonormal matrix and $S^\perp = \mathrm{Id} - SS^\top$ be its orthogonal complement. Let $X \in \mathbb{R}^{d \times d}$ be any matrix where $\|X\| < \frac{1}{3}$. We have:*

$$\| \mathrm{Id}_{(\mathrm{Id} - X)S}(\mathrm{Id} - X)S^\perp + 2\,\mathrm{Id}_S\, XS^\perp \| \lesssim \|X\|^2. \tag{43}$$

Proof By definition, we know that:

$$\mathrm{Id}_{(\mathrm{Id} - X)S} = (\mathrm{Id} - X)S \left(S^\top (\mathrm{Id} - X)^\top (\mathrm{Id} - X)S \right)^{-1} S^\top (\mathrm{Id} - X)^\top$$

$$= (\mathrm{Id} - X)S \left(\mathrm{Id} - S^\top (X^\top + X)S + S^\top X^\top XS \right)^{-1} S^\top (\mathrm{Id} - X)^\top \tag{44}$$

Denote by $Y = S^\top (X^\top + X)S - S^\top X^\top XS$. We have that $\|Y\| \leq 2\|X\| + \|X\|^2 < 7\|X\|/3$. By expanding $(\mathrm{Id} - Y)^{-1}$, we obtain:

$$\left\| (\mathrm{Id} - Y)^{-1} - (\mathrm{Id} + Y) \right\| \leq \sum_{i=2}^{\infty} \|Y\|^i$$

$$\leq \frac{\|Y\|^2}{1 - \|Y\|} \leq 25\|X\|^2.$$

Hence we get:

$$\left\| (\mathrm{Id} - Y)^{-1} - (\mathrm{Id} + S^\top (X^\top + X)S) \right\| \leq 26\|X\|^2.$$

Denote by

$$A = (\mathrm{Id} - X)S \left(\mathrm{Id} + S^\top (X^\top + X)S \right) S^\top (\mathrm{Id} - X^\top)$$

$$= (\mathrm{Id} - X) \left(\mathrm{Id}_S + \mathrm{Id}_S (X^\top + X)\, \mathrm{Id}_S \right) (\mathrm{Id} - X^\top).$$

We separate the terms in A which has degree 1 in X:

$$A_1 = \mathrm{Id}_S - X\,\mathrm{Id}_S - \mathrm{Id}_S\, X^\top + \mathrm{Id}_S (X^\top + X)\,\mathrm{Id}_S$$

Consider the spectral norm of $A(\mathrm{Id} - X)S^\perp$. We have that $A_1 S^\perp = -\mathrm{Id}_S\, XS^\perp$. For $-A_1 XS^\perp$, the only term which has degree 1 in X is $-\mathrm{Id}_S\, XS^\perp$. To summarize, we obtain by triangle inequality that:

$$\| \mathrm{Id}_{(\mathrm{Id} - X)S}(\mathrm{Id} - X)S^\top + 2\,\mathrm{Id}_S\, XS^\perp \| \leq \|A + 2\,\mathrm{Id}_S\, XS^\perp\| + 26\|X\|^2$$

$$\lesssim \|A_1 + 2\,\mathrm{Id}_S\, XS^\perp\| + \|X\|^2 \lesssim \|X\|^2$$

$\blacksquare$

Now we are ready present the proof of Proposition D.3.

Proof [Proof of Proposition D.3] We first consider the term $M_t Z_t$:

$$M_t Z_t = \frac{1}{m} \sum_{i=1}^{m} \langle A_i, U_t U_t^\top - X^\star \rangle A_i Z_t$$

$$= \frac{1}{m} \sum_{i=1}^{m} \left\langle A_i, Z_t Z_t^\top - X^\star \right\rangle A_i Z_t$$

$$+ \frac{1}{m} \sum_{i=1}^{m} \left\langle A_i, E_t Z_t^\top + Z_t E_t^\top \right\rangle A_i Z_t + \frac{1}{m} \sum_{i=1}^{m} \left\langle A_i, E_t E_t^\top \right\rangle A_i Z_t \tag{45}$$

Note that $\nabla f(Z_t) = \frac{1}{m} \sum_{i=1}^{m} \langle A_i, Z_t Z_t^\top - X^\star \rangle A_i Z_t$. Using Lemma 2.3 and F.2 on the two terms in line (45) with the fact that $Z_t^\top E_t = 0$, we have that

$$\left\| M_t Z_t - E_t Z_t^\top Z_t - \nabla f(Z_t) \right\| \le 2\sqrt{r}\delta \|E_t\| \|Z_t\|^2 + \delta d \|E_t\|^2 \|Z_t\| \tag{46}$$

We decompose Z_{t+1} by:

$$\begin{aligned}
Z_{t+1} = \mathrm{Id}_{S_{t+1}} U_{t+1} &= U_{t+1} - (\mathrm{Id} - \mathrm{Id}_{S_{t+1}}) U_{t+1} \\
&= (\mathrm{Id} - \eta M_t) Z_t + (\mathrm{Id} - \eta M_t) E_t - (\mathrm{Id} - \mathrm{Id}_{S_{t+1}})(\mathrm{Id} - \eta M_t) E_t \\
&= Z_t - \eta M_t Z_t + \mathrm{Id}_{S_{t+1}}(\mathrm{Id} - \eta M_t) E_t,
\end{aligned} \tag{47}$$

where in the second equation we use the fact that $(\mathrm{Id} - \mathrm{Id}_{S_{t+1}})(\mathrm{Id} - \eta M_t) Z_t = 0$, since the column subspace of Z_t is S_t. Setting $S = \mathrm{Id}_{S_t}$ and $X = \eta M_t$ in Lemma D.9, we have that $\mathrm{Id}_{S_{t+1}} = \mathrm{Id}_{(\mathrm{Id} - X)S}$. Applying Lemma D.9, we conclude:

$$\| \mathrm{Id}_{S_{t+1}}(\mathrm{Id} - \eta M_t) S^\perp + 2 \mathrm{Id}_S(\mathrm{Id} - \eta M_t) S^\perp \| \le \eta^2 \|M_t\|^2,$$

which implies that:

$$\| \mathrm{Id}_{S_{t+1}}(\mathrm{Id} - \eta M_t) E_t + 2 \mathrm{Id}_S(\mathrm{Id} - \eta M_t) E_t \| \le \eta^2 \|E_t\| \|M_t\|^2.$$

Since $\|M_t\| \le O(1)$ by Lemma D.8, the conclusion follows by combining Equation (46) and (47) with the equation above. ∎

D.3. Proof of Proposition D.4

Towards proving Proposition D.4, we further decompose Z_t into

$$Z_t = (U^\star + F_t) R_t \tag{48}$$

where $R_t \in \mathbb{R}^{r \times d}, F_t \in \mathbb{R}^{d \times r}$ are defined as

$$R_t = U^{\star \top} Z_t, \text{ and } F_t = (\mathrm{Id} - \mathrm{Id}_{U^\star}) Z_t R_t^+ \tag{49}$$

Recall that X^+ denotes the pseudo-inverse of X. We first relate the the spectral norm of F_t with our target in Proposition D.4, the principal angle between Z_t and $U^\star$. Up to third order term, we show that $\|F_t\|$ is effective equal to the principle angle, and this is pretty much the motivation to decompose Z_t in equation (48).

Lemma D.10 *Let F_t be defined as in equation (49). Then, if $\|F_t\| < 1/3$, we have that*

$$\|F_t\| - \|F_t\|^3 \le \sin(Z_t, U^\star) \le \|F_t\|$$

Proof Let $S_t = (U^\star + F_t)(\mathrm{Id} + F_t^\top F_t)^{-1/2}$. By the fact that $U^{\star \top} F_t = 0$, we have $S_t^\top S_t = \mathrm{Id}$ and S_t has the same column span as Z_t. Therefore, the columns of S_t form an orthonormal basis of Z_t, and we have that

$$\sin(Z_t, U^\star) = \|(\mathrm{Id} - \mathrm{Id}_{U^\star}) S_t\| = \left\| F_t(\mathrm{Id} + F_t^\top F_t)^{-1/2} \right\|$$

Suppose F_t has singular value $\sigma_j, j = 1, \ldots, r$, then it's straightforward to show that $F_t(\mathrm{Id} + F_t^\top F_t)^{-1/2}$ has singular values $\frac{\sigma_j}{\sqrt{1+\sigma_j^2}}, j = 1, \ldots, r$. The conclusion then follows basic calculus and the fact that $\max \sigma_j \leq 1/3$. ∎

Therefore, it suffices to bound the spectral norm of F_t. However, the update rules of F_t or Z_t are difficult to reason about. Therefore, we introduce the following intermediate term $\tilde{Z}_t$ that bridges Z_{t+1}, F_{t+1} with Z_t and $F + t$. We define $\tilde{Z}_t$ as:

$$\tilde{Z}_t = (\mathrm{Id} - \eta E_t Z_t^\top) Z_t (\mathrm{Id} - 2\eta Z_t^+ \mathrm{Id}_{S_t} M_t E_t) \tag{50}$$

The motivation of defining such Z_t is that Z_{t+1} depends on Z_t via relatively simple formula as the lemma below shows:

Proposition D.11 *In the setting of Proposition D.4,*

$$\left\| Z_{t+1} - \left(\tilde{Z}_t - \eta \nabla f(Z_t) \right) \right\| = O(\eta \tau_t)$$

The proof of Proposition D.11 is deferred to the later part of this section. We also decompose $\tilde{Z}_t$ same as Z_t to $\tilde{Z}_t = (U^\star + \tilde{F}_t)\tilde{R}_t$, where

$$\tilde{R}_t = U^{\star\top} \tilde{Z}_t, \text{ and } \tilde{F}_t = (\mathrm{Id} - \mathrm{Id}_{U^\star})\tilde{Z}_t \tilde{R}_t^+ \tag{51}$$

We will prove that $\tilde{R}_t$ is close to R_t and $\tilde{F}_t$ is close to F_t in the following sense:

Proposition D.12 *In the setting of Proposition D.4,*

$$\sigma_{\min}(\tilde{R}_t) \geq \sigma_{\min}(R_t) \left(1 - \frac{\eta}{100\kappa} \right)$$

Proposition D.13 *In the setting of Proposition D.4,*

$$\|\tilde{F}_t - F_t\| \lesssim \eta \|E_t\| + \eta \rho$$

We will prove these propositions in the following sections. After that, we will focus on the update from $\tilde{Z}_t$ to Z_{t+1}. In particular, we will bound R_{t+1} and F_{t+1} directly using $\tilde{R}_t$ and $\tilde{F}_t$.

