Noise Regularizes Over-parameterized Rank One Matrix Recovery, Provably

Tianyi Liu
ByteDance

Yan Li
Georgia Tech

Enlu Zhou
Georgia Tech

Tuo Zhao
Georgia Tech

Abstract

We investigate the role of noise in optimization algorithms for learning over-parameterized models. Specifically, we consider the recovery of a rank one matrix $Y^* \in \mathbb{R}^{d\times d}$ from a noisy observation Y using an over-parameterization model. We parameterize the rank one matrix Y^* by $XX^\top$, where $X \in \mathbb{R}^{d\times d}$. We then show that under mild conditions, the estimator, obtained by the randomly perturbed gradient descent algorithm using the square loss function, attains a mean square error of $\mathcal{O}(\sigma^2/d)$, where σ^2 is the variance of the observational noise. In contrast, the estimator obtained by gradient descent without random perturbation only attains a mean square error of $\mathcal{O}(\sigma^2)$. Our result partially justifies the implicit regularization effect of noise when learning over-parameterized models, and provides new understandings of training over-parameterized neural networks.

1 INTRODUCTION

Deep neural networks have revolutionized many research areas, and achieved the state-of-the-art performance in many computer vision (Krizhevsky et al., 2012; Goodfellow et al., 2014; Long et al., 2015), natural language processing (Graves et al., 2013; Bahdanau et al., 2014; Young et al., 2018) and signal processing tasks (Yu and Deng, 2010). Such huge successes cannot be well explained by conventional wisdom. These deep neural networks are significantly over-parameterized – using more parameters than statistically necessary. However, training these neural networks does not require explicit regularization or constraints to control the model complexity.

There have been two major lines of theoretical research on demystifying such an over-parameterization phenomenon. One line of research attempts to investigate the training of deep neural networks from a pure optimization perspective. Liang et al. (2018); Sharifnassab et al. (2020); Liang et al. (2019) show that under properly simplified settings, the over-parameterization can eliminate spurious local optima of the training objective, and all obtained local optima become global. Therefore, over-parameterization makes the optimization landscape benign, which eases the training of neural network. However, these results are not relevant to the generalization performance of deep neural networks.

Another line of research attempts to connects the deep neural networks to reproducing kernel functions. Du et al. (2018); Jacot et al. (2018); Allen-Zhu et al. (2018); Arora et al. (2019) show that under certain conditions, training the over-parameterized neural networks by gradient descent is equivalent to training a kernel machine, which is often referred to as Neural Tangent Kernel (NTK) in existing literature. Therefore, adding more neurons only makes the behavior of deep neural networks behave more close to that of their corresponding reproducing kernel functions. By further exploiting such a connection, they show that the global optima of the training objective can be obtained by the gradient descent (GD) algorithm, However, as shown in E et al. (2020), these results cannot explain the generalization performance well, as the equivalent reproducing kernel functions still suffer from the curse of dimensionality.

Complementary to the aforementioned two lines of research, there have been some empirical investigations on the role of noise in optimization algorithms for training over-parameterized neural networks. For example, Keskar et al. (2016) show that the stochastic gradient descent (SGD) algorithms with small batch sizes yield significantly better generalization performance than those with large batch sizes. This clearly indicates that the noise plays a very important role on implicitly controlling the model complexity of over-parameterized neural networks. Unfortunately, due to the complex

Proceedings of the 25th International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

structures of deep neural networks and current technical limit, establishing theory for understanding the noise in SGD is very challenging. Though some of the aforementioned work consider SGD, they only consider small learning rates and large batch sizes to make the noise negligible such that its behavior is close to GD. Hence, their results cannot justify the advantage of SGD for training over-parameterized models.

To flesh out our understanding the role of noise for training over-parameterized models, we propose to analyze a simpler but nontrivial alternative problem – over-parameterized matrix factorization using perturbed gradient descent (P-GD). Specifically, we consider the recovery of a symmetric rank one matrix $Y^* \in \mathbb{R}^{d\times d}$ from its noisy observation Y under over-parameterization model. Different from existing work, which usually parameterizes Y^* as the outer product of two vectors, we factorize Y^* as the product of two matrices $XX^\top$, where $X \in \mathbb{R}^{d\times d}$. Therefore, we are essentially using d^2 parameters rather than statistically necessary d parameters. To recover Y^*, we solve the following optimization problem,

$$\min_{X\in\mathbb{R}^{d\times d}} \frac{1}{4}\left\|Y - XX^\top\right\|_{\mathrm{F}}^2. \tag{1}$$

We then solve (1) using a perturbed form of gradient descent P-GD, which injects independent noise to iterates, and then evaluates gradient at the perturbed iterates. Note that our algorithm is different from SGD in terms of the noise. For our algorithm, we inject independent noise to the iterate X_t and use the gradient evaluated at the perturbed iterates. The noise of SGD, in contrast, usually comes from the training sample. As a consequence, the noise of SGD has very complex dependence on the iterate, which is difficult to analyze.

We further analyze the computational and statistical properties of the P-GD algorithm. Specifically, at the early stage, noise helps the algorithm to avoid regions with undesired landscape, including saddle points. After entering the region with benign landscape, the noise induces an implicit regularization effect, and P-GD eventually converges to an estimator $\widehat{X}$, which attains a mean square error of $\mathcal{O}(\sigma^2/d)$ with overwhelming probability, i.e.,

$$\frac{1}{d^2}\left\|\widehat{X}\widehat{X}^\top - Y^*\right\|_{\mathrm{F}}^2 = \mathcal{O}_P\Big(\frac{\sigma^2}{d}\Big),$$

where σ^2 is the variance of the observational noise. For comparison, if we solve (1) by GD without random perturbation, and the obtained estimator only attains a mean square error of $\mathcal{O}(\sigma^2)$. To the best of our knowledge, this is the first theoretical result towards understanding the role of noise in training over-parameterized models.

Our work is closely related to Li et al. (2017), which analyze GD for solving over-parameterized matrix sensing problem. Specifically, they show that when initialized at a sufficiently small magnitude, GD also has an implicit regularization effect and can approximately recover low rank matrix under the RIP condition. Their theory, however, only works for noiseless cases. Our theory complements their results under the noisy setting, and demonstrates that the noise of optimization algorithms can also contribute to the implicit regularization effects for training over-parameterized models.

The rest of the paper is organized as follows: Section 2 introduces the rank-1 matrix factorization problem and the perturbed gradient descent algorithm to solve it. Section 3 presents the main theorem showing that P-GD converges to solutions with smaller mean square error than GD. Section 4 verifies our theoretical result numerically on rank-1 matrix recovery, rank-r matrix recovery and also rectangular matrix recovery. The discussion on the extension of our theoretical results and also related literature are presented in Section 5.

Notations: Let $\mathcal{S}$ be a subspace of $\mathbb{R}^d$, we use $\mathrm{Proj}_{\mathcal{S}}(\cdot)$ to denote the projection of a vector or matrix to $\mathcal{S}$. For a vector $v \in \mathbb{R}^d$ and matrix $A \in \mathbb{R}^{d\times d}$, we use $\mathrm{Id}_v A$ to denote the projection of each column of A onto the subspace $\mathrm{span}(v) = \{x \in \mathbb{R}^d | x = \alpha v, \alpha \in \mathbb{R}\}$. Id is the identity matrix. The ball with radius r in $\mathbb{R}^d$ and its sphere are denoted as $\mathbb{B}(1)$ and $\mathbb{S}(1)$, respectively. For matrices $A, B \in \mathbb{R}^{n\times m}$, we use $\langle A, B\rangle$ to denote the Frobenius inner product, i.e., $\langle A, B\rangle = \mathrm{tr}\big(A^\top B\big)$. $\|A\|_{\mathrm{F}}$ and $\|A\|_2$ denotes the Frobenius norm and spectral norm of A, respectively.

2 MODEL AND ALGORITHM

We first describe the over-parameterized rank one matrix factorization problem. Specifically, we observe a matrix $Y \in \mathbb{R}^{d\times d}$, where

$$Y = Y^* + \Gamma,$$

where $Y \in \mathbb{R}^{d\times d}$ is an unknown rank one matrix, and $\Gamma \in \mathbb{R}^{d\times d}$ is a random noise matrix with each entry i.i.d. sampled from some sub-Gaussian distribution with $\mathbb{E}\Gamma_{ij} = 0$ and $\mathbb{E}\Gamma_{ij}^2 = \sigma^2$. We recover Y^* by solving the following problem:

$$\widehat{X} = \operatorname*{argmin}_{X\in\mathbb{R}^{d\times d}} \mathcal{F}(X),$$
$$\text{where } \mathcal{F}(X) = \frac{1}{4}\left\|XX^\top - Y\right\|_{\mathrm{F}}^2. \tag{2}$$

The estimator of Y^* can be obtained by $\widehat{Y} = \widehat{X}\widehat{X}^\top$. Here $\widehat{Y}$ is over-parameterized with d^2 parameters in $\widehat{X}$, while the intrinsic dimension of the rank one matrix Y^* is only d. We do not use any explicit regularizer to control the search space of X.

We then describe the perturbed gradient descent (P-GD) algorithm for solving (2). Specifically, at the $(t+1)$-th iteration, we first inject a random noise matrix $W_t \in \mathbb{R}^{d\times d}$ to X_t,

$$\widetilde{X}_t = X_t + W_t,$$

where each column of W_t is independently sampled from $\mathrm{UNIF}(\mathbb{S}(\nu))$, and $\mathbb{S}(\nu)$ denotes the hypersphere with radius ν centered at 0. Note that we have $\|W_t\|_{\mathrm{F}}^2 = d\nu^2$. We then update X_t using the gradient of $\mathcal{F}(X)$ at $\widetilde{X}_t$,

$$\begin{aligned} X_{t+1} &= X_t - \eta\nabla\mathcal{F}(\widetilde{X}_t) \\ &= X_t - \eta\left(\widetilde{X}_t\widetilde{X}_t^\top - Y_{\mathrm{sym}}\right)\widetilde{X}_t, \end{aligned} \tag{3}$$

where $Y_{\mathrm{sym}} = (Y + Y^\top)/2$. The P-GD algorithm is essentially solving the following stochastic optimization problem,

$$\min_{X\in\mathbb{R}^{d\times d}} \widetilde{\mathcal{F}}(X) = \mathbb{E}_W\mathcal{F}(X+W), \tag{4}$$

where each column of W is independently sampled from $\mathrm{UNIF}(\mathbb{S}(\nu))$. Note that (4) can be viewed as a smooth approximation of (2) by convolution using a uniform kernel. The smoothing effect further induces implicit regularization effect to the estimator.

The P-GD algorithm is also related to the randomized smoothing in existing literature. It was first proposed by Duchi et al. (2012) to handle convex non-smooth optimization. Zhou et al. (2019); Jin et al. (2017); Lu et al. (2019) further show that the random perturbation can also help escape from saddle points and spurious optima.

3 CONVERGENCE ANALYSIS

We study the convergence properties of our proposed perturbed gradient descent (P-GD) algorithm. Before presenting our main results, we first introduce the subspace dissipative condition, which is frequently used in our proof and is defined as follows.

Definition 3.1 (Subspace Dissipativity). *Let $\mathcal{S}$ be a subspace of $\mathbb{R}^d$ and $x_{\mathcal{S}} = \mathrm{Proj}_{\mathcal{S}}(x)$ be the projection of $\forall x \in \mathbb{R}^d$ into $\mathcal{S}$. For any operator $\mathcal{H}: \mathbb{R}^d \to \mathbb{R}^d$, we say that $\mathcal{H}$ is $(c_{\mathcal{S}}, \gamma_{\mathcal{S}}, \mathcal{S})$-subspace dissipative with respect to (w.r.t.) the subset $\mathcal{X}^* \subseteq \mathbb{R}^d$ over the set $\mathcal{X} \supseteq \mathcal{X}^*$, if for every $x \in \mathcal{X}$, there exist an $x^* \in \mathcal{X}^*$ and two positive universal constants $c_{\mathcal{S}}$ and $\gamma_{\mathcal{S}}$ such that*

$$\langle \mathrm{Proj}_{\mathcal{S}}(\mathcal{H}(x)), x_{\mathcal{S}} - x_{\mathcal{S}}^*\rangle \geq c_{\mathcal{S}}\|x_{\mathcal{S}} - x_{\mathcal{S}}^*\|_2^2 - \gamma_{\mathcal{S}}. \tag{5}$$

Here, $\mathcal{X}$ is called the subspace dissipative region of the operator $\mathcal{H}$.

The intuition behind the subspace dissipative condition is that $\mathrm{Proj}_{\mathcal{S}}(\mathcal{H}(x))$ has a positive fraction pointing towards x_S^* up to certain perturbation. When the algorithm iterates along $-\mathcal{H}(x)$, its projection in $\mathcal{S}$ can gradually evolve towards x_S^* and finally converge to a neighborhood of x_S^*.

We then introduce two assumptions on the signal noise ratio and the initialization, respectively. Specifically, the first assumption requires noise Γ not to overwhelm the ground truth low rank matrix Y^*. For notational simplicity, we denote $Y^* = x^*x^{*\top}$.

Assumption 1 (Signal-Noise-Ratio). *There exist some universal constants C_0, C_1 such that*

$$\begin{aligned} &\|x^*\|_2 \geq C_0, \quad \sigma \leq \frac{C_1}{d}, \ \|\Gamma_{\mathrm{sym}}\|_{\mathrm{F}} \leq 2d\sigma, \\ &\max\{\|\Gamma_{\mathrm{sym}}x^*\|_2, \|\Gamma_{\mathrm{sym}}\|_2\} \leq C_1\sqrt{d}\sigma, \end{aligned} \tag{6}$$

where

$$\Gamma_{\mathrm{sym}} = Y_{\mathrm{sym}} - Y^* = (Y + Y^\top)/2 - Y^* = (\Gamma + \Gamma^\top)/2.$$

The spectral and Frobenius norms of the noise Γ_{sym} are of order $\mathcal{O}(1/\sqrt{d})$ and $\mathcal{O}(1)$, respectively, while x^* is non-degenerate and yields a sufficiently large signal noise ratio.

