First we would like to thank the reviewers for their interest on the contributions of the main paper. We share the enthusiasm of the reviewers about the promising theoretical results on the discrete dynamics and the perturbation analysis of the paper and we highly appreciated their interest and their detailed comments.

**Regarding the strength of Assumption 1. (R1 & R3 & R4)** Eq. 13 is always true using a large enough $\epsilon$ (one can always find such decomposition described in Eq. 13), but the matrix $\boldsymbol{B}$ may be large (L 279 we describe how to compute a candidate for $\boldsymbol{B}$ in this decomposition). Formally speaking, Assumption 1 should be stated as a proposition and the informal assumption associated with this proposition is that the matrix $\boldsymbol{B}$ (or equivalently epsilon) is relatively small. As noted by reviewer 1 and reviewer 4 this assumption is significantly weaker than the one done in the related work and thus is a major improvement compare to them. Experiments in section 4.1 assesses that this assumption is actually quite met in practice. We thank thanks the reviewers for pointing this out, we will clarify this in the revision.

**Results for no more than two layers. (R2 & R3 & R4)** We agree with the reviewers that we should remove "deep" from the title. On the question whether or not we can extend our analysis to more than two layers, it is a very interesting question. To our knowledge, proving such result is still an open question, the reason being that there is not closed form solutions for the continuous dynamics when $n \geq 3$ (previous related works used proof technique necessitating closed form solutions of the continuous dynamics). However, our discrete analysis and our perturbation analysis did not use a closed form solution, letting us thinks that we can be optimistic and that we could use similar techniques for $n \geq 3$.

**"the perturbation analysis, [...] not discussed at all in the body of the paper". (R4)** We think that R4 missed how we address perturbation analysis in the body of the paper: we discuss the practical relevance of the assumption $\epsilon$ small (L134-149) providing several application cases and provide intuitions regarding this hypothesis: $\epsilon$ represent to what extent the covariance matrices $\boldsymbol{\Sigma}_x$ and $\boldsymbol{\Sigma}_{\boldsymbol{xy}}$ do not commute (L133).

We provide experimental evidence of the relevance of this assumption ($\epsilon$ small) in §4.1. We consider that the motivations behind the perturbation analysis are well discussed as noted by R1 and R3. We decided not to discuss the details of the proof technique itself for obvious space issues: in term of priority motivating a result comes before the discussion of its proof. However, we understand that our presentation regarding perturbation analysis may be improved and we thank R4 for pointing this out. We will also add some intuitions regarding this proof and its difficulties in the revision.

**Novelty of the discrete case. (R4)** R4 mentioned that "transition from continuous (gradient flow) to discrete (gradient descent) optimization – is also relatively simple". We are conscious that the transition from continuous to discrete may appear easy but we think that it is not an accident that the close related work only addressed the continuous dynamics: working on the discrete dynamics is more challenging. We are quite surprised that R4 do not mention at all the whole paragraph (L238-248) we wrote on "why the discrete analysis is challenging" where we developed some points to explain the new difficulties arising when working with the discrete dynamics. We explained in this paragraph why this transition is difficult. We encourage the reviewer to consider it carefully in their revision of the review.

**Implicit Regularization. (R4)** Regularization is a restriction within the search space of solution in order to improve generalization. A low rank constraint is an explicit regularization. Using a method that finds these low rank solutions without explicitly putting low rank constraint is a restriction in the search space of potential solutions with good generalization and thus is an implicit regularization. Early stopping is described as a regularization technique for deep learning in [Goodfellow, Bengio and Courville, 2016, §7.8] and is still relevant in practice, e.g. with corrupted labels [Li, Soltanolkotabi and Oymak, 2019].

We think that even though early stopping might not be necessary in some specific cases, the study of the optimization path is a conceptual advance in term of understanding of the inductive bias of gradient descent: it help to explain why the test 0-1 loss plateaus while the training optimization loss still decreases. For instance, in [Vaswani, Mishkin et al. 2019, Fig. 3] we can see that the 0-1 test accuracy plateaus while the training optimization loss is still decreasing showing that along the optimization path the solutions have at least as good generalization properties as the final solution.

**About initialization formula. (R1)** In Theorem 1, the matrix $Q$ can be chosen arbitrary. Thus, by density of the invertible matrices, for almost initialization $W_1$, one could find a matrix $Q$ to get the desired factorization. Thus, only $W_2$ requires to be specifically initialized. In practice, practitioners have the freedom to choose the initialization.

This initialization is necessary with the current proof technique (Lemma 5 in appendix is not true anymore if the $W_1$ and $W_2$ are not initialized with different $(\delta_i)_i$). We think that for almost all random initialization the phenomenon of sequential learning still occurs. It is confirmed by our experiments in §4.2 where $W_1$ and $W_2$ are initialized randomly.

Regarding the scaling, having different vanishing $\delta_i$ just rescales the times $T_i$ depending on the relative speed at which each component vanishes. For instance, if $\delta \to 0$, we would have $T_i = {}^{\delta_i}\!/_{\delta\sigma_i}$ and thus the phase transition time depends on the limit of the ratio ${}^{\delta}\!/_{\delta_i}$. We only presented the case $\delta_i = \delta$ for simplicity of the discussion.