D.3.1. PROOFS OF PROPOSITION D.11, D.12, AND D.13

Proof [Proof of Proposition D.11] By definition of $\tilde{Z}_t$,

$$\tilde{Z}_t = (\mathrm{Id} - \eta E_t Z_t^\top) Z_t (\mathrm{Id} - 2\eta Z_t^+ \mathrm{Id}_{S_t} M_t E_t) \tag{52}$$

$$= Z_t - \eta \left(E_t Z_t^\top Z_t + 2\mathrm{Id}_{S_t} M_t E_t \right) + 2\eta^2 E_t Z_t^\top \mathrm{Id}_{S_t} M_t E_t \tag{53}$$

Recall that by Proposition D.3, the update rule of Z_{t+1} satisfies

$$\left\| Z_{t+1} - \left(Z_t - \eta \nabla f(Z_t) - \eta E_t Z_t^\top Z_t - 2\eta \mathrm{Id}_{S_t} M_t E_t \right) \right\| \leq \eta \tau_t$$

Putting the above two formulas together (using the bound of $\|M_t E_t\|$ as in Lemma D.7) we have that

$$\left\|Z_{t+1} - \left(\tilde{Z}_t - \eta\nabla f(Z_t)\right)\right\| = O(\eta^2 \|E_t\| + \eta\tau_t) \tag{54}$$

On the other hand, a direct calculation also shows that $\|\tilde{Z}_t - Z_t\| = O(\eta\|E_t\|)$. Moreover, $\tilde{Z}_t$ is a rank r matrix and $\|Z_t\|, \|\tilde{Z}_t\| = O(1)$. Therefore, let us denote by $\tilde{Z}_t = Z_t + \Delta_t$, we have:

$$
\begin{aligned}
\nabla f(\tilde{Z}_t) &= \frac{1}{m}\sum_{i=1}^m \langle A_i, \tilde{Z}_t\tilde{Z}_t^\top - X^*\rangle A_i\tilde{Z}_t \\
&= \frac{1}{m}\sum_{i=1}^m \langle A_i, Z_tZ_t^\top - X^*\rangle A_i\tilde{Z}_t + \frac{1}{m}\sum_{i=1}^m \langle A_i, Z_t\Delta_t^\top + \Delta_tZ_t^\top\rangle A_i\tilde{Z}_t + \frac{1}{m}\sum_{i=1}^m \langle A_i, \Delta_t\Delta_t^\top\rangle A_i\tilde{Z}_t \\
&= \frac{1}{m}\sum_{i=1}^m \langle A_i, Z_tZ_t^\top - X^*\rangle A_iZ_t + \frac{1}{m}\sum_{i=1}^m \langle A_i, Z_tZ_t^\top - X^*\rangle A_i\Delta_t \\
&\quad + \frac{1}{m}\sum_{i=1}^m \langle A_i, Z_t\Delta_t^\top + \Delta_tZ_t^\top\rangle A_i\tilde{Z}_t + \frac{1}{m}\sum_{i=1}^m \langle A_i, \Delta_t\Delta_t^\top\rangle A_i\tilde{Z}_t
\end{aligned}
$$

Note that Δ_t is at most rank $2r$, therefore, we can apply the RIP property of $\{A_i\}_{i=1}^m$ (Lemma 2.3) and $\|\tilde{Z}_t\|, \|Z_t\| = O(1)$ to conclude that

$$\frac{1}{m}\left\|\sum_{i=1}^m \langle A_i, Z_tZ_t^\top - X^*\rangle A_i\Delta_t + \sum_{i=1}^m \langle A_i, Z_t\Delta_t^\top + \Delta_tZ_t^\top\rangle A_i\tilde{Z}_t + \sum_{i=1}^m \langle A_i, \Delta_t\Delta_t^\top\rangle A_i\tilde{Z}_t\right\| = O(\|\Delta_t\|)$$

Recall that $\|\Delta_t\| = O(\eta\|E_t\|)$, the equation above and the formula for $\nabla f(\tilde{Z})$ implies that

$$\left\|Z_{t+1} - \left(\tilde{Z}_t - \eta\nabla f(\tilde{Z}_t)\right)\right\| = O(\eta^2 \|E_t\| + \eta\tau_t) = O(\eta\tau_t)$$

∎

The above proposition implies that Z_{t+1} is very close to doing one step of gradient descent from $\tilde{Z}_t$. Thus, we will mainly focus on $\tilde{Z}_t$ in the later section.

Proof [Proof of Proposition D.12] By definition of $\tilde{R}_t$, we know that

$$
\begin{aligned}
\tilde{R}_t &= U^{\star\top}\tilde{Z}_t = U^{\star\top}(\mathrm{Id} - \eta E_tZ_t^\top)Z_t(\mathrm{Id} - 2\eta Z_t^+ \mathrm{Id}_{S_t} M_tE_t) \\
&= U^{\star\top}Z_t(\mathrm{Id} - 2\eta Z_t^+ \mathrm{Id}_{S_t} M_tE_t) - \eta U^{\star\top}E_tZ_t^\top Z_t(\mathrm{Id} - 2\eta Z_t^+ \mathrm{Id}_{S_t} M_tE_t)
\end{aligned} \tag{55}
$$

By definition of R_t and the assumption that $\|F_t\| \leq 1/3$, we have that $\sigma_r(Z_t) \geq \frac{1}{2}\sigma_{\min}(R_t)$, which implies that $\|Z_t^+\| \leq 2\|R_t^+\|$. Thus, using the bound of $\|M_tE_t\|$ as in Lemma D.7,

$$\|Z_t^+ \mathrm{Id}_{S_t} M_tE_t\| \lesssim \frac{\|M_tE_t\|}{\sigma_{\min}(R_t)} \lesssim \frac{\|E_t\|(\delta\sqrt{r} + \sin(Z_t, U^\star))}{\sigma_{\min}(R_t)}$$

Similarly, we can get:

$$\|U^{\star\top}E_t\| = \left\|U^{\star\top}(\mathrm{Id} - \mathrm{Id}_{S_t})E_t\right\| \leq \|E_t\|\left\|U^{\star\top}(\mathrm{Id} - \mathrm{Id}_{S_t})\right\| = \|E_t\|\sin(Z_t, U^\star) \tag{56}$$

Therefore, bounding the terms in equation (55) using the bounds above, we have:

$$
\begin{aligned}
\sigma_{\min}(\tilde{R}_t) &\geq \sigma_{\min}(U^{\star\top}Z_t)(1 - \eta\left\|Z_t^+ \operatorname{Id}_{S_t} M_t E_t\right\|) - \eta\left\|U^{\star\top}E_t Z_t^\top Z_t(\operatorname{Id} - 2\eta Z_t^+ \operatorname{Id}_{S_t} M_t E_t)\right\| \\
&\geq \sigma_{\min}(R_t)(1 - \eta\left\|Z_t^+ \operatorname{Id}_{S_t} M_t E_t\right\|) - 2\eta\left\|U^{\star\top}E_t\right\| \\
&\geq \sigma_{\min}(R_t)\left(1 - \eta O\left(\frac{\|E_t\|\,(\delta\sqrt{r} + \sin(Z_t, U^\star))}{\sigma_{\min}(R_t)}\right)\right)
\end{aligned}
\tag{57}
$$

Again, using $\sigma_r(Z_t) \geq \frac{1}{2}\sigma_{\min}(R_t)$ we know that

$$
\frac{\|E_t\|\,(\delta\sqrt{r} + \sin(Z_t, U^\star))}{\sigma_{\min}(R_t)} \leq 2\frac{\|E_t\|\,(\delta\sqrt{r} + \sin(Z_t, U^\star))}{\sigma_{\min}(Z_t)}
$$

By assumption of Proposition D.4 and that $\sigma_{\min}(Z_t) \geq \frac{1}{2}\|E_t\|$, we obtain:

$$
\begin{aligned}
2\frac{\|E_t\|\,(\delta\sqrt{r} + \sin(Z_t, U^\star))}{\sigma_{\min}(Z_t)} &\leq 2(\delta\sqrt{r} + \sin(Z_t, U^\star)) \\
&\lesssim \sqrt{\rho} \qquad\qquad \text{(by Assumption in proposition D.4)}
\end{aligned}
$$

Thus, by equation (57) above, and our choice of parameter such that $\sqrt{\rho} \lesssim 1/\kappa$, we conclude that

$$
\sigma_{\min}(\tilde{R}_t) \geq \sigma_{\min}(R_t)\left(1 - \frac{\eta}{100\kappa}\right)
$$

∎

Next we prove prove Proposition D.13, which focus on $\tilde{F}_t$:

Proof [Proof of Proposition D.13]

We know that right multiply Z_t by an invertible matrix does not change the column subspace as Z_t, thus won't change the definition of F_t, so we can just focus on $(\operatorname{Id} - \eta E_t Z_t^\top)Z_t$. We know that

$$
\begin{aligned}
\tilde{F}_t &= (\operatorname{Id} - \operatorname{Id}_{U^\star})\left(\operatorname{Id} - \eta E_t Z_t^\top\right)Z_t(U^{\star\top}(\operatorname{Id} - \eta E_t Z_t^\top)Z_t)^+ \\
&= (\operatorname{Id} - \operatorname{Id}_{U^\star})Z_t\left(U^{\star\top}Z_t - \eta U^{\star\top}E_t Z_t^\top Z_t\right)^+ \\
&\quad - \eta(\operatorname{Id} - \operatorname{Id}_{U^\star})E_t Z_t^\top\left(U^{\star\top}Z_t - \eta U^{\star\top}E_t Z_t^\top Z_t\right)^+
\end{aligned}
$$