Note that Li et al. (2019) show that 0 is a strict saddle point to (2). The second assumption requires the initialization of the P-GD algorithm to be sufficiently distant from 0.

Assumption 2 (Proper Initialization). *X_0 is bounded and sufficiently away from 0, i.e.,*

$$\|X_0\|_{\mathrm{F}}^2 \leq 1 - C_1\sqrt{d\sigma^2}, \|X_0^\top x^*\|_2^2 \geq C_1^2 d\sigma^2. \tag{7}$$

Assumption 2 can be further relaxed to an arbitrary initialization within a hyperball centered at 0. We do not consider such a relaxation, since it is not directly related to the regularization effect of the noise, but makes the convergence analysis much more involved.

Remark 3.1. *Note that both Assumptions 1 and 2 are deterministic. Later in Lemmas 3.7 and 3.8, we will show that both assumptions hold with high probability, given that Γ is sub-Gaussian and our initialization is random within a ball.*

We then present our main results in the following theorem.

Theorem 3.1 (Convergence Rate of P-GD). *Suppose that Assumptions 1 and 2 hold for Γ and X_0, respectively. For any $\delta \in (0,1)$, we choose*

$$\nu^2 = C_1\sqrt{d\sigma^2} \quad \text{and} \quad \eta \leq \eta_0 = \mathcal{O}\Big(\frac{\sigma^2}{d^2}\Big(\log\frac{1}{\delta}\Big)^{-1}\Big).$$

Then there exists some generic constant c_0 such that with probability at least $1-\delta$, we have

$$\frac{1}{d^2}\|X_tX_t^\top - Y^*\|_{\mathrm{F}}^2 \leq c_0\frac{\sigma^2}{d},$$

for all t's such that $\tau_0 \le t \le T = \mathcal{O}(\eta^{-2})$, *where* $\tau_0 = \mathcal{O}\Big(\frac{1}{\eta}\log\frac{1}{d\sigma^2}\log\frac{1}{\delta}\Big)$.

Theorem 3.1 implies that the noise plays an important role on regularizing the over-parameterized model during training, and induces a bias towards low complexity estimators. The estimation error is optimal for noisy rank one matrix factorization. For comparison, we can invoke the theoretical analyses in Jain et al. (2015) and show that GD does not have such a regularization effect and converges to a solution denoted by X_{GD}, where $X_{\mathrm{GD}}X_{\mathrm{GD}}^\top$ is essentially the positive semidefinite approximation of Y_{sym}. Therefore, X_{GD} only attains a suboptimal estimation error,

$$\frac{1}{d^2}\|X_{\mathrm{GD}}X_{\mathrm{GD}}^\top - Y^*\|_{\mathrm{F}}^2 = \mathcal{O}(\sigma^2).$$

As can be seen, P-GD outperforms GD in terms of mean square error for recovering the underlying low rank matrix Y^* by a factor of d.

The proof of Theorem 3.1 is very involved. Due to the space limit, we only present a proof sketch here. Please see more details in Appendix B.

Proof Sketch. Without loss of generality, we assume $\|x^*\|_2 = 1$. We start with a meta proof plan. Specifically, we decompose X_t into its projections in the subspace spanned by x^* and its orthogonal complement as follows:

$$X_t = \mathrm{Id}_{x^*}X_t + (\mathrm{Id} - \mathrm{Id}_{x^*})X_t = x^* r_t^\top + E_t, \tag{8}$$

where $r_t = X_t^\top x^*$. Note that the signal term $R_t = x^* r_t^\top$ always satisfies $R_t R_t^\top = \|r_t\|_2^2 Y^*$, which is a multiple of the ground truth matrix. Therefore, any solution satisfying $\|r_t\|_2^2 = 1$ and $E_t = 0$ gives the exact recovery. In light of this fact, we show that P-GD can find a solution such that $\|r_t\|_2$ is approximately 1 and E_t stays small. To facilitate our analysis, we write down the update of r_t and E_t as follows.

$$\begin{aligned} r_{t+1} &= X_{t+1}^\top x^* = r_t - \eta\nabla_X\mathcal{F}(X_t+W_t)^\top x^*, \\ E_{t+1} &= (\mathrm{Id} - \mathrm{Id}_{x^*})X_{t+1} \\ &= E_t - \eta(\mathrm{Id} - \mathrm{Id}_{x^*})\nabla_X\mathcal{F}(X_t+W_t). \end{aligned}$$

We further denote the gradient of $\mathcal{F}$ with respect to r and E as $\nabla_r\mathcal{F}(X) = \nabla_X\mathcal{F}(X)^\top x^*$ and $\nabla_E\mathcal{F}(X) = (\mathrm{Id} - \mathrm{Id}_{x^*})\nabla_X\mathcal{F}(X)$, where $r = X^\top x^*$, and $E = (\mathrm{Id} - \mathrm{Id}_{x^*})X$. The key of our proof is to show that $\nabla_r\mathcal{F}$ and $\nabla_E\mathcal{F}$ satisfy the subspace dissipative condition, which is stated in the next lemma.

Lemma 3.1 (Subspace Dissipativity). *For any* $X \in \mathbb{R}^{d\times d}$, $\nabla_E\mathcal{F}$ *satisfies*

$$\begin{aligned} &\langle \mathbb{E}_W[\nabla_E\mathcal{F}(X+W)], E\rangle \\ &\ge \left((2d+1)\frac{\nu^2}{d} - \|\Gamma_{\mathrm{sym}}\|_2\right)\|E\|_{\mathrm{F}}^2 - \frac{1}{4}\|\Gamma_{\mathrm{sym}}x^*\|_2^2. \end{aligned} \tag{9}$$

Let $a = 1-(2d+1)\frac{\nu^2}{d} + x^{*\top}\Gamma_{\mathrm{sym}}x^*$, *then* $\nabla_r\mathcal{F}$ *satisfies the inequality below if* $\|r\|_2^2 \ge a$.

$$\begin{aligned} &\langle \mathbb{E}_W[\nabla_r\mathcal{F}(X+W)], r\rangle \\ &\ge \|r\|_2^2(\|r\|_2^2 - a) - \frac{1}{4}\|\Gamma_{\mathrm{sym}}u_*\|_{\mathrm{F}}^2. \end{aligned} \tag{10}$$

Moreover, when $\|E\|_{\mathrm{F}}^2 \le c^2\|\Gamma_{\mathrm{sym}}x^*\|_2$, $\|\Gamma_{\mathrm{sym}}x^*\|_2^2 \le \|r\|_2^2 \le a$, *for some constant* $c > 0$, *then* $-\nabla_r\mathcal{F}$ *satisfies the following inequality.*

$$\begin{aligned} &\langle \mathbb{E}_W[-\nabla_r\mathcal{F}(X+W)], r\rangle \\ &\ge \|r\|_2^2(a - \|r\|_2^2) - (c^2+c)\|\Gamma_{\mathrm{sym}}u_*\|_{\mathrm{F}}^2. \end{aligned} \tag{11}$$

Lemma 3.1 helps describe the converge pattern of E_t and r_t. Specifically, the subspace dissipativity holds for $\nabla_E\mathcal{F}(X+W)$ globally, which implies that the orthogonal part E_t vanishes independent of r_t. The convergence of $\|r_t\|_2^2$, however, is more complicated. On the one hand, (10) suggests that when $\|r_t\|_2^2$ exceeds a, P-GD tends to decrease the norm $\|r_t\|_2^2$. On the other, when $\|r_t\|_2^2$ is small, (11) suggests $\|r_t\|_2^2$ will increase to a only after $\|E_t\|_{\mathrm{F}}^2$ is sufficiently small. Combining these two aspects, $\|r_t\|_2^2$ will move towards and stay close to $a \approx 1$. We remark that (11) requires $\|E_t\|_{\mathrm{F}}^2$ to be small. Therefore, the convergence of $\|r_t\|_2^2$ happens after that of $\|E_t\|_{\mathrm{F}}^2$.

Before showing the convergence, we provide a lemma showing that the trajectory of P-GD is bounded with high probability. This lemma helps us bound high order terms in the proof.

Lemma 3.2 (Boundedness of Trajectory). *Suppose* Γ *and* X_0 *satisfy Assumptions 1 and 2, respectively. For any* $\delta \in (0,1)$, *we choose* $\nu^2 = C_1\sqrt{d\sigma^2}$ *and*

$$\eta \le \eta_1 = \min\left\{\mathcal{O}\left(\frac{1}{d}\left(\log\frac{1}{\delta}\right)^{-1}\right), \mathcal{O}\left(\frac{1}{d^2}\right)\right\}.$$

Then with probability at least $1-\delta$, *for* $t \le T = \mathcal{O}(\frac{1}{\eta^2})$,

$$\|X_t\|_{\mathrm{F}}^2 \le 4d.$$

Following our discussions, we first show the convergence of $\|E_t\|_{\mathrm{F}}^2$ in the next lemma.

Lemma 3.3 (Convergence of E_t). *Suppose* Γ *and* X_0 *satisfy Assumptions 1 and 2, respectively. For any* $\delta \in (0,1)$, *we choose* $\nu^2 = C_1\sqrt{d\sigma^2}$ *and*

$$\eta \le \eta_2 = \min\left\{\mathcal{O}\left(\frac{\sigma}{d^3}\left(\log\frac{1}{\delta}\right)^{-1}\right), \mathcal{O}\left(\frac{\sigma^2}{d^2}\right)\right\},$$

then with probability at least $1-\delta$,

$$\|E_t\|_{\mathrm{F}}^2 \le \|E_0\|_{\mathrm{F}}^2 + c_1\sqrt{d\sigma^2} \tag{12}$$

holds for all t's such that $t \le T = \mathcal{O}(\eta^{-2})$, *and*

$$\|E_t\|_{\mathrm{F}}^2 \le c_1\sqrt{d\sigma^2} \tag{13}$$

holds for all t's such that $\tau_1 \le t \le T = O(\eta^{-2})$ where c_1 is an absolute constant and

$$\tau_1 = \mathcal{O}\Big(\frac{1}{\eta\sqrt{d\sigma^2}} \log \frac{1}{d\sigma^2} \log \frac{1}{\delta}\Big).$$

In addition to the convergence result (13), the boundedness of $\|E_t\|_{\mathrm{F}}^2$ in (12) will help us show that $\|r_t\|_2^2$ always stays away from the strict saddle point 0 as shown in the following lemma.

Lemma 3.4 (Avoid Strict Saddle)**.** *Suppose Γ and X_0 satisfy Assumptions 1 and 2, respectively. Assume (12) holds for all $t > 0$. For any $\delta \in (0,1)$, we choose $\nu^2 = C_1\sqrt{d\sigma^2}$ and*

$$\eta \le \eta_3 = \min\left\{\mathcal{O}\left(\sigma^4\left(\log\frac{1}{\delta}\right)^{-1}\right), \mathcal{O}\left(\frac{\sigma^2}{d^2}\right)\right\},$$

we then have with probability at least $1-\delta$, for all $t \le \mathcal{O}(1/\eta^2)$,

$$\|r_t\|_2^2 \ge \|\Gamma_{\mathrm{sym}} x^*\|_2^2. \tag{14}$$

Given that $\|E_t\|_{\mathrm{F}}^2$ becomes sufficiently small in Lemma 3.3, and r_t stays distant from zero, we can then invoke subspace dissipative condition (10) and (11) and show that $\|r_t\|_2^2$ will converge to 1 in the following lemma.

Lemma 3.5 (Convergence of r_t)**.** *Suppose Γ satisfies Assumption 1. Assume (13) and (14) hold for all $t > 0$. For any $\delta \in (0,1)$, we choose $\nu^2 = C_1\sqrt{d\sigma^2}$ and*

$$\eta \le \eta_4 = \min\left\{\mathcal{O}\left(\sigma^4\left(\log\frac{1}{\delta}\right)^{-1}\right), \mathcal{O}\left(\frac{\sigma^2}{d^2}\right)\right\}.$$

Then with probability at least $1-\delta$,

$$|\|r_t\|_2^2 - 1| \le c_2\sqrt{d\sigma^2} \tag{15}$$

for all t's such that $\tau_2 \le t \le T = O(\eta^{-2})$, where c_2 is an absolute constant and

$$\tau_2 = \mathcal{O}\Big(\frac{1}{\eta}\log\frac{1}{d\sigma^2}\log\frac{1}{\delta}\Big).$$

Note that the recovering error can be rewritten as follows.

$$\begin{aligned}\|X_tX_t^\top - Y^*\|_{\mathrm{F}}^2 &= (1-\|r_t\|_2^2)^2 + 2\|E_tr_t\|_2^2 + \|E_tE_t^\top\|_{\mathrm{F}}^2 \\ &\le (1-\|r_t\|_2^2)^2 + 2\|E_t\|_{\mathrm{F}}^2\|r_t\|_2^2 + \|E_t\|_{\mathrm{F}}^4.\end{aligned} \tag{16}$$

Combining (13) and (15), we know that P-GD has already entered and stays in the region with small recovery error. We remark that a naive treatment of the cross term $\|E_tr_t\|_2^2$ as in (16) will result in a recovery error $\mathcal{O}(\sqrt{d\sigma^2})$ that dominates (16), with a worse dependency on d. Instead, we take a more refined approach to bound the cross term by directly analyzing its optimization trajectory.

Lemma 3.6 (Convergence of E_tr_t)**.** *Suppose Γ satisfies Assumption 1. Assume (13) and (14) hold for all $t > 0$. For any $\delta \in (0,1)$, we choose $\nu^2 = C_1\sqrt{d\sigma^2}$ and*

$$\eta \le \eta_5 = \mathcal{O}\left(d\sigma^2\left(\log\frac{1}{\delta}\right)^{-1}\right).$$

Then with probability at least $1-\delta$,

$$\|E_tr_t\|_2^2 \le c_3 d\sigma^2$$

holds for all t's such that $\tau_3 \le t \le T = \mathcal{O}(\eta^{-2})$, where $c_3 > 0$ is an absolute constant and

$$\tau_3 = \mathcal{O}\Big(\frac{1}{\eta}\log\frac{1}{d\sigma^2}\log\frac{1}{\delta}\Big).$$

The proof of Lemmas 3.1–3.6 requires supermartingale-based analysis, which is very involved and technical. See more details in Section 3.1 and Appendix B.

