We thank the reviewers for the positive reviews and valuable feedback. First, we provide general comments addressing remarks of multiple reviewers. Then, we reply to other remarks. We will update the manuscript accordingly.

Connection to NTK (Raised by Rev. 1, 3, 4). The idea of the proof is to combine the PL-inequality in Lemma 4.1 with the fact that the matrices $\{F_1, W_3, \dots, W_L\}$ stay full rank during training. To show the latter, we prove that the weights cannot move too far from initialization, see l. 192-195. Our non-convex optimization perspective allows us to consider more general settings than existing NTK analyses. In fact, if the width of one of the layers is constant, then the NTK is not well defined. On the contrary, our paper just requires the first layer to be overparameterized (i.e., all the other layers can have constant widths).

Training data and generalization (Raised by Rev. 2, 3). As noted by Rev. 3, if $x_i = x_j$, then $\lambda_F = 0$. Thus, Assumption 3.2 cannot hold unless $\Phi = 0$ (i.e., we initialize at a global minimum) or $X = 0$ (i.e., the GD iterates do not move). In general, if the data points are not parallel and the activation function is analytic and not polynomial, then $\lambda_* > 0$ (and thus, $\lambda_F > 0$), see [15]. Furthermore, if x_i and x_j are close, then λ_F is small and, therefore, α_0 is small. Thus, GD requires more iterations to converge to a global optimum. This happens regardless of the value of y_i and y_j. Providing results for deep pyramidal networks that depend on the quality of the labels is an outstanding problem. Solving it could also lead to generalization bounds, see e.g. "Fine-Grained Analysis of Optimization . . ." by Arora et al.

Pyramidal network and spectrum of W_l (Raised by Rev. 1, 2). The pyramidal assumption is needed for Lemma 4.1.3 (l. 176). The key idea (see Lemma 4.3 in [27] for the proof) is that the norm of the gradient can be lower bounded by the smallest singular value of $\prod_{p=3}^{L} A_p$ with $A_p = \Sigma_{p-1}(W_p \otimes \mathbb{I}_N) \in \mathbb{R}^{r_{p-1} \times r_p}$. Assuming that $r_2 \geq r_3 \geq \ldots \geq r_L$, one can further lower bound this quantity by the product of the smallest singular values of the A_p's. This is where our assumption on the pyramidal topology comes from. Lemma 4.1 should be seen as providing a sufficient condition for a PL-inequality, rather than suggesting that such a PL-inequality holds only for pyramidal networks. Intuitively, if W_l has large minimum singular value and is well-conditioned, then GD will keep it away from the zero-measure set of low-rank matrices, in which case the loss satisfies the PL-inequality and has Lipschitz gradient, thus leading to convergence.

Rev. 1. *Initialization in Section 3.1:* Concretely, one can use Xavier's initialization for W_1, pick $W_2 = 0$ and $[W_l]_{ij} \sim \mathcal{N}(0, (28c)^2/n_{l-1})$ under the extra assumption $\sqrt{n_{l-1}} \geq 2\sqrt{n_l}$ for sufficiently large c. This fulfils our assumptions w.p. $\geq 1 - 2\sum_{l=3}^{L} e^{-n_{l-1}/32}$. Another option is to pick W_l to be scaled identity matrices (or rectangular matrices whose left block is a scaled identity). *Weakness 1:* The reviewer is right. Lemma 4.1.4 is not used explicitly, it's only meant to add an interpretation to Lemma 4.1.3. *Weakness 2:* Yes, λ_F^2 and λ_F^3 correspond to the second and third powers of λ_F. Our proof of Theorem 3.2 requires both lower bounds on λ_F. *Cross-entropy loss:* The challenge is that this loss may not satisfy our PL inequality. Oftentimes a different analysis is required, which leads to stronger assumptions on the data and weaker convergence guarantees, see e.g. [11, 22, 28, 36]. *ReLU:* Currently, ReLU does not work because *(i)* its derivative is not Lipschitz, which is needed to prove (19), and *(ii)* it can have zero derivative, while we need $\gamma > 0$ for the PL-inequality to hold. The second problem seems to us more fundamental, i.e. how to show a PL-inequality for ReLU and ensure that it holds throughout the trajectory of GD.

Rev. 2. *Clarifications:* $n_1 < n_2$ is allowed; We will explicitly mention the condition $\sqrt{n_{l-1}} \geq 1.01(\sqrt{n_l} + t)$ for the Xavier case; We will define the sub-gaussian norm. *Case $L = 2$:* Conditions (4)-(5) in Assumption 3.1 become $\lambda_F^2 \geq 12 \left\|X\right\|_F \sqrt{2\Phi(\theta_0)} \max(\bar{\lambda}_1, \bar{\lambda}_2)$ and $\lambda_F^3 \geq 24 \left\|X\right\|_2 \left\|X\right\|_F \sqrt{2\Phi(\theta_0)}\bar{\lambda}_2$. To satisfy the first condition, one can scale W_1^0 by a constant c, and set $W_2^0 = 0$. In fact, one can prove that λ_F^2 scales with c^2 for the class of activations in (2), whereas the RHS scales with c. Thus, when c is large enough, the first condition holds. Similarly, the second condition also holds. As for Theorem 3.2, the expressions for α_0, Q_0 and Q_1 simplify as the quantities involving λ_l and $\bar{\lambda}_l$ disappear for $l \in [3, L]$. *About $\phi > 0$:* We haven't worked out our bounds explicitly for this assumption, but this is an interesting direction. One idea is to follow an approach similar to Appendix B of [29] to relate λ_F to ϕ. *Overloaded notation:* Yes, it is intended. We want to apply the lemma for any upper bounds of $\bar{\lambda}_l$.

Rev. 3. *Weakness 2:* The same result holds if the data has unit norm and the weights of the first layer are scaled up by a factor $\sqrt{d}$. The scaling of x_i is chosen so that $\langle x_i, w_j \rangle \sim \mathcal{N}(0, 1)$, w_j being the weight of the first layer. If this is not the case, one would need to extend the Hermite analysis of Lemma D.3 in Appendix D.2. *Lines 171-172:* The PL-inequality follows from the same argument of Lemma 4.3 in [27]. We apologize for the confusion. *Line 177:* Part 4 of Lemma 4.1 basically follows from part 3 by setting the gradient on the LHS to zero and, in the RHS of l. 176, W_2 only appears implicitly via the Σ_l's matrices. *Lines 194-198:* Thanks! We will mention this. We will also fix the typos.

Rev. 4. *Double descent:* Thanks for an intriguing question. It is indeed possible (or even likely) that pyramidal networks exhibit a nonmonotonic behavior in the test loss (double descent or even more complicated multi-scale phenomena as in "The Neural Tangent Kernel in High Dimensions. . ." by Adlam and Pennington). One way forward is to study the spectrum of the feature matrices at the different layers by extending the analysis of the paper mentioned above to the pyramidal architecture. We regard this as a challenging (yet very interesting) open direction.