Since Z_t is rank r, we can do the SVD of Z_t as $Z_t = V\Sigma_Z W^\top$ for column orthonormal matrices $V, W \in \mathbb{R}^{d\times r}$ and diagonal matrix $\Sigma_Z \in \mathbb{R}^{r\times r}$. Then, we can write $\left(U^{\star\top}Z_t - \eta U^{\star\top}E_t Z_t^\top Z_t\right)^+$ as:

$$
\begin{aligned}
\left(U^{\star\top}Z_t - \eta U^{\star\top}E_t Z_t^\top Z_t\right)^+ &= \left(\left(U^{\star\top}V - \eta U^{\star\top}E_t Z_t^\top V\right)\Sigma_Z W^\top\right)^+ \\
&= W^\top \Sigma_Z^{-1}\left(U^{\star\top}V - \eta U^{\star\top}E_t Z_t^\top V\right)^{-1}
\end{aligned}
$$

By our assumption that $\sin(Z_t, U^\star) \leq \frac{1}{3}$ we know that $\sigma_{\min}(U^{\star\top}V) = \Omega(1)$. Thus, by $\|Z_t\| = O(1)$ and Woodbury matrix inversion formula we have:

$$\left\| \left(U^{\star\top}V - \eta U^{\star\top}E_t Z_t^\top V \right)^{-1} - (U^{\star\top}V)^{-1} \right\| \lesssim \eta \left\| U^{\star\top}E_t Z_t^\top V \right\| \lesssim \eta \left\| U^{\star\top}E_t \right\|$$

Recall that $F_t = (\mathrm{Id} - \mathrm{Id}_{U^\star})Z_t R_t^+ = (\mathrm{Id} - \mathrm{Id}_{U^\star})Z_t W^\top \Sigma_Z^{-1}(U^{\star\top}V)^{-1}$. Thus, using $\sigma_r(Z_t) \geq \frac{1}{2}\sigma_{\min}(R_t)$, an elementary calculation gives us:

$$\left\| \tilde{F}_t - F_t \right\| \lesssim \eta \left\| E_t \right\| + \eta \frac{\|F_t\| \left\| U^{\star\top}E_t \right\|}{\sigma_{\min}(R_t)}$$

By assumption $\|E_t\| \lesssim \sigma_{\min}(Z_t)$ in Proposition D.4, together with $\sigma_r(Z_t) \geq \frac{1}{2}\sigma_{\min}(R_t)$ and inequality 56, we have

$$\frac{\|F_t\| \left\| U^{\star\top}E_t \right\|}{\sigma_{\min}(R_t)} \lesssim \frac{\|F_t\| \|E_t\| \sin(Z_t, U^\star)}{\sigma_{\min}(Z_t)} \lesssim \|F_t\|^2$$

By our choice of parameter, we know that $\|F_t\|^2 \leq \rho$, therefore, we proved that $\left\| \tilde{F}_t - F_t \right\| \leq \eta(\|E_t\|_2 + \rho)$ as desired. ∎

Now, we can just focus on $\tilde{Z}_t$. One of the crucial fact about the gradient $\nabla f(\tilde{Z}_t)$ is that it can be decomposed into

$$\nabla f(\tilde{Z}_t) = N_t \tilde{R}_t$$

where N_t is a matrix defined as

$$N_t = \frac{1}{m}\sum_{i=1}^{m} \langle A_i, \tilde{Z}_t \tilde{Z}_t^\top - X^\star \rangle A_i(U^\star + \tilde{F}_t) \tag{58}$$

Therefore, $\tilde{Z}_t$ and $\nabla f(\tilde{Z}_t)$ share the row space and we can factorize the difference between $\tilde{Z}_t$ and $\eta \nabla f_t(\tilde{Z}_t)$ as

$$\tilde{Z}_t N_t - \eta \nabla f(\tilde{Z}_t) = (\tilde{F}_t - \eta N_t)\tilde{R}_t$$

Note that the definition of N_t depends on the random matrices $A_1, \ldots, A_t$. The following lemma show that for our purpose, we can essentially view N_t as its population version — the counterpart of N_t when we have infinitely number of examples. The proof uses the RIP properties of the matrices $A_1, \ldots, A_m$.

Lemma D.14 *In the setting of Proposition D.4, let N_t be defined as in equation* (58). *Then,*

$$\left\| N_t - (\tilde{Z}_t \tilde{Z}_t^\top - X^\star)(U^\star + \tilde{F}_t) \right\| \leq 2\delta \left\| \tilde{Z}_t \tilde{Z}_t^\top - X^\star \right\|_F$$

Proof Recalling the definition of N_t, by Lemma 2.3, we have that

$$\left\| N_t - (\tilde{Z}_t \tilde{Z}_t^\top - X^\star)(U^\star + \tilde{F}_t) \right\| \leq \delta \left\| \tilde{Z}_t \tilde{Z}_t^\top - X^\star \right\|_F \left\| U^\star + \tilde{F}_t \right\| \qquad (59)$$

$$\leq 2\delta \left\| \tilde{Z}_t \tilde{Z}_t^\top - X^\star \right\|_F \qquad \text{(by the assumption } \left\| \tilde{F}_t \right\| < 1/3 \text{)}$$

$\blacksquare$

Lemma D.15 *For any $t \geq 0$, suppose $\left\| Z_{t+1} - (\tilde{Z}_t - \eta \mathcal{G}(\tilde{Z}_t)) \right\| \leq \eta\tau$, we have*

$$\|Z_{t+1}\| \leq \|\tilde{Z}_t\| \left(1 - \frac{1}{2}\eta\|\tilde{Z}_t\|^2 \right) + 2\eta\|\tilde{Z}_t\|\|X^\star\| + \eta\tau .$$

Proof By Lemma 2.3, we have:

$$\left\| \mathcal{G}(\tilde{Z}_t) - (\tilde{Z}_t \tilde{Z}_t^\top - X^\star)\tilde{Z}_t \right\| \leq \delta \left\| \tilde{Z}_t \tilde{Z}_t^\top - X^\star \right\|_F \left\| \tilde{Z}_t \right\| . \qquad (60)$$

Therefore,

$$\|Z_{t+1}\| \leq \left\| \tilde{Z}_t - \eta\mathcal{G}(\tilde{Z}_t) \right\| + \eta\tau$$

$$\leq \left\| \tilde{Z}_t - \eta(\tilde{Z}_t \tilde{Z}_t^\top - X^\star)\tilde{Z}_t \right\| + \eta\delta \left\| \tilde{Z}_t \tilde{Z}_t^\top - X^\star \right\|_F \left\| \tilde{Z}_t \right\| + \eta\tau$$

$$\leq \left\| \left(\mathrm{Id} - \eta\tilde{Z}_t \tilde{Z}_t^\top + \eta X^\star \right) \tilde{Z}_t \right\| + \eta\delta\sqrt{r} \left\| \tilde{Z}_t \right\| \left(\left\| \tilde{Z}_t \right\|^2 + \|X^\star\| \right) + \eta\tau$$

$$\leq (1 - \eta \left\| \tilde{Z}_t \right\|) \left\| \tilde{Z}_t \right\| + \frac{1}{2}\eta \left\| \tilde{Z}_t \right\| \left(\left\| \tilde{Z}_t \right\|^2 + 4 \|X^\star\| \right) + \eta\tau \qquad \text{(by } \delta\sqrt{r} \leq 1/2 \text{)}$$

$\blacksquare$

As a direct corollary, we can inductive control the norm of $\tilde{Z}_t$.

Corollary D.16 *In the setting of Proposition D.4, we have that*

$$\|Z_{t+1}\| \leq 5, \left\| \tilde{R}_t \right\| \leq 6$$

Moreover,

$$\|Z_{t+1}Z_{t+1}^\top - X^\star\|_F \lesssim \sqrt{r}, \text{ and } \|N_t\| \lesssim \sqrt{r}$$

Proof Using the assumption that $\|X^\star\| = 1$ and the assumption that equation (33) holds, then we have that $\left\| \tilde{Z}_t \right\| \leq \|Z_t\| (1 + O(\eta \|E\|_t)) \leq 5(1 + O(\eta \|E\|_t))$. Applying Lemma D.15 with $\tau = O(\tau_t)$, we have that $\|Z_{t+1}\| \leq 5$. We also have that $\left\| \tilde{R}_t \right\| \leq \left\| \tilde{Z}_t \right\| \leq 6$. Moreover, we have $\|Z_{t+1}Z_{t+1}^\top - X^\star\|_F \leq \left\| \tilde{Z}_t \right\|_\star + \|X^\star\|_F \lesssim \sqrt{r}(\left\| \tilde{Z}_t \right\| + \|X^\star\|) \lesssim \sqrt{r}$. As a consequence, $\|N_t\| \leq (1 + \delta) \left\| \tilde{Z}_t \tilde{Z}_t^\top - X^\star \right\|_F \left(\|U^\star\| + \left\| \tilde{F}_t \right\| \right) \lesssim \sqrt{r}$. $\blacksquare$

We start off with a lemma that controls the changes of $\tilde{R}_t$ relatively to R_{t+1}.

Lemma D.17 *In the setting of Proposition D.4, then we have that $\tilde{R}_t R_{t+1}^{-1}$ can be written as*

$$\tilde{R}_t R_{t+1}^+ = \mathrm{Id} + \eta \tilde{R}_t \tilde{R}_t^\top - \eta \Sigma^\star + \xi_t^{(R)}$$

where $\|\xi_t^{(R)}\| \lesssim \eta \delta \sqrt{r} + \eta \rho + \eta \|\tilde{F}_t\| + \eta^2$. It follows that $\left\|\tilde{R}_t R_{t+1}^+\right\| \leq 4/3$ and $\tau \leq 2\rho\sigma_{\min}(R_{t+1})$.