Finally, using the conclusions of Lemmas 3.3, 3.5 and 3.6, we have with high probability that

$$\begin{aligned}\|X_tX_t^\top - Y^*\|_{\mathrm{F}}^2 &= (1-\|r_t\|_2^2)^2 + 2\|E_tr_t\|_2^2 + \|E_tE_t^\top\|_{\mathrm{F}}^2 \\ &\le (c_2^2 + 2c_2 + c_1^2)d\sigma^2\end{aligned}$$

holds when $\tau_1+\tau_2+\tau_3 \le t \le T$. Take $c_0 = c_1^2 + c_2^2 + 2c_3$, $\tau_0 = \tau_1+\tau_2+\tau_3$ and

$$\eta \le \eta_0 = \mathcal{O}\left(\frac{\sigma^2}{d^2}\left(\log\frac{1}{\delta}\right)^{-1}\right) \le \min\{\eta_1,\eta_2,\eta_3,\eta_4,\eta_5\},$$

where the last inequality holds since $\sigma = \mathcal{O}(1/d)$, and we prove that

$$\frac{1}{d^2}\|X_tX_t^\top - Y^*\|_{\mathrm{F}}^2 \le c_0\frac{\sigma^2}{d},$$

holds with high probability for all $\tau_0 \le t \le T$. □

Next, we verify that the noise matrix and the initialization of P-GD satisfy Assumptions 1 and 2, respectively. For noise matrix Γ with i.i.d sub-Gaussian entries, Assumption (1) holds with high probability by applying the standard concentration result as in the next lemma.

Lemma 3.7 (Signal-Noise-Ratio)**.** *For any $\delta \in (0,1)$, with high probability at least $1-\delta$, we have*

$$\begin{aligned}\max\{\|\Gamma_{\mathrm{sym}}x^*\|_2, \|\Gamma_{\mathrm{sym}}\|_2\} &\le C\sqrt{d}\sigma + C\sigma\sqrt{\log\frac{8}{\delta}}, \\ \|\Gamma_{\mathrm{sym}}\|_{\mathrm{F}} &\le d\sigma + \sigma\sqrt{2C\log\frac{8}{\delta}},\end{aligned} \tag{17}$$

where C is some absolute constant. Moreover, take $\delta = \mathcal{O}(\exp(-d))$ and we have

$$\max\{\|\Gamma_{\mathrm{sym}}x^*\|_2, \|\Gamma_{\mathrm{sym}}\|_2\} \le C_1\sqrt{d}\sigma, \quad \|\Gamma_{\mathrm{sym}}\|_{\mathrm{F}} \le 2d\sigma.$$

Furthermore, P-GD with random initialization in a unit ball satisfies Assumption 2 with high probability as shown in the next lemma.

Lemma 3.8 (Proper Initialization). *Given a random initialization $X_0 = x_0 x_0^\top$ where $x_0 \sim \mathrm{UNIF}(\mathbb{B}(1))$, then with probability at least $1 - \mathcal{O}\left(d^{-\frac{1}{4}}\right)$,*

$$\|X_0\|_{\mathrm{F}}^2 \le 1 - C_1\sqrt{d\sigma^2}, \|X_0^\top x^*\|_2^2 \ge C^2 d\sigma^2.$$

3.1 Super-martingale Theorem

In this section, we briefly introduce the key technique behind our analysis. We first provide a super-martingale based theorem frequently used in our ensuing analysis of perturbed gradient descent algorithm. Such a theorem can also be adapted to analyzing other stochastic recursive algorithm satisfying certain conditions, and hence can be of independent interest.

Theorem 3.2. *Given a random sequence $\{x_t\} \in \mathbb{R}^d$ satisfying $x_{t+1} = x_t - \eta f(x_t, \xi_t), \quad \forall t \ge 0$, where $x_0 \in \mathbb{R}^d$ is known, f is some bounded real valued function and $\xi_t \in \mathbb{R}^d$ represents the randomness in the update. Let g be some real valued function on $\mathbb{R}^d$. If there exist constants $\alpha_0 \in \mathbb{R}, \beta > 0, \lambda > 0, \phi > 0$ such that $(1-\eta\beta)^{-t}(g(x_t) - \alpha_0 - \eta\lambda)$ is a super-martingale for any $\eta > 0$ satisfying $\eta\beta < 1$, i.e.,*

$$\begin{aligned}&\mathbb{E}[(1-\eta\beta)^{-t-1}(g(x_{t+1}) - \alpha_0 - \eta\lambda)|\mathcal{F}_t] \\ &\qquad \le (1-\eta\beta)^{-t}(g(x_t) - \alpha_0 - \eta\lambda), \qquad (18)\end{aligned}$$

where $\mathcal{F}_t = \sigma\{x_\tau, \tau \le t\}$. Then for any $\delta \in [0,1], \alpha \ge \alpha_0$, we have the following conclusions.

Part I. *If we take $\eta \le \min\left\{\frac{\alpha_0}{2(\max f)\mathcal{I}_{g(x_0)>2\alpha}}, \frac{\alpha_0}{4\lambda}\right\}$. With probability at least $1-\delta$, there exists $t \le \tau'$, such that*

$$g(x_t) \le 2\alpha,$$

where

$$\tau' = \begin{cases} \mathcal{O}\left(\frac{1}{\eta\beta}\log\frac{4(g(x_0)-\frac{5}{4}\alpha_0)}{\alpha}\log\frac{1}{\delta}\right), & g(x_0) > 2\alpha; \\ 0, & \text{o.w.} \end{cases}$$

Part II. *Moreover, if we further have*

$$|g(x_{t+1}) - \mathbb{E}[g(x_{t+1})|\mathcal{F}_t]|\mathcal{I}_{\{g(x_t)\le 4\alpha\}} \le \phi\eta, \qquad (19)$$

and take

$$\eta \le \min\left\{\mathcal{O}\left(\frac{\alpha^2\beta}{\phi^2}\left(\log\frac{1}{\delta}\right)^{-1}\right), \frac{\alpha_0}{2(\max f)\mathcal{I}_{g(x_0)>2\alpha}}, \frac{\alpha_0}{4\lambda}\right\},$$

then with probability at least $1-\delta$,

$$g(x_t) \le 4\alpha,$$

for any $\tau' \le t \le T = \mathcal{O}(1/\eta^2)$.

Theorem 3.2 has two parts of results. Part I states that when the update satisfies (18), with properly chosen step size, the sequence $\{x_t\}$ can enter the region where g is bounded by a pre-specified constant α in polynomial time. Part II ensures that the sequence will stay in this region for long enough time. Please refer to Section A for the detailed proof.

To utilize Theorem 3.2 in analyzing our P-GD algorithm, we only need to check whether $g(x) = \|x - x^*\|_2^2$ meets the condition stated in (18). In fact, when the subspace dissipativity condition (5) is satisfied, we have

$$\begin{aligned}&\mathbb{E}\left[\|x_{t+1} - x^*\|_2^2|\mathcal{F}_t\right] \\ =&\|x_t - x^*\|_2^2 - 2\eta\mathbb{E}\left[\langle f(x_t,\xi_t), x_t - x_{\mathcal{S}}^*\rangle|\mathcal{F}_t\right] \\ &\qquad + \eta^2\mathbb{E}\left[\|f(x_t,\xi_t)\|_2^2|\mathcal{F}_t\right] \\ \le&(1-2\eta c_S)\|x_t - x^*\|_2^2 + 2\eta(\gamma_S + \mathcal{O}(\eta))\end{aligned}$$

where c_S, γ_S are the constants of subspace dissipativity conditions. By simple manipulation, the above inequality can be shown to be equivalent to the following inequality

$$\begin{aligned}&\mathbb{E}\left[(1-2\eta c_S)^{-t-1}\left(\|x_{t+1} - x^*\|_2^2 - \left(\frac{\gamma_S}{c_S} + \mathcal{O}(\eta)\right)\right)|\mathcal{F}_t\right] \\ &\quad \le (1-2\eta c_S)^{-t}\left(\|x_t - x^*\|_2^2 - \left(\frac{\gamma_S}{c_S} + \mathcal{O}(\eta)\right)\right),\end{aligned}$$

which is in the same form as in (18). Therefore, by exploiting subspace dissipativity in conjunction with our developed super-martingale theorem, we are poised to prove the key elements Lemma 3.1–3.6 in the analysis of P-GD.

4 NUMERICAL EXPERIMENTS

In this section, we demonstrate the regularization effect of noise using numerical experiments. Specifically, we compare P-GD algorithm with gradient descent (GD) with small and large initialization to show that noise induces a bias towards low complexity estimators.

4.1 Noisy Positive Semidefinite Matrix Recovery

We consider recovering a positive semidefinite (PSD) matrix $Y^* = X^* X^{*\top}$, with $X^* \in \mathbb{R}^{d\times r}$. We first present experiments for $r = 1$ to support our theory, then we conduct experiments on $r = 3$ and show that the general rank-r PSD matrix recovery exhibits similar behavior. **Rank-1.** We set the ground truth matrix $Y^* = x^* x^{*\top}$, where $x^* = (1, 1, \cdots, 1) \in \mathbb{R}^d$, and $d = 30$. The noise matrix Γ has i.i.d. Gaussian entries with mean 0 and variance $\sigma^2 = 0.1$. We run P-GD, GD with small and large initializations to solve (2). Specifically, GD-Small is initialized at $\frac{1}{d}A$, where A is a random orthogonal matrix as suggested in Li et al. (2017), while

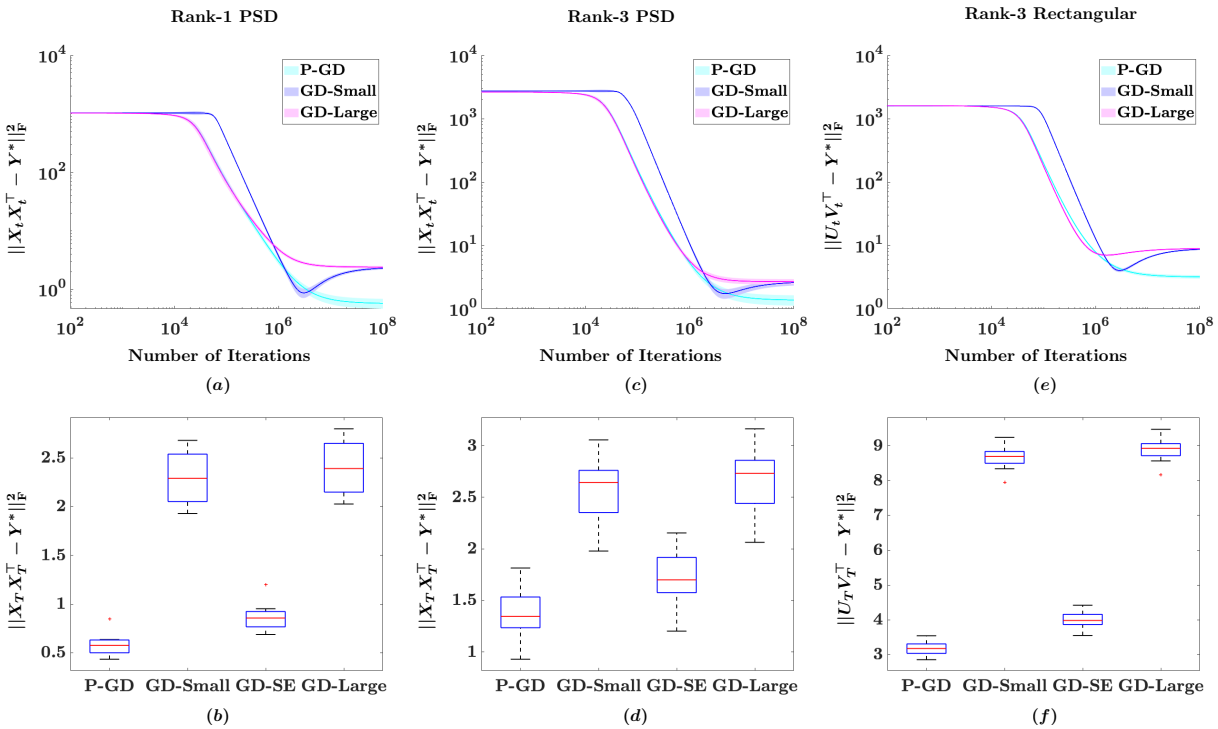

Figure 1: Average learning curves and final recovery error box plots of P-GD, GD with small initialization (GD-Small) and with large initialization (GD-Large). X,Y-axes are in log scale. The band in (a), (c), (e) represents standard deviation. (a)-(b): Rank-1 positive semidefinite matrix recovery. (c)-(d): Rank-3 positive semidefinite matrix recovery. (e)-(f): Rank-3 rectangular matrix recovery. GD-Small shows a regularization effect in the early stage but overfits later; P-GD performs the best. GD-Small with Early Stopping (GD-SE) achieves significantly better recovery error than other GD's, but still worse than P-GD.

GD-Large is initialized at $\frac{1}{\sqrt{d}}B$, where B is a random matrix with i.i.d. standard normal entries. P-GD takes the same initialization as GD-Large. In every iteration of P-GD , we perturb the iterate with Γ having i.i.d. Unif($\mathbb{S}(\nu)$) columns, where $\nu^2 = 0.4\sqrt{d\sigma^2}$. All three algorithms are run with $\eta = 0.25\sigma^2/d^2 = 2.7 \times 10^{-6}$ for $T = 1 \times 10^8$ iterations.

The results of 20 repeated runs are summarized in Figure 1.(a) and (b). The average learning curve in Figure 1.(a) shows that the convergence of GD with small initialization has two phases. Specifically, it first iterates towards the low complexity solutions and achieve a recovery error 0.9 in around 3×10^6 iterations, which is consistent with the algorithmic regularization effect of GD shown in Li et al. (2017). In the second stage, however, GD-small overfits the observational noise and finally attains a larger recovery error about 2, which is similar to GD-Large. In Figure 1.(b), we plot the final recovery error of the three algorithms. We also plot the minimal recovery error obtained by GD-Small and name it GD-Small with early stopping (GD-SE). It can be seen GD-SE avoids overfitting and can obtain a significantly lower recovery error than GD-Small. This observation justifies the regularization effect of early stopping in gradient descent learning. However, GD-SE still performs worse than P-GD. Different from GD, P-GD always converges to the estimators with lower recovery error around 0.5, even with large initialization. This suggests that noise induces implicit bias towards the low complexity solutions in training over-paramterized models.

