We appreciate reviewers' valuable comments. We will correct typos and reply to comments in the following.

To Reviewer #1: • We would like to correct a typo that may cause confusion. The equation under line 67 should be $\frac{1}{d^2}\|\widehat{X}\widehat{X}^\top - Y^*\| = \mathcal{O}_P(\sigma^2/d)$. Thus, PGD does yield better test error than GD. We will clear this typo and add more discussion in our future version.

• The implicit PSD constraint in our formulation cannot improve the convergence rate of GD and PGD. Moreover, our over-parameterization is the most natural way, especially when the rank of the signal matrix is unknown. We will add more discussion and clarify this point in our next version.

• Asymmetric Noise: Our result also works when noise is symmetric. As shown in (3), PGD and GD actually try to minimize $||XX^\top - Y_{sym}||_F^2$ and thus whether Γ is symmetric or not will not change our result.

• Jain et al, 2017 consider a slightly different problem , i.e., computing matrix square root. There is no direct comparable results, but the proof technique (Lemmas 4.1 -4.5, Jain et al, 2017) can be exactly followed to show the convergence of GD in our setting. We will add a proof of GD in the appendix in our next version.

To Reviewer #2: • Extension to Rank-r case. Our proof technique can be applied to rank-r case but will be much more involved. Specifically, we can extend our subspace dissipativity condition (Lemma 1) to the projection of $X_t's$ on the eigenspace and its orthogonal, respectively. This can be done by considering the projection on the span of each eigenvector. 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. Due to the space limit, we cannot present full details here .

• We first briefly discuss the main difference between our work and Blanc et al., 2020. (a). They consider a different problem setup, i.e., a 2-layer neural network without over-parameterization, and do not provide any explicit recovery error bound. (b). The noise perturbation in this paper is on labels, while we consider perturbing parameters. (c). the l_2 regularizer in. Blanc et al., 2020 is the l_2 norm of gradient. If we directly adopt their regularizer to our problem setting, i.e., $||(XX^\top - Y)X||_F^2$, to our best knowledge, there is not existing literature that shows this regularizer can help find solutions with low complexity. Given these differences, we do not think applying the result from Blanc et al. 2020 on low rank matrix recovery can recover our results and explain the implicit bias of noise when over-parameterized.

• Related Literature. All of these papers study related problems but have fundamental differences with our work.

[Haochen et al, 2020]. We have to mention that this paper was first released on Jun 15th, while we submit our paper to NeurIPS on Jun 5th. Beside, Haochen et al, 2020 consider perturbing labels while our work consider perturbing parameters.

[Du and Lee, 2018] consider a different problem setting. They consider a problem without underlying low complexity generating models and provide a generalization bound and optimization landscape analysis. However, our problem has an underlining low rank generating model and we provide an estimation error bound analysis. Most important, we study the implicit regularization effect of noise without any explicit regularizers used in their work.

• Proof Technique. We must remark that our proof technique is very different from that of Li et al. Both of these two works apply projection on eigen-space, but that is a quite common technique in low rank matrix factorization analysis. Our key analytics tools are sub-space dissapativity condition and super-martingale theory which cannot be found in Li et al, which considers gradient descent without noise.

To Reviewer #3: We thank the reviewer for carefully reading our paper. We will polish our supplementary material in our next version to make our proof easier to understand. Due to space limit, we will only clarify some major issues here and will clean other issues in our next version.

• Weakness 3. We thank the reviewer for this great suggestion. We can try to solve the smoothed problem (4) directly in the future. Moreover, in our next version, we will provide more intuitive explanation and highlight where the regularization effect happens to make our theory easier to understand.

• We thank the reviewer for carefully check the proof. We actually use a uniform learning rate in all of our lemmas and theorems. Thus, in some results, such as Lemma 2, Lemma 3, the learning rate used is much smaller than required. We will clarify this in our next version.

• We are sorry for this typo here that confuses you. The first dot point should be $\|\mathbb{E}_t\|_F^2 \leq \|\mathbb{E}_0\|_F^2 + c_1\sqrt{d\sigma^2}$. Here, we take $\alpha = \frac{\|\mathbb{E}_0\|_F^2 + c_1\sqrt{d\sigma^2}}{2}$ with proper c_1 such that $\alpha \geq \alpha_2$. Then directly applying Theorem 2 Part I, we can prove the inequality. The argument $\|\mathbb{E}_0\|_F^2 + c_1\sqrt{d\sigma^2} \leq 1$ comes from our Assumption 2.

• The last statement of Lemma 1 is used in Lemma 4 to prove that when $\|r_t\|_2^2 \leq 1$, PGD will increase $\|r_t\|_2^2$ to 1 (Line 459). Moreover, line 164 provides the conditions such that (11) holds. In our proof, we explicitly show that these conditions can be achieved by PGD, see (13), (14) and also (15). In our next version, we will provide more intuitions to these conditions.