Proof By the definition of $\tilde{R}_t$ and equation (33), we have that

$$\begin{aligned}
\eta\tau &\geq \|R_{t+1} - \tilde{R}_t - \eta U^{\star\top} \nabla f(\tilde{Z}_t)\| \\
&= \|R_{t+1} - (\mathrm{Id} - \eta U^{\star\top} N_t)\tilde{R}_t\|
\end{aligned} \tag{61}$$

Form this we can first obtain a very weak bound on $\sigma_{\min}(R_{t+1})$:

$$\begin{aligned}
\sigma_{\min}(R_{t+1}) &\geq \sigma_{\min}((\mathrm{Id} - \eta U^{\star\top} N_t)\tilde{R}_t) - \eta\tau \\
&\geq \frac{3}{4} \cdot \sigma_{\min}(\tilde{R}_t) - O(\eta\rho\sigma_{\min}(\tilde{R}_t)) \\
&\quad \text{(by } \tau \leq O(\rho\sigma_{\min}(\tilde{R}_t)) \text{ and } \|\eta U^{\star\top} N_t\| \leq \|\eta N_t\| \lesssim \eta\sqrt{r} \leq 1/4) \\
&\geq \frac{1}{2} \cdot \sigma_{\min}(\tilde{R}_t)
\end{aligned} \tag{62}$$

Re-using equation (61), we have

$$\begin{aligned}
\|\mathrm{Id} - (\mathrm{Id} - \eta U^{\star\top} N_t)\tilde{R}_t R_{t+1}^+\| &\leq \eta\tau \cdot \left\|R_{t+1}^+\right\| = \eta\tau/\sigma_{\min}(R_{t+1}) \\
&\leq 2\eta\tau/\sigma_{\min}(\tilde{R}_t) \leq O(\eta\rho)
\end{aligned} \tag{63}$$

where we used equation (62) and $\tau \leq O(\rho\sigma_{\min}(\tilde{R}_t))$. This also implies a weak bound for $\tilde{R}_t R_{t+1}^+$ that $\|\tilde{R}_t R_{t+1}^+\| \leq 2$. By Lemma D.14, we have that

$$\|\eta U^{\star\top} N_t - \eta U^{\star\top}(\tilde{Z}_t \tilde{Z}_t^\top - X^\star)(U^\star + \tilde{F}_t)\| \leq 2\delta\eta\|\tilde{Z}_t \tilde{Z}_t^\top - X^\star\|_F \lesssim \delta\eta\sqrt{r}$$

where we use the fact that $\|\tilde{Z}_t \tilde{Z}_t^\top - X^\star\|_F \lesssim \sqrt{r}$. Note that $X^\star = U^\star \Sigma^\star U^{\star\top}$ and $\tilde{Z}_t = (U^\star + \tilde{F}_t)\tilde{R}_t$, we have $\|\eta U^{\star\top} N_t - \eta(\tilde{R}_t \tilde{R}_t^\top - \Sigma^\star)(U^\star + \tilde{F}_t)^\top(U^\star + \tilde{F}_t)\| \lesssim \delta\eta\sqrt{r}$.

Bounding the higher-order term, we have that

$$\left\|\eta(\tilde{R}_t \tilde{R}_t^\top - \Sigma^\star)(U^\star + \tilde{F}_t)^\top(U^\star + \tilde{F}_t) - \eta(\tilde{R}_t \tilde{R}_t^\top - \Sigma^\star)\right\| \lesssim \eta\left\|\tilde{F}_t\right\| \qquad \text{(by } \left\|\tilde{R}_t\right\| \leq 6)$$

which implies that

$$\|\eta U^{\star\top} N_t - \eta(\tilde{R}_t \tilde{R}_t^\top - \Sigma^\star)\| \lesssim \delta\eta\sqrt{r} + \eta\|\tilde{F}_t\| \tag{64}$$

Combining the equation above with equation (63) and $\|\tilde{R}_t R_{t+1}^+\| \leq 2$, we have that

$$\|\mathrm{Id} - (\mathrm{Id} - \eta \tilde{R}_t \tilde{R}_t^\top + \eta\Sigma^\star)\tilde{R}_t R_{t+1}^+\| \lesssim \eta\delta\sqrt{r} + \eta\rho + \eta\left\|\tilde{F}_t\right\| \tag{65}$$

For $\eta \lesssim 1$, we know that

$$\left\|(\mathrm{Id} - \eta \tilde{R}_t \tilde{R}_t^\top + \eta\Sigma^\star)(\mathrm{Id} + \eta \tilde{R}_t \tilde{R}_t^\top - \eta\Sigma^\star) - \mathrm{Id}\right\| \lesssim \eta^2$$

This implies that

$$\|(\mathrm{Id} + \eta \tilde{R}_t \tilde{R}_t^\top - \eta \Sigma^\star) - \tilde{R}_t R_{t+1}^+\| \lesssim \eta \delta \sqrt{r} + \eta \rho + \eta \left\| \tilde{F}_t \right\| + \eta^2$$

which completes the proof. ∎

We express F_{t+1} as a function of $\tilde{F}_t$ and other variables.

Lemma D.18 *In the setting of Proposition D.4, let N_t be defined as in equation (58). Then,*

$$F_{t+1} = \tilde{F}_t (\mathrm{Id} - \eta \tilde{R}_t \tilde{R}_t^\top) \tilde{R}_t R_{t+1}^+ + \xi_t^{(F)}$$

where $\|\xi_t^{(F)}\| \lesssim \delta \eta \sqrt{r} + \eta \rho + \eta \left\| \tilde{F}_t \right\|^2$.

Proof By equation (33), we have that

$$\left\| (\mathrm{Id} - \mathrm{Id}_{U^\star}) \cdot \left(Z_{t+1} - (\tilde{Z}_t - \eta \nabla f(\tilde{Z}_t)) \right) \right\| \le \eta \tau_t$$

which, together with the decomposition (49), implies

$$\eta \tau \ge \left\| F_{t+1} R_{t+1} - \tilde{F}_t \tilde{R}_t + \eta (\mathrm{Id} - \mathrm{Id}_{U^\star}) \nabla f(\tilde{Z}_t) \right\|$$
$$= \left\| F_{t+1} R_{t+1} - (\tilde{F}_t - \eta (\mathrm{Id} - \mathrm{Id}_{U^\star}) N_t) \tilde{R}_t \right\|$$

Recall that $\tau_t \le 2\rho \sigma_{\min}(R_{t+1})$ (by Lemma D.17), we conclude

$$\left\| F_{t+1} - (\tilde{F}_t - \eta (\mathrm{Id} - \mathrm{Id}_{U^\star}) N_t) \tilde{R}_t R_{t+1}^+ \right\| \le 2\eta \rho \tag{66}$$

Note that $(\mathrm{Id} - \mathrm{Id}_{U^\star}) X^\star = 0$ and that $(\mathrm{Id} - \mathrm{Id}_{U^\star}) \tilde{Z}_t \tilde{Z}_t^\top = \tilde{F}_t \tilde{R}_t \tilde{R}_t^\top (U^\star + \tilde{F}_t)^\top$. We obtain that

$$\left\| (\mathrm{Id} - \mathrm{Id}_{U^\star}) N_t - \tilde{F}_t \tilde{R}_t \tilde{R}_t^\top (U^\star + \tilde{F}_t)^\top (U^\star + \tilde{F}_t) \right\| \le 2\delta \left\| \tilde{Z}_t \tilde{Z}_t^\top - X^\star \right\|_F \lesssim \delta \sqrt{r} \tag{67}$$

where we used the fact that $\left\| \tilde{Z}_t \tilde{Z}_t^\top - X^\star \right\|_F \lesssim \sqrt{r}$ (by Lemma D.16).

Bounding the higher-order term, we have that

$$\left\| \tilde{F}_t \tilde{R}_t \tilde{R}_t^\top (U^\star + \tilde{F}_t)^\top (U^\star + \tilde{F}_t) - \tilde{F}_t \tilde{R}_t \tilde{R}_t^\top \right\| \lesssim \left\| \tilde{F}_t \right\|^2 \quad \text{(by } \left\| \tilde{R}_t \right\| \le 6 \text{ from Corollary D.16)}$$

Combining equation (66), (67) and the equation above, and using the fact that $\left\| \tilde{R}_t R_{t+1}^+ \right\| \le 2$, we complete the proof. ∎

Combining Lemma D.18 and Lemma D.17, we can relate the $\|F_{t+1}\|$ with $\left\| \tilde{F}_t \right\|$:

Lemma D.19 *In the setting of Proposition D.4, we have that $\tilde{F}_t$ can be written as*

$$F_{t+1} = \tilde{F}_t (\mathrm{Id} - \eta \Sigma^\star) + \xi_t$$

where $\|\xi_t\| \le O(\eta \delta \sqrt{r} + \eta \left\| \tilde{F}_t \right\|^2 + \eta \rho + \eta^2)$. *As a consequence,*

$$\|F_{t+1}\| \le \left\| \tilde{F}_t \right\| + O(\eta \left\| \tilde{F}_t \right\|^2 + \eta \rho)$$

Proof Combine Lemma D.18 and Lemma D.17, we have that

$$F_{t+1} = \tilde{F}_t(\mathrm{Id} - \eta\tilde{R}_t\tilde{R}_t^\top)\tilde{R}_t R_{t+1}^+ + \xi_t^{(F)}$$
$$= \tilde{F}_t(\mathrm{Id} - \eta\tilde{R}_t\tilde{R}_t^\top)\left(\mathrm{Id} + \eta\tilde{R}_t\tilde{R}_t^\top - \eta\Sigma^\star + \xi_t^{(R)}\right) + \xi_t^{(F)}$$

Thus, with the bound on $\|\xi_t^{(R)}\|$ and $\|\xi_t^{(F)}\|$ from Lemma D.18 and Lemma D.17, we know that

$$\left\|F_{t+1} - (\mathrm{Id} - \eta\Sigma^\star)\tilde{F}_t\right\| \lesssim \eta\|\tilde{F}_t\| + \|\tilde{F}_t\|\|\xi_t^{(R)}\| + \|\xi_t^{(F)}\| \lesssim \eta\|\tilde{F}_t\|^2 + \eta\delta\sqrt{r} + \eta\rho + \eta^2$$
$$\lesssim \eta\left\|\tilde{F}_t\right\|^2 + \eta\rho\,.$$

■

The proof of Proposition D.4 follow straightforwardly from Lemma D.10 and Lemma D.19.
Proof [Proof of Proposition D.4] Using the assumption that $\|F_t\| \lesssim \sqrt{\rho}$ (Thus $\|F_t\|^2 \lesssim \rho$). Since we have showed that $\left\|\tilde{F}_t - F_t\right\| \lesssim \eta(\rho + \|E_t\|)$, the proof of this proposition followings immediately from Lemma D.19. ■

D.4. Proof of Proposition D.5

We first prove the following technical lemma that characterizes how much the least singular value of a matrix changes when it got multiplied by matrices that are close to identity.