Rank-3. We then consider rank-3 PSD matrix recovery. We choose set $Y^* = X^* X^{*\top}$, and $X^* \in \mathbb{R}^{d\times 3}$ with i.i.d. standard Gaussian entries. One can verify that Y^* is a rank-3 PSD matrix with probability 1. We choose other experiment settings the same as those of the rank-1 case except $\nu^2 = 0.25\sqrt{d\sigma^2}$. The results of 20 repeated runs are summarized in Figure 1.(c) and (d). We have similar observations as those in the rank-1 case from the learning curve and boxplot of GD and P-GD.

4.2 Noisy Rectangular Matrix Recovery

We perform experiments on rectangular matrix factorization, to show that the regularization effect of noise is not limited to symmetric matrix factorization problems. We set ground truth matrix $Y^* = U^* V^{*\top}$, where $U^* \in \mathbb{R}^{d\times 3}$ and $V^* \in \mathbb{R}^{d\times 3}$. We recover Y^*

by solving the following over-parameterized nonconvex optimization problem.

$$\left(\widehat{U}, \widehat{V}\right) = \underset{U\in\mathbb{R}^{d\times d}, V\in\mathbb{R}^{d\times d}}{\operatorname{argmin}} \frac{1}{2}\left\|UV^\top - Y\right\|_{\mathrm{F}}^2, \qquad (20)$$

where $Y = Y^* + \Gamma$ is a noisy observation of Y^*. P-GD solving (20) takes the following update:

$$\begin{aligned} U_{t+1} &= U_t - \eta\nabla_U \mathcal{F}(U_t + W_t, V_t + Z_t), \\ V_{t+1} &= V_t - \eta\nabla_V \mathcal{F}(U_t + W_t, V_t + Z_t), \end{aligned} \qquad (21)$$

where $\nabla_U \mathcal{F}(U,V) = (UV^\top - Y)V$, $\nabla_V \mathcal{F}(U,V) = (UV^\top - Y)^\top U$. Note that without perturbation, GD will converge exactly to an optimal solution such that $U_tV_t = Y$. In our experiment, we choose $d = 30$ and U^*, V^* to be two random rectangular matrices with i.i.d. standard Gaussian entries. The noise matrix Γ has i.i.d. Gaussian entries with mean 0 and variance $\sigma^2 = 0.1$. We run P-GD, GD-Small and GD-large to solve (20). GD-Small is initialized at $\frac{1}{d}(A_1, A_2)$, where A_1, A_2 are a random orthogonal matrix, while GD-Large is initialized at $\frac{1}{\sqrt{d}}(B_1, B_2)$, where B_1, B_2 is a random matrix with i.i.d. standard normal entries. P-GD takes the same initialization as GD-Large. We run P-GD with perturbation noise W_t, Z_t taking i.i.d. Unif($\mathbb{S}(\nu)$) columns, where $\nu^2 = 0.6\sqrt{d\sigma^2}$. All three algorithms are run with $\eta = 0.25\sigma^2/d^2 = 2.7\times 10^{-6}$ for $T = 1\times 10^8$ iterations.

The results of 20 repeated experiments are summarized in Figure 1.(e) and (f). We observe similar phenomenon of GD and P-GD as that in the rank-1 PSD matrix recovery, which advocates that the regularization effect of noise appears in general rectangular matrix recovery.

5 DISCUSSIONS

Extension to Rank-r Matrix Recovery: We can extend our theoretical analysis to rank-r PSD matrix recovery. Similar to the projection in (8), we project each iterate into the subspaces spanned by each eigenvectors of Y^*, and the orthogonal complement. The subspace dissipative conditions of each subspace can be obtained following similar lines to the proof of Lemma 3.1. We can then apply our super-martingale type analysis and show that P-GD can achieve the optimal convergence rate $\mathcal{O}(\frac{r\sigma^2}{d})$ for the rank-r case under some conditions on the eigen-value. The analysis, however, will be much more involved. We believe our results on the rank-1 case has already unveiled the regularization effect of noise and left technical extensions as our future work.

Extension to Rectangular Matrix Recovery: Our theoretical analysis can potentially extend to rectangular matrix recovery (20) by reducing the problem to symmetric PSD matrices as in Ge et al. (2017). Denote $W_t^\top = (U_t^\top, V_t^\top)$ and $W^{*\top} = (U^{*\top}, V^{*\top})$, where $Y^* = U^*V^{*\top}$. One can verify $N^* = W^*W^{*\top}$ is a PSD matrix. Recovering Y^* by P-GD (21) can be viewed as recovering N^* by applying P-GD on W_t. The problem is then reduced to rank-r PSD matrix recovery. To complete the analysis, we need theoretical guarantees on the equivalence of this reduction in our noisy observation case, which is left for future research.

Biased Stochastic Gradient Approximation: In our P-GD algorithm, the random perturbation to the iterates makes the gradient approximation biased. We remark that the biased stochastic gradient approximation also appears in training neural networks. Specifically, neural nets are often trained by SGD combining with many regularization techniques such as batch normalization (BN), weight decay, dropout and etc. These tricks help overcome overfitting. Meanwhile, since they essentially change the network structure or the loss function, the stochastic gradient in SGD becomes biased with respect to the original objective (Helmbold and Long, 2015, 2017; Mianjy et al., 2018; Luo et al., 2018). Such a biased approximation is worth further investigation to unveil their importance in learning over-parameterized models.

Regularization Effect: Our theoretical results provide new insights towards understanding the regularization effect of SGD in training deep neural networks. Specifically, besides the algorithmic regularization induced by deterministic first order algorithms (such as GD as shown in Li et al. (2017)), our theory suggests that noise also plays an important role in regularizing over-parameterized models.

Related Literature: Blanc et al. (2020) also study the implicit regularization for SGD type of algorithms based on a different problem setup, i.e., a 2-layer neural network without over-parameterization. They consider noise perturbation on labels instead of on parameters as in our paper and do not provide any explicit recovery error bound. Moreover, Blanc et al. (2020) consider regularizing the l_2 norm of gradient, which is equivalent to adding a regularizer defined by $||(XX^\top - Y)X||_F^2$ in our setting. To our best knowledge, there is no existing literature that shows this regularizer help find solutions with low complexity.

Some other papers study related problems but have fundamental differences with our work. HaoChen et al. (2020) consider perturbing labels while our work consider perturbing parameters. Du and Lee (2018) consider a problem without the underlying low complexity generating models, while we take advantage of an underlining low rank generating model and provide an estimation error bound analysis. Most importantly, we study the implicit regularization effect of noise without any explicit regularizers used in their work.

References

Allen-Zhu, Z., Li, Y. and Song, Z. (2018). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962* .

Arora, S., Du, S., Hu, W., Li, Z. and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*. PMLR.

Bahdanau, D., Cho, K. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* .

Blanc, G., Gupta, N., Valiant, G. and Valiant, P. (2020). Implicit regularization for deep neural networks driven by an ornstein-uhlenbeck like process. In *Conference on learning theory*. PMLR.

Du, S. and Lee, J. (2018). On the power of over-parametrization in neural networks with quadratic activation. In *International Conference on Machine Learning*. PMLR.

Du, S. S., Zhai, X., Poczos, B. and Singh, A. (2018). Gradient descent provably optimizes over-parameterized neural networks. In *International Conference on Learning Representations*.

Duchi, J. C., Bartlett, P. L. and Wainwright, M. J. (2012). Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization* **22** 674–701.

E, W., Ma, C. and Wu, L. (2020). A comparative analysis of optimization and generalization properties of two-layer neural network and random feature models under gradient descent dynamics. *Science China Mathematics* 1–24.

Ge, R., Jin, C. and Zheng, Y. (2017). No spurious local minima in nonconvex low rank problems: A unified geometric analysis. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014). Generative adversarial nets. In *Advances in neural information processing systems*.

Graves, A., Mohamed, A.-r. and Hinton, G. (2013). Speech recognition with deep recurrent neural networks. In *2013 IEEE international conference on acoustics, speech and signal processing*. IEEE.

HaoChen, J. Z., Wei, C., Lee, J. D. and Ma, T. (2020). Shape matters: Understanding the implicit bias of the noise covariance. *arXiv preprint arXiv:2006.08680* .

Helmbold, D. P. and Long, P. M. (2015). On the inductive bias of dropout. *The Journal of Machine Learning Research* **16** 3403–3454.

Helmbold, D. P. and Long, P. M. (2017). Surprising properties of dropout in deep networks. *The Journal of Machine Learning Research* **18** 7284–7311.

Jacot, A., Gabriel, F. and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*.

Jain, P., Jin, C., Kakade, S. M. and Netrapalli, P. (2015). Global convergence of non-convex gradient descent for computing matrix squareroot. *arXiv preprint arXiv:1507.05854* .

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M. and Jordan, M. I. (2017). How to escape saddle points efficiently. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*. JMLR. org.

Keskar, N. S., Mudigere, D., Nocedal, J., Smelyanskiy, M. and Tang, P. T. P. (2016). On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836* .

Krizhevsky, A., Sutskever, I. and Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*.

Li, X., Lu, J., Arora, R., Haupt, J., Liu, H., Wang, Z. and Zhao, T. (2019). Symmetry, saddle points, and global optimization landscape of non-convex matrix factorization. *IEEE Transactions on Information Theory* **65** 3489–3514.

Li, Y., Ma, T. and Zhang, H. (2017). Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. *arXiv preprint arXiv:1712.09203* .

Liang, S., Sun, R., Lee, J. D. and Srikant, R. (2018). Adding one neuron can eliminate all bad local minima. In *Advances in Neural Information Processing Systems*.

Liang, S., Sun, R. and Srikant, R. (2019). Revisiting landscape analysis in deep neural networks: Eliminating decreasing paths to infinity. *arXiv preprint arXiv:1912.13472* .

Long, J., Shelhamer, E. and Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Lu, S., Hong, M. and Wang, Z. (2019). Pa-gd: On the convergence of perturbed alternating gradient descent to second-order stationary points for structured nonconvex optimization. In *International Conference on Machine Learning*.

Luo, P., Wang, X., Shao, W. and Peng, Z. (2018). Towards understanding regularization in batch normalization. *arXiv preprint arXiv:1809.00846* .

Mianjy, P., Arora, R. and Vidal, R. (2018). On the implicit bias of dropout. In *International Conference on Machine Learning*. PMLR.

Sharifnassab, A., Salehkaleybar, S. and Golestani, S. J. (2020). Bounds on over-parameterization for guaranteed existence of descent paths in shallow relu networks. In *International Conference on Learning Representations*. URL https://openreview.net/forum?id=BkgXHTNtvS

Vershynin, R. (2018). *High-dimensional probability: An introduction with applications in data science*, vol. 47. Cambridge university press.

Young, T., Hazarika, D., Poria, S. and Cambria, E. (2018). Recent trends in deep learning based natural language processing. *ieee Computational intelligenCe magazine* **13** 55–75.

Yu, D. and Deng, L. (2010). Deep learning and its applications to signal and information processing [exploratory dsp]. *IEEE Signal Processing Magazine* **28** 145–154.

Zhou, M., Liu, T., Li, Y., Lin, D., Zhou, E. and Zhao, T. (2019). Towards understanding the importance of noise in training neural networks. *arXiv preprint arXiv:1909.03172* .

A PROOF OF THEOREM 3.2

Proof. **Part I.** We first show that with probability at least $1-\delta$, there exists $t \le \tau_0$ such that $g(x_t) \le 2\alpha$. We only need to consider the case where $g(x_0) > 2\alpha_0$. Let $\mathcal{E}_t = \{g(x_\tau) \ge 2\alpha, \forall \tau \le t\}, G_t = (1-\eta\beta)^{-t}(g(x_t) - \alpha_0 - \eta\lambda)$. Then by (18), we have

$$\mathbb{E}[G_{t+1}\mathcal{I}_{\mathcal{E}_t}\mathcal{F}_t] \le \mathbb{E}[G_t\mathcal{I}_{\mathcal{E}_t}] \le \mathbb{E}[G_t\mathcal{I}_{\mathcal{E}_{t-1}}].$$

The last inequality holds since when $\mathcal{I}_{\mathcal{E}_{t-1}} = 1$ while $\mathcal{I}_{\mathcal{E}_t} = 0$, $G_t \ge 0$ when $\eta \le \min\{\frac{\alpha_0}{2\max f}, \frac{\alpha_0}{4\lambda}\}$. Then $\{G_t\mathcal{I}_{\mathcal{E}_{t-1}}\}$ are a supermartingale sequence. Then

$$\mathbb{P}(\mathcal{E}_t) \le \mathbb{P}(g(x_t) \ge 2\alpha) \le \frac{\mathbb{E}[g(x_t)]}{2\alpha} \le \frac{(1-\eta\beta)^t(g(x_0) - \alpha_0 - \eta\lambda) + \alpha_0 + \eta\lambda}{2\alpha} \le \frac{3}{4},$$

when $t \ge \frac{1}{\eta\beta}\log\frac{4(g(x_0)-\frac{5}{4}\alpha_0)}{\alpha}$. Recursively applying the above lines for $\mathcal{O}(\log\frac{1}{\delta})$ times, we know there exists $t \le \tau$ such that $g(x_t) \le 2\alpha$ with probability at least $1-\delta$, where

$$\tau_0 = \mathcal{O}\left(\frac{1}{\eta\beta}\log\frac{4(g(x_0)-\frac{5}{4}\alpha_0)}{\alpha}\log\frac{1}{\delta}\right).$$

Part II. Then we show, with high probability, for any $\alpha \ge \alpha_0$, $g(x_t) \le 4\alpha$ for long enough time. Let $\mathcal{H}_t = \{g(x_\tau) \le 4\alpha, \forall \tau \le t\}$. By (18), we have

$$\mathbb{E}[G_{t+1}\mathcal{I}_{\mathcal{H}_t}\mathcal{F}_t] \le \mathbb{E}[G_t\mathcal{I}_{\mathcal{H}_t}] \le \mathbb{E}[G_t\mathcal{I}_{\mathcal{H}_{t-1}}]. \tag{22}$$

Then $\{G_t\mathcal{I}_{\mathcal{E}_{t-1}}\}$ are a super-martingale sequence. We then bound the difference between $G_t\mathcal{I}_{\mathcal{E}_{t-1}}$ and $\mathbb{E}[G_t\mathcal{I}_{\mathcal{E}_{t-1}}|\mathcal{F}_t]$.

$$d_t = |G_t\mathcal{I}_{\mathcal{E}_{t-1}} - \mathbb{E}[G_t\mathcal{I}_{\mathcal{E}_{t-1}}|\mathcal{F}_t]| = (1-\eta\beta)^{-t}|g(x_{t+1}) - \mathbb{E}[g(x_{t+1})|\mathcal{F}_t])| \le (1-\eta\beta)^{-t}\phi\eta. \tag{23}$$

Denote $D_t = \sqrt{\sum_{i=0}^t d_i^2}$. By Azuma's Inequality, we get

$$\mathbb{P}\left(G_t\mathcal{I}_{\mathcal{H}_{t-1}} - G_0 \ge \mathcal{O}(1)\, D_t \log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right)\right) \le \exp\left(-\frac{\mathcal{O}(1)\, D_t^2 \log\left(\frac{1}{\eta^2\delta}\right)}{2\sum_{i=0}^t d_i^2}\right) = \mathcal{O}\left(\eta^2\delta\right).$$

Therefore, with at least probability $1-\mathcal{O}\left(\eta^2\delta\right)$, we have

$$\begin{aligned} g(x_t) &\le (1-\eta\beta)^t\left(g(x_0) - \alpha_0 - \eta\lambda\right) + \mathcal{O}(1)\,(1-\eta\beta)^t D_t \log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right) + \alpha_0 + \eta\lambda \\ &\le g(x_0) + \mathcal{O}(1)\,\frac{\phi\eta}{\sqrt{\eta\beta}}\log^{\frac{1}{2}}\left(\frac{1}{\eta^2\delta}\right) + \frac{5}{4}\alpha \le 4\alpha, \end{aligned}$$

where the last line holds, since we can always find $\eta = \min\left\{\mathcal{O}\left(\frac{\alpha^2\beta}{\phi^2}\left(\log\frac{1}{\delta}\right)^{-1}\right), \frac{\alpha_0}{4\lambda}\right\}$ to satisfy the condition.

The above inequality shows that if $\mathcal{H}_t$ holds, then $\mathcal{H}_{t+1}$ holds with at least probability $1-\mathcal{O}\left(\eta^2\delta\right)$. Hence, with at least probability $1-\delta$, we have $g(x_t) \le 4\alpha$ for all $t \le T = \mathcal{O}\left(\frac{1}{\eta^2}\right)$.

Combine Part I and Part 2, properly rescale δ, and we prove the theorem. □

B PROOF OF TECHNICAL LEMMAS

B.1 Proof of Lemma 3.1

Proof. For notational simplicity, we denote $\gamma = \frac{\nu}{\sqrt{d}}$. We start from the subspace dissipative condition for E_t. We will use the fact $Y^*E = 0$.

$$\begin{aligned}
&\langle E, \mathbb{E}_W(\mathrm{Id} - \mathrm{Id}_{x^*})\nabla_X \mathcal{F}(X+W)\rangle \\
&= \left\langle E, (\mathrm{Id} - \mathrm{Id}_S)\left((XX^\top - Y)X + (2d+1)\gamma^2 X\right)\right\rangle \\
&= \left\langle E, (2d+1)\gamma^2 E + (\mathrm{Id} - \mathrm{Id}_S)\left((XX^\top - Y)X\right)\right\rangle \\
&= (2d+1)\gamma^2\|E\|_{\mathrm{F}}^2 + \left\langle XX^\top - Y, (\mathrm{Id} - \mathrm{Id}_S)EX^\top\right\rangle \\
&= (2d+1)\gamma^2\|E\|_{\mathrm{F}}^2 + \left\langle XX^\top - Y, EX^\top\right\rangle \\
&= (2d+1)\gamma^2\|E\|_{\mathrm{F}}^2 + \left\langle XX^\top - Y^*, EX^\top\right\rangle - \left\langle \Gamma_{\mathrm{sym}}, EX^\top\right\rangle \\
&= (2d+1)\gamma^2\|E\|_{\mathrm{F}}^2 + \left\langle XX^\top, EX^\top\right\rangle - \left\langle \Gamma_{\mathrm{sym}}, EX^\top\right\rangle \\
&= (2d+1)\gamma^2\|E\|_{\mathrm{F}}^2 + \|EX^\top\|_{\mathrm{F}}^2 - \left\langle \Gamma_{\mathrm{sym}}, EX^\top\right\rangle \\
&= (2d+1)\gamma^2\|E\|_{\mathrm{F}}^2 + \|EE^\top\|_{\mathrm{F}}^2 + \|EZ^\top\|_{\mathrm{F}}^2 - \left\langle \Gamma_{\mathrm{sym}}, EE^\top\right\rangle - \left\langle \Gamma_{\mathrm{sym}}, EZ^\top\right\rangle \\
&\geq ((2d+1)\gamma^2 - \|\Gamma_{\mathrm{sym}}\|_2)\|E\|_{\mathrm{F}}^2 + \|EZ^\top\|_{\mathrm{F}}^2 - \|\Gamma_{\mathrm{sym}}\|_2\|EZ^\top\|_2 \\
&\geq ((2d+1)\gamma^2 - \|\Gamma_{\mathrm{sym}}\|_2)\|E\|_{\mathrm{F}}^2 - \frac{1}{4}\|\Gamma_{\mathrm{sym}}\|_2^2.
\end{aligned}$$

The last inequality holds since given $b > 0$, $x^2 - bx \geq -\frac{b^2}{4}$ for $\forall x > 0$. Then we obtain the inequality (9).

We next prove the subspace dissipative condition for r.

$$\begin{aligned}
&\left\langle r_t, \mathbb{E}\left[\nabla_X \mathcal{F}(X_t + W_t)^\top x^*\right]\right\rangle \\
&= \left\langle r_t, X_t^\top(X_tX_t^\top - Y_{\mathrm{sym}})x^* + (2d+1)\gamma^2 X_t^\top x^*\right\rangle \\
&= (2d+1)\gamma^2\|r_t\|_2^2 + \left\langle r_t, X_t^\top(X_tX_t^\top - Y^*)x^*\right\rangle - \left\langle r_t, X_t^\top\Gamma_{\mathrm{sym}}x^*\right\rangle. \qquad (24)
\end{aligned}$$

We calculate the last two terms separately. Note that $X_t^\top(X_tX_t^\top - Y^*)x^* = X_t^\top X_t r_t - r_t = (\|r_t\|_2^2 - 1)r_t + E_t^\top E_t r_t$. Insert this equation in the second term in (24) and we have

$$\left\langle r_t, X_t^\top(X_tX_t^\top - Y^*)x^*\right\rangle = (\|r_t\|_2^2 - 1)\|r_t\|_2^2 + \|E_t r_t\|_2^2.$$

Moreover, the last term in(24) can be calculated as follows.

$$\begin{aligned}
\left\langle r_t, X_t^\top\Gamma_{\mathrm{sym}}x^*\right\rangle &= r_t^\top X_t^\top\Gamma_{\mathrm{sym}}x^* \\
&= r_t^\top(r_t x^{*\top} + E_t^\top)\Gamma_{\mathrm{sym}}x^* \\
&= \|r_t\|_2^2 x^{*\top}\Gamma_{\mathrm{sym}}x^* + r_t^\top E_t^\top\Gamma_{\mathrm{sym}}x^*.
\end{aligned}$$

Let $a = 1 - (2d+1)\gamma^2 + x^{*\top}\Gamma_{\mathrm{sym}}x^*$. Then we have

$$\begin{aligned}
\left\langle r_t, \mathbb{E}\left[\nabla_X \mathcal{F}(X_t + W_t)^\top x^*\right]\right\rangle &= (2d+1)\gamma^2\|r_t\|_2^2 + (\|r_t\|_2^2 - 1)\|r_t\|_2^2 \\
&\quad + \|E_t r_t\|_2^2 - \|r_t\|_2^2 x^{*\top}\Gamma_{\mathrm{sym}}x^* - r_t^\top E_t^\top\Gamma_{\mathrm{sym}}x^* \\
&= \left(\|r_t\|_2^2 - 1 + (2d+1)\gamma^2 - x^{*\top}\Gamma_{\mathrm{sym}}x^*\right)\|r_t\|_2^2 + \|E_t r_t\|_2^2 - r_t^\top E_t^\top\Gamma_{\mathrm{sym}}x^* \\
&\geq (\|r_t\|_2^2 - a)\|r_t\|_2^2 + \|E_t r_t\|_2^2 - \|E_t r_t\|_2\|\Gamma_{\mathrm{sym}}x^*\|_2 \\
&\geq (\|r_t\|_2^2 - a)\|r_t\|_2^2 - \frac{1}{4}\|\Gamma_{\mathrm{sym}}x^*\|_2^2.
\end{aligned}$$

This proves the inequality (10). On the other hand,

$$\begin{aligned}
\left\langle r_t, -\mathbb{E}\left[\nabla_X \mathcal{F}(X_t + W_t)^\top x^*\right]\right\rangle &= -(2d+1)\gamma^2\|r_t\|_2^2 - (\|r_t\|_2^2 - 1)\|r_t\|_2^2 \\
&\quad - \|E_t r_t\|_2^2 + \|r_t\|_2^2 x^{*\top}\Gamma_{\mathrm{sym}}x^* + r_t^\top E_t^\top\Gamma_{\mathrm{sym}}x^* \\
&= \left(a - \|r_t\|_2^2\right)\|r_t\|_2^2 - \|E_t r_t\|_2^2 + r_t^\top E_t^\top\Gamma_{\mathrm{sym}}x^* \\
&\geq (a - \|r_t\|_2^2)\|r_t\|_2^2 - \|E_t r_t\|_2^2 - \|E_t r_t\|_2\|\Gamma_{\mathrm{sym}}x^*\|_2 \\
&\geq (a - \|r_t\|_2^2)\|r_t\|_2^2 - (c^2 + c)\|\Gamma_{\mathrm{sym}}x^*\|_2^2.
\end{aligned}$$

We prove the inequality (11). □

B.2 Proof of Lemma 3.2

Proof. Given our choice of ν, we have with probability at least $1-\delta$, $d\gamma^2 \geq \|\Gamma_{\text{sym}}\|_2$. Given our initialization, we have $\|X_0\|_{\text{F}}^2 = 1 \leq 4d$.

$$\begin{aligned}
\mathbb{E}[\|X_{t+1}\|_{\text{F}}^2|\mathcal{F}_t] &= \mathbb{E}\left[\|X_t - \eta\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2|\mathcal{F}_t\right] \\
&= \|X_t\|_{\text{F}}^2 - 2\eta\mathbb{E}[\langle X_t, \nabla_X\mathcal{F}(X_t+W_t)\rangle|\mathcal{F}_t] + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2|\mathcal{F}_t\right] \\
&= \|X_t\|_{\text{F}}^2 - 2\eta\left\langle X_t, \mathbb{E}[\nabla_X\mathcal{F}(X_t+W_t)|\mathcal{F}_t]\right\rangle + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2|\mathcal{F}_t\right] \\
&= \|X_t\|_{\text{F}}^2 - 2\eta\left\langle X_t, \nabla_X\mathcal{F}(X_t) + (2d+1)\gamma^2X_t\right\rangle + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2|\mathcal{F}_t\right] \\
&= \|X_t\|_{\text{F}}^2 - 2\eta\left\langle X_t, (X_tX_t^\top - Y^*)X_t - \Gamma_{\text{sym}}X_t + (2d+1)\gamma^2X_t\right\rangle \\
&\qquad + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2|\mathcal{F}_t\right] \\
&= (1-2\eta(2d+1)\gamma^2)\|X_t\|_{\text{F}}^2 - 2\eta\|X_t^\top X_t\|_{\text{F}}^2 + 2\eta\langle X_t, Y^*X_t\rangle \\
&\qquad + 2\eta\langle X_t, \Gamma_{\text{sym}}X_t\rangle + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2|\mathcal{F}_t\right].
\end{aligned}$$

Note that

$$\langle X_t, Y^*X_t\rangle = \text{tr}\big(X_t^\top Y^*X_t\big) \leq \|X_t\|_{\text{F}}^2,$$

$$\|X_t^\top X_t\|_{\text{F}}^2 = \sum_{i,j}(X_t^\top X_t)_{i,j}^2 \geq \sum_i (X_t^\top X_t)_{i,i}^2 \geq \frac{1}{d}\left(\sum_i (X_t^\top X_t)_{i,i}\right)^2 = \frac{1}{d}\|X_t\|_{\text{F}}^4,$$

and

$$\langle X_t, \Gamma_{\text{sym}}X_t\rangle \leq \|\Gamma_{\text{sym}}\|_2\|X_t\|_{\text{F}}^2.$$