Lemma D.20 *Suppose $Y_1 \in \mathbb{R}^{d\times r}$ and Σ is a PSD matrix in $\mathbb{R}^{r\times r}$. For some $\eta > 0$, let*

$$Y_2 = (\mathrm{Id} + \eta\Sigma)Y_1(\mathrm{Id} - \eta Y_1^\top Y_1)$$

Then, we have:

$$\sigma_{\min}(Y_2) \geq (1 + \eta\sigma_{\min}(\Sigma))\left(1 - \eta\sigma_{\min}(Y_1)^2\right)\sigma_{\min}(Y_1)$$

Proof First let's consider the matrix $Y := Y_1(\mathrm{Id} - \eta Y_1^\top Y_1)$. We have that that $\sigma_{\min}(Y) = \left(1 - \eta\sigma_{\min}(Y_1)^2\right)\sigma_{\min}(Y_1)$. Next, we bound the least singular value of $(\mathrm{Id} + \eta\Sigma)Y$:

$$\sigma_{\min}((\mathrm{Id} + \eta\Sigma)Y) \geq \sigma_{\min}(\mathrm{Id} + \eta\Sigma)\sigma_{\min}(Y) = (1 + \sigma_{\min}(\Sigma))\sigma_{\min}(Y)$$

where we used the facts that $\sigma_{\min}(AB) \geq \sigma_{\min}(A)\sigma_{\min}(B)$, and that for any symmetric PSD matrix B, $\sigma_{\min}(\mathrm{Id} + B) = 1 + \sigma_{\min}(B)$. ■

Now we are ready to prove Proposition D.5. Note that the least singular value of Z_t is closely related to the least singular value of $\tilde{R}_t$ because F_t is close to 0. Using the machinery in the proof of Lemma D.17, we can write R_{t+1} as some transformation of $\tilde{R}_t$, and then use the lemma above to bound the least singular value of R_{t+1} from below.
Proof [Proof of Proposition D.5] Recall that in equation (65) in the the proof of Lemma D.17, we showed that

$$\|\mathrm{Id} - (\mathrm{Id} - \eta\tilde{R}_t\tilde{R}_t^\top + \eta\Sigma^\star)\tilde{R}_t R_{t+1}^+\| \lesssim \eta\delta\sqrt{r} + \eta\rho + \eta\|F_t\|\,.$$

Together with $\left\|\tilde{R}_t\right\| \leq 6$ and $\left\|\tilde{R}_t R_{t+1}^+\right\| \leq 2$ by Corollary D.16 and Lemma D.17 respectively, we have:

$$\| \operatorname{Id} -(\operatorname{Id} +\eta\Sigma^\star)\tilde{R}_t(\operatorname{Id} -\eta\tilde{R}_t^\top \tilde{R}_t)R_{t+1}^+\| \lesssim \eta\delta\sqrt{r} + \eta\rho + \eta \left\|F_t\right\| + \eta^2 .$$

Denote by $\xi = \operatorname{Id} -(\operatorname{Id} +\eta\Sigma^\star)\tilde{R}_t(\operatorname{Id} -\eta\tilde{R}_t^\top \tilde{R}_t)R_{t+1}^+$, we can rewrite

$$R_{t+1} - (\operatorname{Id} +\eta\Sigma^\star)\tilde{R}_t(\operatorname{Id} -\eta\tilde{R}_t^\top \tilde{R}_t) = \xi R_{t+1} ,$$

Without loss of generality, let us assume that $\sigma_{\min}(\tilde{R}_t) \leq \frac{1}{1.9\sqrt{\kappa}}$. By Lemma D.20 that

$$\sigma_{\min}(R_{t+1}) \geq \frac{(1 + \eta\sigma_{\min}(\Sigma^\star)) \left(1 - \eta\sigma_{\min}(\tilde{R}_t)^2\right) \sigma_{\min}(\tilde{R}_t)}{1 + O\left(\eta\delta\sqrt{r} + \eta\rho + \eta \left\|F_t\right\| + \eta^2\right)} .$$

Since $\|F_t\| \lesssim \eta\rho t$ by Proposition D.4, we have:

$$\begin{aligned}
\sigma_{\min}(R_{t+1}) &\geq (1 + \eta/\kappa) \left(1 - \eta\sigma_{\min}(\tilde{R}_t)^2\right) \sigma_{\min}(\tilde{R}_t) \left(1 - O\left(\eta\delta\sqrt{r} + \eta\rho + \eta^2\rho t\right)\right) \\
&\geq \left(1 + \eta\left(\frac{1}{3\kappa} - O\left(\delta\sqrt{r} + \rho + \eta(\delta\rho)t\right)\right)\right) \sigma_{\min}(\tilde{R}_t) \quad \text{(by } \sigma_{\min}(\tilde{R}_t) \leq \frac{1}{1.9\sqrt{\kappa}}) \\
&\geq \left(1 + \frac{\eta}{4\kappa}\right) \sigma_{\min}(\tilde{R}_t) \quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad\quad \text{(by } t \leq \frac{c}{\eta\kappa\rho})
\end{aligned}$$

Since $\sigma_{\min}(\tilde{R}_t) \geq \sigma_{\min}(R_t) \left(1 - \frac{\eta}{100\kappa}\right)$ by Proposition D.12, the conclusion follows. ∎

D.5. Proof of Proposition 4.3

Since $U_{t+1} = U_t - \eta\nabla f(U_t)$, we first write down the error for step $t + 1$:

$$\begin{aligned}
\left\|U_{t+1}U_{t+1}^\top - X^\star\right\|_F^2 &= \left\|(U_t - \eta\nabla f(U_t))(U_t^\top - \eta\nabla f(U_t)^\top) - X^\star\right\|_F^2 \\
&= \|U_tU_t^\top - X^\star\|_F^2 - 2\eta\langle\nabla f(U_t)U_t^\top, U_tU_t^\top - X^\star\rangle &\text{(68)} \\
&\quad + \langle 2\eta^2(U_tU_t^\top - X^\star), \nabla f(U_t)\nabla f(U_t)^\top\rangle + 4\eta^2\|\nabla f(U_t)U_t^\top\|_F^2 &\text{(69)} \\
&\quad - \langle 4\eta^3\nabla f(U_t)U_t^\top, \nabla f(U_t)\nabla f(U_t)^\top\rangle + \eta^4\|\nabla f(U_t)\nabla f(U_t)^\top\|_F^2 &\text{(70)}
\end{aligned}$$

Denote by

$$\Delta = U_tU_t^\top - E_tE_t^\top = E_tZ_t^\top + Z_tE_t^\top + E_tE_t^\top .$$

First we have that

$$\|\Delta\| \leq 2\|E_t\|\|Z_t\| + \|E_t\|_2^2 \leq O(\|E_t\|).$$

We first consider the degree one term of η in equation (68).

Claim D.21 *In the setting of this subsection, we have that:*

$$\langle\nabla f(U_t)U_t^\top, U_tU_t^\top - X^\star\rangle \geq (1 - O(\eta))\|(Z_tZ_t^\top - X^\star)Z_t\|_F^2 - O(\delta)\|Z_tZ_t^\top - X^\star\|_F^2 - O(\sqrt{dr}\|E_t\|).$$

Proof First, we have that

$$\langle \nabla f(U_t)U_t^\top, U_t U_t^\top - X^\star \rangle \geq \langle \nabla f(U_t)U_t^\top, Z_t Z_t^\top - X^\star \rangle - \|\nabla f(U_t)U_t^\top\|_F \|\Delta\|_F \qquad (71)$$

Since $\|U_t\| \leq O(1)$, we focus on the gradient of U_t. We divide $\nabla f(U_t)$ into the sum of three parts:

$$Y_1 = \frac{1}{m} \sum_{i=1}^m \langle A_i, Z_t Z_t^\top - X^\star \rangle A_i Z_t,$$

$$Y_2 = \frac{1}{m} \sum_{i=1}^m \langle A_i, Z_t Z_t^\top - X^\star \rangle A_i E_t,$$

$$Y_3 = \frac{1}{m} \sum_{i=1}^m \langle A_i, \Delta \rangle A_i U_t$$

By Lemma 2.3, we have that:

$$\|Y_1\| \leq \left\| (Z_t Z_t^\top - X^\star) Z_t \right\| + \delta \times \|Z_t Z_t^\top - X^\star\|_F \|Z_t\| \leq O(1), \qquad (72)$$

where $\|Z_t Z_t^\top - X^\star\|_F \leq O(\sqrt{r})$ by Corollary D.16. Since Y_1 has rank at most r, we get that $\|Y_1\|_F \leq O(\sqrt{r})$. For Y_2, we apply Lemma 2.3 again:

$$\|Y_2\| \leq \|Z_t Z_t^\top - X^\star\|\|E_t\| + \delta \times \|Z_t Z_t^\top - X^\star\|_F \|E_t\| \leq O(\|E_t\|).$$