Then we have

$$\mathbb{E}[\|X_{t+1}\|_{\text{F}}^2|\mathcal{F}_t] \leq \big(1-2\eta(2d+1)\gamma^2 + 2\eta(1+\|\Gamma_{\text{sym}}\|_2)\big)\|X_t\|_2^2 - 2\eta\frac{1}{d}\|X_t\|_{\text{F}}^4 + \eta^2\mathbb{E}[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2].$$

or equivalently,

$$\begin{aligned}
&\mathbb{E}[\|X_{t+1}\|_{\text{F}}^2 - d|\mathcal{F}_t] \\
\leq\ &\big(1-2\eta(2d+1)\gamma^2 + 2\eta\|\Gamma_{\text{sym}}\|_2\big)\left(\|X_t\|_{\text{F}}^2 - d\right) - 2\eta\frac{1}{d}\|X_t\|_{\text{F}}^2(\|X_t\|_{\text{F}}^2 - d) \\
&\quad - 2\eta d((2d+1)\gamma^2 - \|\Gamma_{\text{sym}}\|_2) + \eta^2\mathbb{E}[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2] \\
\leq\ &\left(1-2\eta(2d+1)\gamma^2 + 2\eta\|\Gamma_{\text{sym}}\|_2 - 2\eta\frac{1}{d}\|X_t\|_{\text{F}}^2\right)(\|X_t\|_{\text{F}}^2 - d) \\
&\quad - 2\eta d\left((2d+1)\gamma^2 - \|\Gamma_{\text{sym}}\|_2\right) + \eta^2\mathbb{E}[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2] \\
\leq\ &\left(1-2\eta(2d+1)\gamma^2 + 2\eta\|\Gamma_{\text{sym}}\|_2 - 2\eta\frac{1}{d}\|X_t\|_{\text{F}}^2\right)(\|X_t\|_{\text{F}}^2 - d) + \eta^2\mathbb{E}[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2] \\
\leq\ &\big(1-2\eta(2d+1)\gamma^2 + 2\eta\|\Gamma_{\text{sym}}\|_2 - 2\eta\big)(\|X_t\|_{\text{F}}^2 - d) + \eta^2\mathbb{E}[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2].
\end{aligned}$$

Let $\mathcal{E}_t$ be the event $\left\{\|X_\tau\|_{\text{F}}^2 \leq 4d, \forall\tau\leq t\right\}$. Then

$$\begin{aligned}
\mathbb{E}\left[(\|X_{t+1}\|_{\text{F}}^2 - d)\mathcal{I}_{\mathcal{E}_t}|\mathcal{F}_t\right] &\leq \big(1-2\eta(2d+1)\gamma^2 + 2\eta\|\Gamma_{\text{sym}}\|_2 - 2\eta\big)(\|X_t\|_{\text{F}}^2 - d)\mathcal{I}_{\mathcal{E}_t} + \\
&\qquad \eta^2\mathbb{E}[\|\nabla_X\mathcal{F}(X_t+W_t)\|_{\text{F}}^2]\mathcal{I}_{\mathcal{E}_t} \\
&\leq \big(1-2\eta(2d+1)\gamma^2 + 2\eta\|\Gamma_{\text{sym}}\|_2 - 2\eta\big)(\|X_t\|_{\text{F}}^2 - d)\mathcal{I}_{\mathcal{E}_t} + \eta^2C_1\mathcal{I}_{\mathcal{E}_t}.
\end{aligned}$$

where $C_1 = \mathcal{O}(d^3)$. Let $\lambda_1 = \frac{C_1}{2(2d+1)\gamma^2 - 2\|\Gamma_{\text{sym}}\|_2 + 2} = \mathcal{O}(d^3), \beta_1 = 2(2d+1)\gamma^2 - 2\|\Gamma_{\text{sym}}\|_2 + 2$. Equivalently, we have the following inequality.

$$\mathbb{E}\left[(\|X_{t+1}\|_{\text{F}}^2 - d - \eta\lambda_1)\mathcal{I}_{\mathcal{E}_t}|\mathcal{F}_t\right] \leq (1-\beta_1)\left(\|X_t\|_{\text{F}}^2 - d - \eta\lambda_1\right)\mathcal{I}_{\mathcal{E}_t}.$$

We further denote $G_t = (1-\beta_1)^{-t}(\|X_t\|_{\text{F}}^2 - d - \eta\lambda_1)$. Then we have

$$\mathbb{E}[G_{t+1}\mathcal{I}_{\mathcal{E}_t}] \leq \mathbb{E}[G_t\mathcal{I}_{\mathcal{E}_t}] \leq \mathbb{E}[G_t\mathcal{I}_{\mathcal{E}_{t-1}}].$$

Then (22) is satisfied. We then bound the difference between $G_t\mathcal{I}_{\mathcal{E}_{t-1}}$ and the conditional expectation $\mathbb{E}[G_t\mathcal{I}_{\mathcal{E}_{t-1}}|\mathcal{F}_{t-1}]$.

$$\begin{aligned}
d_t &= \left|G_t\mathcal{I}_{\mathcal{E}_{t-1}} - \mathbb{E}[G_t\mathcal{I}_{\mathcal{E}_{t-1}}|\mathcal{F}_{t-1}]\right| \\
&= (1-\beta_1)^{-t}\Big|2\eta\left(\mathbb{E}[\langle X_{t-1}, \nabla_X\mathcal{F}(X_{t-1}+W_{t-1})\rangle\,|\mathcal{F}_{t-1}] - \langle X_{t-1}, \nabla_X\mathcal{F}(X_{t-1}+W_{t-1})\rangle\right) \\
&\quad - \eta^2\left(\mathbb{E}[\|\nabla_X\mathcal{F}(X_{t-1}+W_{t-1})\|_{\mathrm{F}}^2|\mathcal{F}_{t-1}] - \|\nabla_X\mathcal{F}(X_{t-1}+W_{t-1})\|_{\mathrm{F}}^2\right)\Big| \\
&\le (1-\beta_1)^{-t}\Big[2\eta\big((2d+1)\gamma^2\|X_{t-1}\|_{\mathrm{F}}^2 + 3\|X_{t-1}\|_{\mathrm{F}}^3\|W_{t-1}\|_{\mathrm{F}} + 3\|X_{t-1}\|_{\mathrm{F}}^2\|W_{t-1}\|_{\mathrm{F}}^2 + \|X_{t-1}\|_{\mathrm{F}}\|W_{t-1}\|_{\mathrm{F}}^3 \\
&\qquad + \|X_{t-1}\|_{\mathrm{F}}\|X_{t-1}\|_{\mathrm{F}}\|W_{t-1}\|_{\mathrm{F}}\big) + 2\eta^2 C_1\Big] \\
&\le (1-\beta_1)^{-t}\eta\phi_1,
\end{aligned}$$

where $\phi_1 = \mathcal{O}(d^{1.5})$. Then (23) is satisfied. Directly applying Part II of Theorem 3.2, we can get the result. Specifically, we choose

$$\begin{aligned}
\eta &= \min\left\{\mathcal{O}\left(\frac{d^2}{d^3}\left(\log\frac{1}{\delta}\right)^{-1}\right), \mathcal{O}\left(\frac{d}{d^3}\right)\right\} \\
&= \mathcal{O}\left(\frac{1}{d^2}\left(\log\frac{1}{\delta}\right)^{-1}\right).
\end{aligned}$$

With at least probability $1-\delta$, we have $\|X_t\|_{\mathrm{F}}^2 \le 4d$ for all $t \le \mathcal{O}\left(\frac{1}{\eta^2}\right)$.

□

B.3 Proof of Lemma 3.3

Proof. Let $\mathcal{F}_t = \sigma\{X_\tau, \tau \le t\}$ be the $\sigma-$field generated by past t iterations. We first calculate the conditional expectation of $\|X_{t+1}\|_{\mathrm{F}}^2$ given $\mathcal{F}_t$. Note that the update of E_t can be written as follows:

$$\begin{aligned}
E_{t+1} = (\mathrm{Id} - \mathrm{Id}_S)X_{t+1} &= (\mathrm{Id} - \mathrm{Id}_S)\left(X_t - \eta\nabla_X\mathcal{F}(X_t+W_t)\right) \\
&= E_t - \eta(\mathrm{Id} - \mathrm{Id}_S)\nabla_X\mathcal{F}(X_t+W_t)
\end{aligned} \tag{25}$$

Then we calculate the conditional expectation of $\|E_{t+1}\|_{\mathrm{F}}^2$.

$$\begin{aligned}
\mathbb{E}[\|E_{t+1}\|_{\mathrm{F}}^2|\mathcal{F}_t] = \|E_t\|_{\mathrm{F}}^2 &- 2\eta\mathbb{E}[\langle E_t, (\mathrm{Id} - \mathrm{Id}_{x^*})\nabla_X\mathcal{F}(X_t+W_t)\rangle\,|\mathcal{F}_t] \\
&+ \eta^2\mathbb{E}[\|(\mathrm{Id} - \mathrm{Id}_{x^*})\nabla_X\mathcal{F}(X_t+W_t)\|_{\mathrm{F}}^2|\mathcal{F}_t].
\end{aligned}$$

Applying the subspace dissipative condition (9), we get the following inequality.

$$\begin{aligned}
\mathbb{E}[\|E_{t+1}\|_{\mathrm{F}}^2|\mathcal{F}_t] \le &(1 - 2\eta((2d+1)\gamma^2 - \|\Gamma_{\mathrm{sym}}\|_2))\|E_t\|_{\mathrm{F}}^2 + \eta\frac{\|\Gamma_{\mathrm{sym}}x^*\|_2^2}{2} \\
&+ \eta^2\mathbb{E}[\|(\mathrm{Id} - \mathrm{Id}_{x^*})\nabla_X\mathcal{F}(X_t+W_t)\|_{\mathrm{F}}^2|\mathcal{F}_t].
\end{aligned}$$

Since both X_t and W_t are bounded, we can verify $\mathbb{E}[\|(\mathrm{Id} - \mathrm{Id}_{x^*})\nabla_X\mathcal{F}(X_t+W_t)\|_{\mathrm{F}}^2|\mathcal{F}_t] \le Cd^3$. Let $\beta_2 = 2((2d+1)\gamma^2 - \|\Gamma_{\mathrm{sym}}\|_2)$, $\alpha_2 = \frac{\|\Gamma_{\mathrm{sym}}x^*\|_2^2}{4((2d+1)\gamma^2 - \|\Gamma_{\mathrm{sym}}\|_2)}$ and $\lambda_2 = C\frac{d^3}{2((2d+1)\gamma^2 - \|\Gamma_{\mathrm{sym}}\|_2)}$, then we have

$$\mathbb{E}[(\|E_{t+1}\|_{\mathrm{F}}^2 - \alpha_2 - \eta\lambda)|\mathcal{F}_t] \le (1-\eta\beta_2)\left(\|E_t\|_{\mathrm{F}}^2 - \alpha_2 - \eta\lambda\right).$$

Then (18) holds for $\|E_t\|_{\mathrm{F}}^2$. Moreover, based on the boundedness proved in Lemma 3.2, one can easily verify.

$$\left|\|E_{t+1}\|_{\mathrm{F}}^2 - \mathbb{E}[\|E_{t+1}\|_{\mathrm{F}}^2|\mathcal{F}_t]\right|\mathcal{I}_{\{\|E_t\|_{\mathrm{F}}^2 \le 4\alpha_2\}} \le \eta C_2 d^{2.25}\sigma^{1.5}.$$

Thus, if we take $\phi_2 = C_2 d^{2.25}\sigma^{1.5}$, (19) holds. By Theorem 3.2, if we take

$$\eta = \min\left\{\mathcal{O}\left(\frac{\sigma}{d^3}\left(\log\frac{1}{\delta}\right)^{-1}\right), \mathcal{O}\left(\frac{\sigma^2}{d^2}\right)\right\},$$

we then have with probability at least $1-\delta$,

- $\|E_t\|_{\mathrm{F}}^2 \le \|E_0\|_{\mathrm{F}}^2 + c_1\sqrt{d\sigma^2} \le 1$, for all $t's$ such that $t \le T = \mathcal{O}(1/\eta^2)$;
- $\|E_t\|_{\mathrm{F}}^2 \le c_1\sqrt{d\sigma^2}$, for all $t's$ such that $\tau_1 \le t \le T = \mathcal{O}(1/\eta^2)$, where c_1 is a constant and

$$\tau_1 = \mathcal{O}\Big(\frac{1}{\eta\sqrt{d\sigma^2}} \log \frac{\|E_0\|_{\mathrm{F}}^2}{\sqrt{d\sigma^2}} \log \frac{1}{\delta}\Big).$$

□

B.4 Proof of Lemma 3.4

Proof. With our initialization, we have $\|r_0\|_2^2 \le a$. Then we prove $\|r_0\|_2^2 \le a + \mathcal{O}(\sqrt{d\sigma^2})$ for long enough time.

Recall that $a = 1 - (2d+1)\gamma^2 + x^{*\top}\Gamma_{\mathrm{sym}}x^*$. By (10), the subspace dissipative condition of r_t, we can upper bound the conditional expectation of $\|r_{t+1}\|_2^2 - a$ given the trajectory history.

$$\begin{aligned}
\mathbb{E}[\|r_{t+1}\|_2^2 - a\big|\mathcal{F}_t] &= (\|r_t\|_2^2 - a) - 2\eta\mathbb{E}\left[\left\langle r_t, \nabla_X\mathcal{F}(X_t+W_t)^\top x^*\right\rangle\big|\mathcal{F}_t\right] \\
&\qquad + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2\big|\mathcal{F}_t\right] \\
&= (1-2\eta\|r_t\|_2^2)(\|r_t\|_2^2 - a) - 2\eta(\|E_t r_t\|_2^2 - r_t^\top E_t^\top\Gamma_{\mathrm{sym}}x^*) \\
&\qquad + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2\big|\mathcal{F}_t\right] \\
&\le (1-2\eta a)(\|r_t\|_2^2 - a) + \eta\frac{\|\Gamma_{\mathrm{sym}}x^*\|_2^2}{2} + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2\big|\mathcal{F}_t\right] \\
&\le (1-2\eta a)(\|r_t\|_2^2 - a) + \eta\frac{\|\Gamma_{\mathrm{sym}}x^*\|_2^2}{2} + \eta^2 C d^3,
\end{aligned}$$

where C is a constant. Let $\beta_3 = 2a$, $\alpha_3 = \frac{\|\Gamma_{\mathrm{sym}}x^*\|_2^2}{4a}$ and $\lambda_3 = \frac{Cd^3}{2a}$. This is equivalent to

$$\mathbb{E}[\|r_{t+1}\|_2^2 - a - \alpha_3\big|\mathcal{F}_t] \le (1-\beta_3)\left(\|r_t\|_2^2 - a - \alpha_3 - \eta\lambda_3\right).$$

Then (18) holds for $\|r_t\|_2^2 - a$. Denote $\mathcal{H}_t = \{\forall\tau \le t, \|r_t\|_2^2 - a \le 4\alpha_3\}$, Then we have

$$\mathbb{E}[G_{t+1}\mathcal{I}_{\mathcal{H}_t}|\mathcal{F}_t] \le G_t\mathcal{I}_{\mathcal{H}_t} \le G_t\mathcal{I}_{\mathcal{H}_{t-1}}.$$

We then bound the difference between $\|r_{t+1}\|_2^2\mathcal{I}_{\mathcal{H}_t}$ and $\mathbb{E}[\|r_{t+1}\|_2^2\mathcal{I}_{\mathcal{H}_t}|\mathcal{F}_t]$.

$$\left|\|r_{t+1}\|_2^2\mathcal{I}_{\mathcal{H}_t} - \mathbb{E}[\|r_{t+1}\|_2^2\mathcal{I}_{\mathcal{H}_t}|\mathcal{F}_t]\right| \le \eta C_3 d = \eta\phi_3.$$

where $\phi_3 = \mathcal{O}(d)$. Thus, (19) holds for $\|r_t\|_2^2 - a$. We can then apply Theorem 3.2. Choose

$$\eta = \min\left\{\mathcal{O}\left(\sigma^4\left(\log\frac{1}{\delta}\right)^{-1}\right), \mathcal{O}\left(\frac{\sigma^2}{d^2}\right)\right\},$$

then with probability $1-\delta$, we have $\|r_t\|_2^2 \le a + 4\alpha_3$ for all t's such that $t \le \mathcal{O}(\eta^{-2})$.

We next prove (14) . Suppose there exists some time t such that $\|r_t\|^2 \ge a - 2/3$, our following analysis will show that the algorithm will stay in the region such that $\|r_t\|^2 \ge a - 2/3$, for long enough time. Then we can move to Lemma 3.5. Suppose such t does not exists, i.e., $\|r_t\|^2 \le a - 2/3$, for all $t's$. Then we have the following inequality.

$$\begin{aligned}
\mathbb{E}[\|r_{t+1}\|_2^2\big|\mathcal{F}_t] &= \|r_t\|_2^2 - 2\eta\mathbb{E}\left[\left\langle r_t, \nabla_X\mathcal{F}(X_t+W_t)^\top x^*\right\rangle\big|\mathcal{F}_t\right] + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2\big|\mathcal{F}_t\right] \\
&= \left(1 - 2\eta(\|r_t\|_2^2 - a)\right)\|r_t\|_2^2 - 2\eta(\|E_t r_t\|_2^2 - r_t^\top E_t^\top\Gamma_{\mathrm{sym}}x^*) \\
&\qquad + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2\big|\mathcal{F}_t\right] \\
&\ge \left(1 - \eta(2\|E_t\|_2^2)\right)\|r_t\|_2^2 + \eta(2(a - \|r_t\|_2^2)\|r_t\|_2^2 - 2\|r_t\|_2\|E_t^\top\Gamma_{\mathrm{sym}}x^*\|_2) \\
&\ge \left(1 - \eta\left(2 + 2.5 - 2\right)\right)\|r_t\|_2^2 + \eta\left(2(a - \|r_t\|_2^2)\|r_t\|_2^2 - 2\|E_t^\top\Gamma_{\mathrm{sym}}x^*\|_2^2\right) \\
&\ge (1 - 2.5\eta)\|r_t\|_2^2 + \eta\left(3\|r_t\|_2^2 - 4r^2\|\Gamma_{\mathrm{sym}}x^*\|_2^2\right),
\end{aligned}$$

where $r^2 = \frac{c_1}{2}\sqrt{d\sigma^2}$. Let $\mathcal{E}_t = \left\{\|r_\tau\|_2^2 \geq r^2\|\Gamma_{\text{sym}}x^*\|_2, \forall \tau \leq t\right\}$. Then we have

$$\begin{aligned}
&\mathbb{E}\left[(1-2.5\eta)^{-t-1}\left(\|r_{t+1}\|_2^2 - \frac{3-4\|\Gamma_{\text{sym}}x^*\|_2}{2.5}r^2\|\Gamma_{\text{sym}}x^*\|_2\right)\mathcal{I}_{\mathcal{E}_t}]\big|\mathcal{F}_t\right] \\
&\qquad \geq (1-2.5\eta)^{-t}\left(\|r_t\|_2^2 - \frac{3-4\|\Gamma_{\text{sym}}x^*\|_2}{2.5}r^2\|\Gamma_{\text{sym}}x^*\|_2\right)\mathcal{I}_{\mathcal{E}_t} \\
&\qquad \geq (1-2.5\eta)^{-t}\left(\|r_t\|_2^2 - \frac{3-4\|\Gamma_{\text{sym}}x^*\|_2}{2.5}r^2\|\Gamma_{\text{sym}}x^*\|_2\right)\mathcal{I}_{\mathcal{E}_{t-1}}.
\end{aligned}$$

The last inequality comes from the fact $\frac{3-4\|\Gamma_{\text{sym}}x^*\|_2}{2.5} \geq 1$. The above inequality actually shows that

$$G_t = (1-2.5\eta)^{-t}\left(\|r_t\|_2^2 - \frac{3-4\|\Gamma_{\text{sym}}x^*\|_2}{2.5}r^2\|\Gamma_{\text{sym}}x^*\|_2\right)\mathcal{I}_{\mathcal{E}_{t-1}}.$$

is a submartingale. Following the same proof of Part II of Theorem 3.2, we can show that with our choice of small η, with high probability, $\|r_t\|_2^2 \geq r^2\|\Gamma_{\text{sym}}x^*\|_2 \geq \|\Gamma_{\text{sym}}x^*\|_2^2$. □

B.5 Proof of Lemma 3.5

Proof. We first show that there must exist some τ_{21} such that $\|r_t\|_2^2 > \frac{a}{3}$. We first have the following inequality:

$$\begin{aligned}
\mathbb{E}[\|r_{t+1}\|_2^2|\mathcal{F}_t] &= \|r_t\|_2^2 - 2\eta\mathbb{E}\left[\left\langle r_t, \nabla_X\mathcal{F}(X_t+W_t)^\top x^*\right\rangle|\mathcal{F}_t\right] + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2|\mathcal{F}_t\right] \\
&= \left(1-2\eta(\|r_t\|_2^2-a)\right)\|r_t\|_2^2 - 2\eta(\|E_tr_t\|_2^2 - r_t^\top E_t^\top\Gamma_{\text{sym}}x^*) \\
&\qquad + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2|\mathcal{F}_t\right] \\
&\geq \left(1-2\eta(\|r_t\|_2^2-a)\right)\|r_t\|_2^2 - 2\eta(\frac{3}{2}\|E_t\|_{\text{F}}^2\|r_t\|_2^2 + \frac{\|\Gamma_{\text{sym}}x^*\|_2^2}{2}) \\
&\qquad + \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2|\mathcal{F}_t\right] \\
&\geq \left(1-2\eta(\|r_t\|_2^2 - a + \frac{3}{2}c_1\epsilon)\right)\|r_t\|_2^2 - \eta\|\Gamma_{\text{sym}}x^*\|_2^2.
\end{aligned}$$

Denote $\mathcal{E}_t = \{\forall\tau \leq t, \|r_\tau\|_2^2 \leq \frac{a}{3}\}$. Then we have

$$\mathbb{E}[\|r_{t+1}\|_2^2\mathcal{I}_{\mathcal{E}_t}|\mathcal{F}_t] \geq \left(1-2\eta(-\frac{2}{3}a + \frac{3}{2}c_1\epsilon)\right)\|r_t\|_2^2\mathcal{I}_{\mathcal{E}_t} - \eta\|\Gamma_{\text{sym}}x^*\|_2^2\mathcal{I}_{\mathcal{E}_t}.$$

Let $G_t = \left(1+\eta(\frac{4}{3}a - 3c_1\epsilon)\right)^{-t}\left(\|r_t\|_2^2 - \frac{\|\Gamma_{\text{sym}}x^*\|_2^2}{\frac{4}{3}a-3c_1\epsilon}\right)$. Thus, we have

$$\mathbb{E}\left[G_{t+1}\mathcal{I}_{\mathcal{E}_t}|\mathcal{F}_t\right] \geq G_t\mathcal{I}_{\mathcal{E}_t} \geq G_t\mathcal{I}_{\mathcal{E}_{t-1}}.$$

The last inequality must hold, otherwise we have found a t such that $\|r_t\|_2^2 > \frac{a}{3}$. We have constructed a submartingale sequence. Following similar lines to our previous proof, with probability at least $1-\delta$, there exists $t \leq \tau_{21} = \frac{1}{a\eta}\log\frac{4a}{d\sigma^2}\log\frac{1}{\delta}$, such that $\|r_t\|_2^2 > \frac{a}{3}$.

Next, we show that the solution trajectory will stay in this region $\{\|r_t\|_2^2 > \frac{a}{3}\}$. Let $\mathcal{E}_t = \{\forall\tau \leq t, \|r_\tau\|_2^2 > \frac{a}{3}\}$.

$$\begin{aligned}
\mathbb{E}[a - \|r_{t+1}\|_2^2|\mathcal{F}_t] &= (a - \|r_t\|_2^2) + 2\eta\mathbb{E}\left[\left\langle r_t, \nabla_X\mathcal{F}(X_t+W_t)^\top x^*\right\rangle|\mathcal{F}_t\right] \\
&\qquad - \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2|\mathcal{F}_t\right] \\
&= (1-2\eta\|r_t\|_2^2)(a-\|r_t\|_2^2) + 2\eta(\|E_tr_t\|_2^2 - r_t^\top E_t^\top\Gamma_{\text{sym}}x^*) \\
&\qquad - \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2|\mathcal{F}_t\right].
\end{aligned}$$

Let $a' = a + 4\alpha_3$, where $\alpha_3 = \frac{\|\Gamma_{\text{sym}}x^*\|_2^2}{4a}$ comes from the proof of Lemma 3.4, and thus $\|r_t\|_2^2 \leq a'$. Then the above equality is equivalent to the following:

$$\begin{aligned}
\mathbb{E}[a' - \|r_{t+1}\|_2^2|\mathcal{F}_t] &= (1-2\eta\|r_t\|_2^2)(a'-\|r_t\|_2^2) + 2\eta(a'-a)\|r_t\|_2^2 \\
&\qquad + 2\eta(\|E_tr_t\|_2^2 - r_t^\top E_t^\top\Gamma_{\text{sym}}x^*) - \eta^2\mathbb{E}\left[\|\nabla_X\mathcal{F}(X_t+W_t)^\top x^*\|_2^2|\mathcal{F}_t\right] \\
&\leq (1-2\eta\|r_t\|_2^2)(a'-\|r_t\|_2^2) + 2\eta(a'-a)a' + 2\eta(\|E_tr_t\|_2^2 - r_t^\top E_t^\top\Gamma_{\text{sym}}x^*).
\end{aligned}$$

We further have

$$\mathbb{E}[a' - \|r_{t+1}\|_2^2 \mathcal{I}_{\mathcal{E}_t} | \mathcal{F}_t] \leq \left(1 - \eta \frac{2}{3} a\right) (a' - \|r_t\|_2^2) \mathcal{I}_{\mathcal{E}_t} + C_4 \sqrt{d\sigma^2} \mathcal{I}_{\mathcal{E}_t},$$

which is equivalent to the following equations.

$$\mathbb{E}[(a' - \|r_{t+1}\|_2^2 - C_4\sqrt{d\sigma^2}) \mathcal{I}_{\mathcal{E}_t} | \mathcal{F}_t] \leq \left(1 - \eta \frac{2}{3} a\right) \left(a' - \|r_t\|_2^2 - C_4 a \sqrt{d\sigma^2}\right) \mathcal{I}_{\mathcal{E}_{t-1}}.$$

Then we can construct a supermartingale $G_t \mathcal{I}_{\mathcal{E}_{t-1}} = \frac{1}{(1-\eta\frac{2}{3}a)^t}(a' - \|r_t\|_2^2 - C_4\sqrt{d\sigma^2})\mathcal{I}_{\mathcal{E}_{t-1}}$. Applying Theorem 3.2, one can show with probability $1-\delta$, we have $a' - \|r_t\|_2^2 \leq 4\frac{a' - \frac{a}{3}}{4}$ or equivalently $\|r_t\|_2^2 \geq \frac{a}{3}$ for all t's such that $\tau_{21} \leq t \leq \mathcal{O}(\eta^{-2})$. Then the following inequality always holds.

$$\mathbb{E}[(a' - \|r_{t+1}\|_2^2 - C_4\sqrt{d\sigma^2}) | \mathcal{F}_t] \leq \left(1 - \eta \frac{2}{3} a\right) \left(a' - \|r_t\|_2^2 - C_4 \sqrt{d\sigma^2}\right).$$

Following similar lines to the proof of $\|r_t\|_2^2 \leq a + 4\alpha_3$, one can show with probability $1-\delta$, we have $\|r_t\|_2^2 \geq a' - C_5\sqrt{d\sigma^2}$ for all t's such that $\tau_{22} \leq t \leq \mathcal{O}(\eta^{-2})$, where $\tau_{22} = \mathcal{O}\left(\frac{1}{\eta}\log\frac{1}{d\sigma^2}\log\frac{1}{\delta}\right)$. Take $\tau_2 = \tau_{21} + \tau_{22}$, we have when $t \geq \tau_2$,

$$a' - C_5\sqrt{d\sigma^2} \leq \|r_t\|_2^2 \leq a + 4\alpha_3.$$

Therefore there exists some constant $c_2 > 0$ such that when $t \geq \tau_2$,

$$|\|r_t\|_2^2 - 1| \leq c_2\sqrt{d\sigma^2}.$$

□

C PROOF OF LEMMA 3.6

Proof. Note that we can refine the upper bound of the norm of X_t as follows: $\|X_t\|_{\mathrm{F}}^2 = \|E_t\|_{\mathrm{F}}^2 + \|r_t\|_2^2 \leq (c1 + c2)\sqrt{d\sigma^2}$. We first write down the update of $E_t r_t$:

$$E_{t+1}r_{t+1} = E_t r_t - \eta E_t \nabla_r \mathcal{F}(X_t + W_t) - \eta \nabla_E \mathcal{F}(X_t + W_t) r_t + \eta^2 \nabla_E \mathcal{F}(X_t + W_t) \nabla_r \mathcal{F}(X_t + W_t).$$