By assumption $\|E_t\| \leq 1/d$, hence $\|Y_2\|_F \leq O(1)$. Finally we apply Lemma F.2 for Y_3:

$$\|Y_3\| \leq \|\Delta\|\|U_t\| + \delta \times (2\|E_t Z_t^\top\|_F + \|E_t E_t^\top\|_*)\|U_t\| \leq O(\|E_t\|)$$

To summarize, we have shown that $\|\nabla f(U_t)\|_F \leq O(\sqrt{r})$. Back to equation (71), we have shown that:

$$\|\nabla f(U_t)U_t^\top\|_F \|\Delta\|_F \leq O(\sqrt{dr})\|E_t\|.$$

For the other part in equation (71),

$$\begin{aligned} \langle \nabla f(U_t)U_t^\top, Z_t Z_t^\top - X^\star \rangle &\geq \langle M_t Z_t Z_t^\top, Z_t Z_t^\top - X^\star \rangle - \|M_t \Delta\|_F \|Z_t Z_t^\top - X^\star\|_F \\ &\geq \langle M_t Z_t Z_t^\top, Z_t Z_t^\top - X^\star \rangle - O(\sqrt{dr})\|E_t\|, \end{aligned}$$

because $\|M_t\| \leq O(1)$ from Lemma D.8. Lastly, we separate out the Δ term in M_t as follows:

$$\langle M_t Z_t Z_t^\top, Z_t Z_t^\top - X^\star \rangle = \langle \nabla f(Z_t)Z_t, Z_t Z_t^\top - X^\star \rangle + \qquad (73)$$

$$\frac{1}{m} \sum_{i=1}^m \langle A_i, E_t Z_t^\top + Z_t E_t^\top \rangle \langle A_i, Z_t Z_t^\top (Z_t Z_t^\top - X^\star) \rangle + \qquad (74)$$

$$\frac{1}{m} \sum_{i=1}^m \langle A_i, E_t E_t^\top \rangle \langle A_i, Z_t Z_t^\top (Z_t Z_t^\top - X^\star) \rangle \qquad (75)$$

Since Z_t is a rank-r matrix, by Lemma 2.2, Equation (73) is at least:

$$\left\langle Z_t Z_t^\top - X^\star, Z_t Z_t^\top \left(Z_t Z_t^\top - X^\star \right) \right\rangle - \delta \|Z_t\|^2 \|Z_t Z_t^\top - X^\star\|_F^2. \tag{76}$$

Equation (74) is similarly bounded by $O(\sqrt{dr}\|E_t\|)$ using Lemma 2.2, since $E_t Z_t^\top$ also has rank at most r. Finally for equation (75), while $E_t E_t^\top$ may have rank d, we can still apply Lemma F.1 and afford to lose a factor of d, because $\|E_t E_t^\top\| \le O(1/d^2)$. To summarize, we have shown that:

$$\langle \nabla f(U_t) U_t^\top, U_t U_t^\top - X^\star \rangle \ge \langle \nabla f(Z_t) Z_t^\top, Z_t Z_t^\top - X^\star \rangle - O(\sqrt{dr})\|E_t\|.$$

The conclusion follows by combining the above equation with (76). ∎

Next we work out the degree two term in equation (70).

Claim D.22 *In the setting of this subsection, we have that both*

$$|\langle (U_t U_t^\top - X^\star), \nabla f(U_t) \nabla f(U_t)^\top \rangle| \text{ and } \|\nabla f(U_t) U_t^\top\|_F^2$$
$$\lesssim \|(Z_t Z_t^\top - X^\star) Z_t\|_F^2 + \delta \|Z_t Z_t^\top - X^\star\|_F^2 + r\|E_t\|$$

Proof The idea of the proof is similar to that for equation (68) and (69). First of all, we have

$$\langle U_t U_t^\top - X^\star, \nabla f(U_t) \nabla f(U_t)^\top \rangle \ge \langle Z_t Z_t - X^\star, \nabla f(U_t) \nabla f(U_t)^\top \rangle - \|\Delta\| \times \|\nabla f(U_t) \nabla f(U_t)^\top\|_*$$

The second term is at most $O(r) \times \|E_t\|$ from our proof. Next,

$$\langle Z_t Z_t^\top - X^\star, \nabla f(U_t) \nabla f(U_t)^\top \rangle = \langle Z_t Z_t^\top - X^\star, M_t Z_t Z_t^\top M_t^\top \rangle + \langle Z_t Z_t^\top - X^\star, M_t \Delta M_t^\top \rangle$$

The second term is at most:

$$\|\Delta\| \times \|M_t^\top (Z_t Z_t^\top - X^\star) M_t\|_* \lesssim r\|E_t\|$$

Lastly, we expand out M_t to obtain:

$$\langle Z_t Z_t^\top - X^\star, M_t Z_t Z_t^\top M_t^\top \rangle = \frac{1}{m} \sum_{i=1}^m \langle A_i, U_t U_t^\top - X^\star \rangle \langle Z_t Z_t^\top - X^\star, A_i Z_t Z_t^\top M_t^\top \rangle$$

$$= \frac{1}{m} \sum_{i=1}^m \langle A_i, Z_t Z_t^\top - X^\star \rangle \langle A_i, W \rangle + \frac{1}{m} \sum_{i=1}^m \langle A_i, \Delta \rangle \langle A_i, W \rangle$$

where we denote by $W = (Z_t Z_t^\top - X^\star) M_t Z_t Z_t^\top$. Clearly, the rank of W is at most r and it is not hard to see that the spectral norm of W is $O(1)$. For the first part, we apply Lemma 2.2 to obtain:

$$|\langle Z_t Z_t^\top - X^\star, W \rangle| + \delta \times \|Z_t Z_t^\top - X^\star\|_F \|W\|_F \lesssim \|(Z_t Z_t^\top - X^\star) Z_t\|_F^2 + \|Z_t Z_t^\top - X^\star\|_F^2 \tag{77}$$

where we used that $\|M_t\| \le O(1)$ and $\delta \lesssim 1/\sqrt{r}$. For the second part, we apply Lemma 2.2 and Lemma F.1 together on Δ to obtain:

$$|\langle \Delta, W \rangle| + \delta \times (2\|E_t Z_t^\top\|_F + \|E_t E_t^\top\|_*)\|W\|_F \le O(r) \times \|E_t\|$$

where we used the assumption that $\|E_t\| \le 1/d$. The proof for $\|\nabla f(U_t)U_t^\top\|_F^2$ is similar. The difference is that we will obtain $W' = Z_t Z_t^\top M_t Z_t Z_t^\top$ instead. To bound $\langle Z_t Z_t - X^\star, W' \rangle$, we use Lemma 2.2 and Lemma F.1 to control M_t. The details are left to the readers. ∎

Finally we consider the degree three and four terms of η in equation (70):

$$\langle 4\eta^3 \nabla f(U_t)U_t^\top, \nabla f(U_t)\nabla f(U_t)^\top \rangle \le O(\eta^3) \times \|\nabla f(U_t)\|_F^2,$$

because $\|U_t U_t^\top - X^\star\| \le O(1)$ and $\|\nabla f(U_t)\| \le O(1)$. For the gradient of U_t, we have already decomposed it to the sum of Y_1, Y_2 and Y_3. And our proof already implies that:

$$\|Y_1\|_F \le O(\sqrt{r}) \times \|Z_t Z_t^\top - X^\star\|, \text{ and}$$
$$\|Y_2\|_F, \|Y_3\|_F \le O(\sqrt{d}) \times \|E_t\|$$

Combining all results together, we get that:

$$\|\nabla f(U_t)\|_F^2 \le O(r)\|Z_t Z_t^\top - X^\star\|^2 + O(d\|E_t\|^2)$$

Hence equation (70) is at most:

$$O(r\eta^3)\|Z_t Z_t^\top - X^\star\|_F^2 + O(\|E_t\|)$$

Combining the above equation with Claim D.21 and D.22, we have shown that:

$$\|U_{t+1}U_{t+1}^\top - X^\star\|_F^2 \le \|U_t U_t^\top - X^\star\|_F^2 - (\eta - O(\eta^2))\|(Z_t Z_t^\top - X^\star)Z_t\|_F^2 + O(\sqrt{dr})\|E_t\| + O(\eta\delta + \eta^2\delta + r\eta^3)\|Z_t Z_t^\top - X^\star\|_F^2$$

Lastly we show that:

$$\|(Z_t Z_t^\top - X^\star)Z_t\|_F^2 \ge \frac{1}{4\kappa}\|Z_t Z_t^\top - X^\star\|_F^2. \tag{78}$$

The conclusion follows since it is not hard to show that

$$\|U_t U_t^\top - X^\star\|_F^2 = \|Z_t Z_t^\top - X^\star\|_F^2 \pm O(\sqrt{dr})\|E_t\|.$$

The rest of the proof is dedicated to equation (78). Without loss of generality, let us assume that [12]

$$X^* = \begin{pmatrix} \Sigma^\star & 0 \\ 0 & 0 \end{pmatrix} \quad \text{and} \quad Z_t Z_t^\top = U\Sigma U^\top,$$

where $U = \begin{pmatrix} U_1 \\ U_2 \end{pmatrix} \in \mathbb{R}^{d\times r}$ is column orthogonal, and $U_1 \in \mathbb{R}^{r\times r}$. We have:

$$\left\|Z_t Z_t^\top - X^\star\right\|_F^2 = \left\|U_1 \Sigma U_1^\top - \Sigma^\star\right\|_F^2 + 2\left\|U_2 \Sigma U_1^\top\right\|_F^2 + \left\|U_2 \Sigma U_2^\top\right\|_F^2,$$

---

12. To see this, we can extend U^* to a d by d column orthogonal matrix. Let $A\Sigma A^\top$ be the SVD of $Z_t Z_t^\top$. Then $U = U^{*\top}A$ is column orthogonal.

and

$$\left\|\left(Z_t Z_t^\top - X^\star\right) Z_t\right\|_F^2$$
$$= \left\|\left(\Sigma^\star U_1 - U_1 \Sigma\right) \Sigma^{1/2}\right\|_F^2 + \left\|\left(U_2 \Sigma\right) \Sigma^{1/2}\right\|_F^2$$
$$\geq \sigma_{\min}(\Sigma) \left(\|\Sigma^\star U_1 - U_1 \Sigma\|_F^2 + \|U_2 \Sigma\|_F^2\right)$$

Moreover, using the fact that $\begin{pmatrix} U_1 \\ U_2 \end{pmatrix}$ is column orthonormal, we can get $\|U_1\|, \|U_2\| \leq 1$, which implies:

$$\left\|U_1 \Sigma U_1^\top - \Sigma^\star\right\|_F^2 + 2 \left\|U_2 \Sigma U_1^\top\right\|_F^2 + \left\|U_2 \Sigma U_2^\top\right\|_F^2$$
$$\leq \left\|U_1 \Sigma U_1^\top - \Sigma^\star\right\|_F^2 + 2 \|U_2 \Sigma\|_F^2 \left\|U_1^\top\right\|^2 + \|U_2 \Sigma\|_F^2 \|U_2^\top\|^2$$
$$\lesssim \left\|U_1 \Sigma U_1^\top - \Sigma^\star\right\|_F^2 + \|U_2 \Sigma\|_F^2$$

By assumption we have $\sin(Z_t, U^*) \leq \frac{1}{3}$, which gives us $\sigma_{\min}(U_1) \geq 1/4$. Thus, considering $\left\|U_1 \Sigma U_1^\top - \Sigma^\star\right\|_F^2$ we have:

$$\left\|U_1 \Sigma U_1^\top - \Sigma^\star\right\|_F^2 = \operatorname{tr}\left(U_1 \Sigma U_1^\top U_1 \Sigma U_1^\top\right) - 2\operatorname{tr}\left(U_1 \Sigma U_1^\top \Sigma^\star\right) + \operatorname{tr}\left(\Sigma^{\star 2}\right)$$

Using $U_1^\top U_1 \preceq I$ we have $\operatorname{tr}\left(U_1 \Sigma U_1^\top U_1 \Sigma U_1^\top\right) \leq \operatorname{tr}\left(U_1 \Sigma^2 U_1^\top\right)$. Using $\sigma_{\min}(U_1) \geq 1/4$ we also have $\operatorname{tr}\left(\Sigma^{\star 2}\right) \lesssim \operatorname{tr}(\Sigma^\star U_1 U_1^\top \Sigma^\star)$. Therefore we obtain:

$$\left\|U_1 \Sigma U_1^\top - \Sigma^\star\right\|_F^2 \lesssim \left(\operatorname{tr}\left(U_1 \Sigma^2 U_1^\top\right) - 2\operatorname{tr}\left(U_1 \Sigma U_1^\top \Sigma^\star\right) + \operatorname{tr}\left(\Sigma^\star U_1 U_1^\top \Sigma^\star\right)\right)$$
$$= \|U_1 \Sigma - \Sigma^\star U_1\|_F^2$$

Putting everything together, we get:

$$\left\langle Z_t Z_t^\top - X^\star, Z_t Z_t^\top \left(Z_t Z_t^\top - X^\star\right)\right\rangle = \left\|\left(Z_t Z_t^\top - X^\star\right) Z_t\right\|_F^2$$
$$\geq \sigma_{\min}(\Sigma) \left(\left\|U_1 \Sigma - \Sigma^\star U_1^\top\right\|_F^2 + \|U_2 \Sigma\|_F^2\right)$$
$$= \Omega(\sigma_{\min}(\Sigma)) \left(\left\|U_1 \Sigma - \Sigma^\star U_1^\top\right\|_F^2 + \|U_2 \Sigma\|_F^2\right)$$
$$= \Omega(\sigma_{\min}(Z_t)) \left\|Z_t Z_t^\top - X^\star\right\|_F^2$$

By the assumption that $\sigma_{\min}(Z_t)^2 \geq 1/(4\kappa) = \Omega(\delta)$, we complete the proof of equation (78).

Appendix E. Missing proofs in Section A

Proof [Proof of Lemma A.1]

Let us first consider the case when X is a rank-1 matrix. Suppose $X = aa^\top$ with $\|a\| = 1$. We then have:

$$\frac{1}{m}\sum_{i=1}^{m}\langle A_i, X\rangle A_i 1_{|\langle A_i, X\rangle|\leq R} = \frac{1}{m}\sum_{i=1}^{m}\langle A_i, aa^\top\rangle A_i$$

$$= \frac{1}{m}\sum_{i=1}^{m}\langle x_i, a\rangle^2 x_i x_i^\top 1_{\langle x_i, a\rangle^2 \leq R^2}$$

We define:

$$H(x_1, \cdots, x_m) := \sup_{u,v\in\mathbb{R}^d, \|u\|_2=\|v\|_2=1}\left|\frac{1}{m}\sum_{i=1}^{m}\langle x_i, u\rangle^2\langle x_i, v\rangle^2 1_{\langle x_i,u\rangle^2\leq R^2} - 2\langle u, v\rangle^2 - 1\right|$$

It suffices to bound H because by definition, for every $X = aa^\top$ being a rank one matrix, with $\|a\| = 1$, we have that

$$\left\|\frac{1}{m}\sum_{i}\langle x_i, a\rangle^2 x_i x_i^\top 1_{\langle x_i, a\rangle^2\leq R^2} - 2X - I\right\| \leq H(x_1, \cdots, x_m)$$

Let us further decompose H into two terms:

$$H(x_1, \cdots, x_m) \leq \sup_{u,v\in\mathbb{R}^d, \|u\|_2=\|v\|_2=1}\left|\frac{1}{m}\sum_{i=1}^{m}\langle x_i, u\rangle^2\langle x_i, v\rangle^2 1_{\langle x_i,u\rangle^2\leq R^2}1_{\langle x_i,v\rangle^2\leq R^2} - 2\langle u, v\rangle^2 - 1\right|$$

$$+ \sup_{u,v\in\mathbb{R}^d, \|u\|_2=\|v\|_2=1}\left|\frac{1}{m}\sum_{i=1}^{m}\langle x_i, u\rangle^2\langle x_i, v\rangle^2 1_{\langle x_i,u\rangle^2\leq R^2}1_{\langle x_i,v\rangle^2> R^2}\right| \tag{79}$$

Let us bound the two term separately. For the first term, for every unit vectors u, v, we define functions $f_{u,v} : \mathbb{R}^d \to \mathbb{R}$ as $f_{u,v}(x) = \langle u, x\rangle\langle v, x\rangle 1_{\langle x_i,u\rangle^2\leq R^2}1_{\langle x_i,v\rangle^2\leq R^2}$. We have that $f_{u,v}(x) \leq R^4$. Thus by the symmetrization technique and the contraction principle(e.g., see Corollary 4.7 in Adamczak et al. (2010)), we have:

$$\mathbb{E}\left[\sup_{u,v\in\mathbb{R}^d, \|u\|_2=\|v\|_2=1}\left|\sum_{i=1}^{m}\left(f_{u,v}(x_i)^2 - \mathbb{E}\left[f_{u,v}(x_i)^2\right]\right)\right|\right] \leq 8R^2\mathbb{E}\left[\sup_{u,v\in\mathbb{R}^d, \|u\|_2=\|v\|_2=1}\left|\sum_{i=1}^{m}\varepsilon_i f_{u,v}(x_i)\right|\right]$$

where $\{\varepsilon_i\}_{i=1}^{m}$ is a set of i.i.d. Rademacher random variables. We can further bound the right hand side of the inequality above by:

$$\mathbb{E}\left[\sup_{u,v\in\mathbb{R}^d, \|u\|_2=\|v\|_2=1}\left|\sum_{i=1}^{m}\varepsilon_i f_{u,v}(x_i)\right|\right] \leq \mathbb{E}\left[\left\|\sum_{i=1}^{m}\varepsilon_i x_i x_i^\top\right\|\right]$$

A standard bound on the norm of Gaussian random matrices gives us: $\mathbb{E}\left[\left\|\sum_{i=1}^{m}\varepsilon_i x_i x_i^\top\right\|\right] \lesssim \sqrt{md}$. Therefore, we conclude that

$$\mathbb{E}\left[\sup_{u,v\in\mathbb{R}^d, \|u\|_2=\|v\|_2=1}\left|\sum_{i=1}^{m}\left(f_{u,v}(x_i)^2 - \mathbb{E}\left[f_{u,v}(x_i)^2\right]\right)\right|\right] \lesssim R^2\sqrt{md}$$

Now let us consider the expectation of $f_{u,v}(x_i)^2$, a direct calculation shows that

$$\left|\mathbb{E}[f_{u,v}(x_i)^2] - 2\langle u, v\rangle^2 - 1\right| = \left|\mathbb{E}\left[\langle x_i, u\rangle^2 \langle x_i, v\rangle^2 1_{\langle x_i, u\rangle^2 \leq R^2} 1_{\langle x_i, v\rangle^2 \leq R^2}\right] - \mathbb{E}\left[\langle x_i, u\rangle^2 \langle x_i, v\rangle^2\right]\right|$$
$$\leq \mathbb{E}\left[\langle x_i, u\rangle^2 \langle x_i, v\rangle^2 1_{\langle x_i, u\rangle^2 \leq R^2} 1_{\langle x_i, v\rangle^2 > R^2}\right] + \mathbb{E}\left[\langle x_i, u\rangle^2 \langle x_i, v\rangle^2 1_{\langle x_i, u\rangle^2 > R^2}\right]$$
$$\leq 2\mathbb{E}[\langle x_i, u\rangle^2 \langle x_i, v\rangle^2 1_{\langle x_i, u\rangle^2 > R^2}]$$

An elementary calculation of Gaussian variables gives us:

$$\mathbb{E}[\langle x_i, u\rangle^2 \langle x_i, v\rangle^2 1_{\langle x_i, u\rangle^2 > R^2}] \lesssim R^4 e^{-R^2/2} \tag{80}$$