For notational simplicity, denote $D_{1,t} = \nabla_E \mathcal{F}(X_t + W_t)\nabla_r \mathcal{F}(X_t + W_t)$. By simple calculation, we know that $\|D_{1,t}\|_2$ is at most $\mathcal{O}(d\sigma^2)$. Then the update of the squared norm of $E_t r_t$ is as follows:

$$\begin{aligned}
\|E_{t+1}r_{t+1}\|_2^2 &= \|E_t r_t\|_2^2 - 2\eta (E_t r_t)^\top E_t \nabla_r \mathcal{F}(X_t + W_t) - 2\eta (E_t r_t)^\top \nabla_E \mathcal{F}(X_t + W_t) r_t \\
&\quad + \eta^2 \left(\|E_t \nabla_r \mathcal{F}(X_t + W_t)\|_2^2 + \|\nabla_E \mathcal{F}(X_t + W_t) r_t\|_2^2 + 2(E_t r_t)^\top D_{1,t}\right) \\
&\quad - 2\eta^3 D_{1,t}^\top \left(E_t \nabla_r \mathcal{F}(X_t + W_t) + \nabla_E \mathcal{F}(X_t + W_t) r_t\right) + \eta^4 \|D_{1,t}\|_2^2 \\
&= \|E_t r_t\|_2^2 - 2\eta (E_t r_t)^\top E_t \nabla_r \mathcal{F}(X_t + W_t) - 2\eta (E_t r_t)^\top \nabla_E \mathcal{F}(X_t + W_t) r_t + \eta^2 D_{2,t},
\end{aligned}$$

where

$$\begin{aligned}
D_{2,t} = &\left(\|E_t \nabla_r \mathcal{F}(X_t + W_t)\|_2^2 + \|\nabla_E \mathcal{F}(X_t + W_t) r_t\|_2^2 + 2(E_t r_t)^\top D_{1,t}\right) \\
&- 2\eta D_{1,t}^\top \left(E_t \nabla_r \mathcal{F}(X_t + W_t) + \nabla_E \mathcal{F}(X_t + W_t) r_t\right) + \eta^2 \|D_{1,t}\|_2^2.
\end{aligned}$$

By simple calculation, we know $D_{2,t}$ is at most $\mathcal{O}(1)$. Thus, the last three terms is $\eta^2 D_{2,t} \le C_6\eta^2$, and the update is dominated by the $\mathcal{O}(\eta)$ terms. We next calculate the $\mathcal{O}(\eta)$ terms as follows

$$
\begin{aligned}
&\mathbb{E}[(E_t r_t)^\top \nabla_E \mathcal{F}(X_t+W_t) r_t|\mathcal{F}_t] \\
=&(E_t r_t)^\top (\mathrm{Id}-\mathrm{Id}_{x^*})\left((X_tX_t^\top - Y^*)X_t - \Gamma_{\mathrm{sym}}X_t + (2d+1)\gamma^2 X_t\right) r_t \\
=&r_t^\top E_t^\top \Big((x^* r_t^\top + E_t)(E_t^\top E_t r_t + r_t r_t^\top r_t) - x^* r_t^\top r_t - \Gamma_{\mathrm{sym}} E_t r_t \\
&\qquad - \Gamma_{\mathrm{sym}} x^* r_t^\top r_t + (2d+1)\gamma^2 E_t r_t + (2d+1)\gamma^2 x^* r_t^\top r_t\Big) \\
=&\left(\|r_t\|_2^2 + r_t^\top E_t^\top x^* + (2d+1)\gamma^2\right)\|E_t r_t\|_2^2 - r_t^\top E_t^\top \Gamma_{\mathrm{sym}} E_t r_t \\
&+ \left(\|r_t\|_2^4 - \|r_t\|_2^2 + (2d+1)\gamma^2\|r_t\|_2^2\right) r_t^\top E_t^\top x^* + \|E_t^\top E_t r_t\|_2^2 + \|r_t\|_2^2 r_t^\top E_t^\top \Gamma_{\mathrm{sym}} x^* \\
\ge&\left(\frac{3}{4}\|r_t\|_2^2 - \|\Gamma_{\mathrm{sym}}\|_2 + r_t^\top E_t^\top x^* + (2d+1)\gamma^2\right)\|E_t r_t\|_2^2 + \|r_t\|_2^2\left(\|r_t\|_2^2 - 1 + (2d+1)\gamma^2\right) r_t^\top E_t^\top x^* \\
&\qquad + \|r_t\|_2^2\left(\frac{1}{4}\|E_t r_t\|_2^2 - r_t^\top E_t^\top \Gamma_{\mathrm{sym}} x^*\right) \\
\ge&\left(\frac{3}{4}\|r_t\|_2^2 + \|\Gamma_{\mathrm{sym}}\|_2 + r_t^\top E_t^\top x^* + (2d+1)\gamma^2\right)\|E_t r_t\|_2^2 \\
&\qquad + \|r_t\|_2^2\left(\|r_t\|_2^2 - 1 + (2d+1)\gamma^2\right) r_t^\top E_t^\top x^* - \frac{1}{2}\|\Gamma_{\mathrm{sym}} x^*\|_2^2,
\end{aligned}
$$

and

$$
\begin{aligned}
&\mathbb{E}[(E_t r_t)^\top E_t \nabla_r \mathcal{F}(X_t+W_t)|\mathcal{F}_t] \\
=&r_t^\top E_t^\top E_t\left((X_t^\top X_t X_t^\top - Y^*)x^* - X_t^\top \Gamma_{\mathrm{sym}} x^* + (2d+1)\gamma^2 X_t^\top x^*\right) \\
=&\|E_t^\top E_t r_t\|_2^2 + \left(\|r_t\|_2^2 - 1 - x^{*\top}\Gamma_{\mathrm{sym}} x^* + (2d+1)\gamma^2\right)\|E_t r_t\|_2^2 - r_t^\top E_t^\top E_t E_t^\top \Gamma_{\mathrm{sym}} x^* \\
\ge&\left(\|r_t\|_2^2 - a\right)\|E_t r_t\|_2^2 - r_t^\top E_t^\top E_t E_t^\top \Gamma_{\mathrm{sym}} x^*.
\end{aligned}
$$

Combine the above two inequalities together and we have:

$$
\begin{aligned}
\mathbb{E}[\|E_{t+1} r_{t+1}\|_2^2|\mathcal{F}_t] \le& \left(1 - 2\eta\left(\frac{7}{4}\|r_t\|_2^2 - a - \|\Gamma_{\mathrm{sym}}\|_2 + r_t^\top E_t^\top x^* + (2d+1)\gamma^2\right)\right)\|E_t r_t\|_2^2 \\
&- 2\eta\Big(\|r_t\|_2^2\left(\|r_t\|_2^2 - 1 + (2d+1)\gamma^2\right) r_t^\top E_t^\top x^* \\
&\qquad - \frac{1}{2}\|\Gamma_{\mathrm{sym}} x^*\|_2^2 - r_t^\top E_t^\top E_t E_t^\top \Gamma_{\mathrm{sym}} x^*\Big) + \eta^2 D_{2,t} \\
\le& (1-\eta)\|E_t r_t\|_2^2 + \eta C_7 d\sigma^2 + C_6\eta^2.
\end{aligned}
$$

Let $\alpha_4 = C_7 d\sigma^2$, $\lambda_4 = C_6$ and $\beta = 1$, then (18) holds. Moreover, one can also check $|\|E_t r_t\|_2^2 - \mathbb{E}[\|E_t r_t\|_2^2|\mathcal{F}_{t-1}] \le \eta C_8$, where C_8 is some constant. Then we can apply Theorem 3.2. Choose

$$
\eta = \mathcal{O}\left(d\sigma^2\left(\log\frac{1}{\delta}\right)^{-1}\right),
$$

then with probability at least $1-\delta$, there exists some constant $c_3 > 0$ such that

$$
\|E_{t+1} r_{t+1}\|_2^2 \le c_3 d\sigma^2,
$$

for all $t's$ such that $\tau_3 \le t \le \mathcal{O}(\frac{1}{\eta^2})$, where $\tau_3 = \mathcal{O}(\frac{1}{\eta}\log\frac{1}{d\sigma^2}\log\frac{1}{\delta})$. □

C.1 Proof of Lemma 3.7

Proof. Note that the Frobenius norm of Γ can be written as a sum of d^2 squared subGaussian random variable: $\|\Gamma\|_{\mathrm{F}}^2 = \sum_{i,j}\Gamma_{ij}^2$. Since $\Gamma_{i,j}$ is subGaussian, $\Gamma_{i,j}^2$ is sub-exponential. Then we have the following concentration inequality.

$$
\mathbb{P}\left(\left|\frac{\|\Gamma\|_{\mathrm{F}}}{d} - \sigma\right| \ge t\right) \le 2\exp\left(-\frac{d^2t^2}{2C\sigma^2}\right),
$$

for any $t > 0$. Take $t = \frac{\sigma}{d}\sqrt{2C\log\frac{8}{\delta}}$, we have with probability at least $1 - \frac{\delta}{4}$, we have

$$\|\Gamma\|_{\mathrm{F}} \le d\sigma + \sigma\sqrt{2C\log\frac{8}{\delta}}.$$

Then we have

$$\|\Gamma_{\mathrm{sym}}\|_{\mathrm{F}} \le \frac{1}{2}(\|\Gamma\|_{\mathrm{F}} + \|\Gamma^\top\|_{\mathrm{F}}) = \|\Gamma\|_{\mathrm{F}} \le d\sigma + \sigma\sqrt{2C\log\frac{8}{\delta}}.$$

Moreover, since $\|x^*\|_2 = 1$, we have

$$\mathbb{P}\left(\left|\frac{\|\Gamma x^*\|_2}{\sqrt{d}} - \sigma\right| \ge t\right) \le 2\exp\left(-\frac{dt^2}{2C\sigma^2}\right),$$

$$\mathbb{P}\left(\left|\frac{\|\Gamma^\top x^*\|_2^2}{\sqrt{d}} - \sigma\right| \ge t\right) \le 2\exp\left(-\frac{dt^2}{2C\sigma^2}\right).$$

Take $t = \frac{\sigma}{\sqrt{d}}\sqrt{2c\log\frac{8}{\delta}}$, we have with probability at least $1 - \frac{\delta}{4}$, we have

$$\|\Gamma x^*\|_2 \le \sqrt{d}\sigma + \sigma\sqrt{2C\log\frac{8}{\delta}},$$

$$\|\Gamma^\top x^*\|_2 \le \sqrt{d}\sigma + \sigma\sqrt{2C\log\frac{8}{\delta}}.$$

Then we have

$$\|\Gamma_{\mathrm{sym}} x^*\|_2 \le \frac{1}{2}(\|\Gamma x^*\|_{\mathrm{F}} + \|\Gamma^\top x^*\|_{\mathrm{F}}) \le \sqrt{d}\sigma + \sigma\sqrt{2C\log\frac{8}{\delta}}.$$

By Theorem 4.4.5 in Vershynin (2018), we have for any $t > 0$,

$$\|\Gamma\|_2 \le C\sigma(2\sqrt{d} + t),$$

with probability at least $1 - 2\exp(-t^2)$, where C is some absolute constant. Take $t = \sqrt{\log\frac{2}{\delta}}$, we have with probability at least $1 - \delta$,

$$\|\Gamma\|_2 \le C\sigma\left(2\sqrt{d} + \sqrt{\log\frac{2}{\delta}}\right).$$

Then we have

$$\|\Gamma_{\mathrm{sym}}\|_2 \le \frac{1}{2}(\|\Gamma\|_2 + \|\Gamma^\top\|_2) = \|\Gamma\|_2 \le C\sigma\left(2\sqrt{d} + \sqrt{\log\frac{2}{\delta}}\right).$$

Take $\delta = \mathcal{O}(\exp(-d))$ and we prove the result. □

C.2 Proof of Lemma 3.8

Proof. Note that our initialization can be rewritten as $X_0 = r x_0' x_0'^\top$, where $r^2 \sim \mathrm{UNIF}[0,1]$ and $x_0' \sim \mathrm{UNIF}(\mathbb{S}(1))$. Then

$$\|r_0\|_2^2 = \|X_0^\top x^*\|_2^2 = r^2(x_0'^\top x^*)^2 = r^2\cos(\angle(x_0', x^*))^2.$$

Note that the probability $\|r_0\|_2^2 \ge C^2 d\sigma^2$ can then be bounded as follows.

$$\begin{aligned}
\mathbb{P}\left(r^2\cos(\angle(x_0', x^*))^2 \ge C^2 d\sigma^2\right) &\ge \mathbb{P}\left(r^2 \ge C\sqrt{d\sigma^2}, \cos(\angle(x_0', x^*))^2 \ge C\sqrt{d\sigma^2}\right) \\
&\ge \mathbb{P}\left(r^2 \ge C\sqrt{d\sigma^2}\right) + \mathbb{P}\left(\cos(\angle(x_0', x^*))^2 \ge C\sqrt{d\sigma^2}\right) - 1 \\
&= 1 - \mathbb{P}\left(r^2 \le C\sqrt{d\sigma^2}\right) - \mathbb{P}\left(\cos(\angle(x_0', x^*))^2 \le C\sqrt{d\sigma^2}\right) \\
&= 1 - \mathbb{P}\left(r^2 \le C\sqrt{d\sigma^2}\right) - 4\mathbb{P}\left(\arccos\left(\sqrt{C\sqrt{d\sigma^2}}\right) \le \theta \le \frac{\pi}{2}\right) \\
&= 1 - \mathcal{O}\left(\frac{1}{d^{0.25}}\right),
\end{aligned}$$

where $\theta \sim \mathrm{UNIF}[0, \pi/2]$. That is with high probability, we have $\|r_0\|_2^2 \ge \|\Gamma_{\mathrm{sym}} x^*\|_2^2$.

Moreover, $\mathbb{P}(\|X_0\|_{\mathrm{F}}^2 \le 1 - C_1\sqrt{d\sigma^2}) = \mathbb{P}(r^2 \le 1 - C_1\sqrt{d\sigma^2}) = 1 - C_1\sqrt{d\sigma^2} = 1 - \mathcal{O}(\frac{1}{d^{0.25}})$. We finish the proof. □