Putting everything together, for $R \geq 1$ we are able to bound the first term of equation (79) by:

$$\mathbb{E}\left[\sup_{u,v\in\mathbb{R}^d, \|u\|_2 = \|v\|_2 = 1} \left|\frac{1}{m}\sum_{i=1}^{m} \langle x_i, u\rangle^2 \langle x_i, v\rangle^2 1_{\langle x_i, u\rangle^2 \leq R^2} 1_{\langle x_i, v\rangle^2 \leq R^2} - 2\langle u, v\rangle^2 - 1\right|\right] \lesssim R^4 \left(\sqrt{\frac{d}{m}} + e^{-R^2/2}\right) \tag{81}$$

Moreover, for every $u, v \in \mathbb{R}^d, \|u\| = \|v\| = 1$ we know that $f_{u,v}(x) \leq R^2$, we can apply (Adamczak et al., 2010, Lemma 4.8) to transform the bound above into a high probability bound. We have that for every $s \in [0, 1]$, with probability at least $1 - e^{-\Omega(s^2 m/(dR^4))}$,

$$\sup_{u,v\in\mathbb{R}^d, \|u\|_2 = \|v\|_2 = 1} \left|\frac{1}{m}\sum_{i=1}^{m} \langle x_i, u\rangle^2 \langle x_i, v\rangle^2 1_{\langle x_i, u\rangle^2 \leq R^2} 1_{\langle x_i, v\rangle^2 \leq R^2} - 2\langle u, v\rangle^2 - 1\right| \lesssim R^4 \left(\sqrt{\frac{d}{m}} + e^{-R^2/2}\right) + s$$

Picking $s = \left(\frac{R^2\sqrt{d}}{\sqrt{m}} \log \frac{1}{q}\right)$ with $R = \Theta\left(\log\left(\frac{1}{\delta}\right)\right)$ we obtain that

$$\sup_{u,v\in\mathbb{R}^d, \|u\|_2 = \|v\|_2 = 1} \left|\frac{1}{m}\sum_{i=1}^{m} \langle x_i, u\rangle^2 \langle x_i, v\rangle^2 1_{\langle x_i, u\rangle^2 \leq R^2} 1_{\langle x_i, v\rangle^2 \leq R^2} - 2\langle u, v\rangle^2 - 1\right| \leq \delta$$

For the second term of equation (79), we have that

$$\sup_{u,v\in\mathbb{R}^d, \|u\|_2 = \|v\|_2 = 1} \left|\frac{1}{m}\sum_{i=1}^{m} \langle x_i, u\rangle^2 \langle x_i, v\rangle^2 1_{\langle x_i, u\rangle^2 \leq R^2} 1_{\langle x_i, v\rangle^2 > R^2}\right|$$
$$\leq \sup_{v\in\mathbb{R}^d, \|v\|_2 = 1} \left|\frac{1}{m}\sum_{i=1}^{m} R^2 \langle x_i, v\rangle^2 1_{\langle x_i, v\rangle^2 > R^2}\right|$$

By (Adamczak et al., 2010, Theorem 3.6 and Remark 3.10), we have that for every $s > 0$, with probability at $1 - e^{-\Omega(s\sqrt{d})}$:

$$\sup_{v\in\mathbb{R}^d, \|v\|_2 = 1} \left|\frac{1}{m}\sum_{i=1}^{m} R^2 \langle x_i, v\rangle^2 1_{\langle x_i, v\rangle^2 > R^2}\right| \lesssim R^2 s^2 \left(\frac{d^2}{m} + \frac{d^2}{m} s^2 R^{-2} \log^2 \frac{m}{n}\right) \tag{82}$$

Taking $s = \Omega\left(\frac{\log\frac{1}{q}}{\sqrt{d}}\right)$, putting everything together we prove the Lemma for the case when $X = aa^\top$ is rank one.

For X of general rank, the proof follows by decomposing X to a sum of rank one singular vectors and apply triangle inequality directly. ∎

Appendix F. Restricted Isometry Properties

In this section we list additional properties we need for the set of measurement matrices $\{A_i\}_{i=1}^m$. Lemma 2.2 follows from the definition of RIP matrices. The rest three Lemmas are all direct implications of Lemma 2.2.

Proof [Proof of Lemma 2.3] For every $x \in \mathbb{R}^d, y \in \mathbb{R}^{d'}$ of norm at most 1, we have:

$$
\frac{1}{m}\sum_{i=1}^m \langle A_i, X\rangle x^\top A_i Ry - x^\top XRy = \frac{1}{m}\sum_{i=1}^m \langle A_i, X\rangle\langle A_i, xy^\top R^\top\rangle - x^\top XRy
$$
$$
\leq \langle X, xy^\top R^\top\rangle + \delta\|X\|_F\|xy^\top R^\top\| - x^\top XRy
$$
$$
\leq \delta\|X\|_F\|R\|_2
$$

The first inequality uses Lemma 2.2. ∎

The following Lemmas deal with matrices that may have rank bigger than r. The idea is to decompose the matrix into a sum of rank one matrices via SVD, and then apply Lemma 2.2.

Lemma F.1 *Let $\{A_i\}_{i=1}^m$ be a family of matrices in $\mathbb{R}^{d\times d}$ that satisfy (r, δ)-restricted isometry property. Then for any matrices $X, Y \in \mathbb{R}^{d\times d}$, where the rank of Y is at most r, we have:*

$$
\left|\frac{1}{m}\sum_{i=1}^m \langle A_i, X\rangle\langle A_i, Y\rangle - \langle X, Y\rangle\right| \leq \delta\|X\|_*\|Y\|_F
$$

Proof Let $X = UDV^\top$ be its SVD. We decompose $D = \sum_{i=1}^d D_i$ where each D_i contains only the i-th diagonal entry of D, and let $X_i = UD_iV^\top$ for each $i = 1, \ldots, d$. Then we have:

$$
\frac{1}{m}\sum_{i=1}^m \langle A_i, X\rangle\langle A_i, Y\rangle = \sum_{j=1}^d\left(\frac{1}{m}\sum_{i=1}^m \langle A_i, X_j\rangle\langle A_i, Y\rangle\right)
$$
$$
\leq \sum_{j=1}^d\left(\langle X_j, Y\rangle + \delta\|X_j\|_F\|Y\|_F\right) = \langle X, Y\rangle + \delta\|X\|_*\|Y\|_F
$$

∎

Lemma F.2 *Let $\{A_i\}_{i=1}^m$ be a family of matrices in $\mathbb{R}^{d\times d}$ that satisfy $(1, \delta)$-restricted isometric property. Then for any matrix $X \in \mathbb{R}^{d\times d}$ and matrix $R \in \mathbb{R}^{d\times d'}$, where d' can be any positive integer, we have:*

$$
\left\|\frac{1}{m}\sum_{i=1}^m \langle A_i, X\rangle A_i R - XR\right\| \leq \delta\|X\|_* \times \|R\|.
$$

The following variant is also true:

$$
\left\|\frac{1}{m}\sum_{i=1}^m \langle A_i, X\rangle UA_i R - UXR\right\| \leq \delta\|X\|_* \times \|U\| \times \|R\|,
$$

where U is any matrix in $\mathbb{R}^{d\times d}$.

Proof Let $X = UDV^\top$ be its SVD. We define X_i and D_i the same as in the proof of Lemma F.1, for each $i = 1, \ldots, d$.

For every $x \in \mathbb{R}^d$, $y \in \mathbb{R}^{d'}$ with norm at most one, we have:

$$\frac{1}{m} \sum_{i=1}^m \langle A_i, X \rangle \, x^\top A_i R y - x^\top X R y$$

$$= \sum_{j=1}^d \left(\frac{1}{m} \sum_{i=1}^m \langle A_i, X_j \rangle \langle A_i, xy^\top R^\top \rangle \right) - x^\top X R y$$

$$\leq \sum_{j=1}^d \left(\langle X_j, xy^\top R^\top \rangle + \delta \|X_j\|_F \|R\| \right) - x^\top X R y = \delta \|X\|_* \|R\|.$$

The variant can be proved by the same approach (details omitted). ∎

Asymmetric sensing matrices. Recall that when each A_i is asymmetric, we simply use $(A_i + A_i^\top)/2$ instead of A_i as our sensing matrix. While $\{(A_i + A_i)/2\}_{i=1}^m$ may only ensure the restricted isometry property for symmetric matrices, we have the same inequality when the matrix X in Lemma 2.3 is symmetric, which is the case for all our applications of Lemma 2.3: [13]

$$\left\| \frac{1}{m} \sum_{i=1}^m \langle \frac{A_i + A_i^\top}{2}, X \rangle (A_i + A_i^\top) R / 2 - X R \right\| \leq \delta \|X\|_F \|R\|.$$

Since X is symmetric, $\langle A_i, X \rangle = \langle A_i^\top, X \rangle$. The above equation then follows by applying Lemma 2.3 twice, with $\{A_i\}_{i=1}^m$ and $\{A_i^\top\}_{i=1}^m$ as sensing matrices respectively.

For the applications of Lemma 2.2 in Equations (28), (29), (74), (75) and (77), we note that either X or Y is symmetric in all applications. Suppose that X is symmetric, then we have the following when we use $(A_i + A_i^\top)/2$ as the i-th sensing matrix:

$$\left| \frac{1}{m} \sum_{i=1}^m \langle \frac{A_i + A_i^\top}{2}, X \rangle \langle \frac{A_i + A_i^\top}{2}, Y \rangle - \langle X, Y \rangle \right| \leq \delta \|X\|_F \left\| \frac{Y + Y^\top}{2} \right\|_F$$

It is straightforward to verify that our proof still holds using the above inequality instead. The details for left for the readers.

---

13. More precisely, X corresponds to any one of $U_t U_t^\top - X^\star$, $E_t E_t^\top$, $Z_t Z_t^\top$, or their linear combinations.