We thank all reviewers for their insightful comments and acknowledging the importance of this work. Reviewers 1, 2, and 4 recommended our paper for "clear accept" or "accept". Although our insufficient explanation seems to have made Reviewer 5 a bit confused, we expect that the following description will clear up his/her misunderstandings.

To Reviewer 5: (i) On singularity of Σ. "the analysis in the paper implies the instability of exact NGD". Our analysis does NOT imply the instability of the exact NGD. We guess you would be missing some of the following points. Theorem 4.1 assumed the positive definiteness of Σ and says nothing on NGDs with singular Σ. When Σ is singular, we need a careful look at how to calculate the pseudo-inverse. In Theorem 4.1 and Section 4, we considered the NGD with the layer-wise block approximation $G_{\text{layer},t}$ (15) and took its pseudo-inverse in the form of (S.72,73) (or (S.87,88)). When Σ is positive definite, we can use the pseudo-inverse of the zero damping limit (S.73) without any instability. When Σ is singular, we can see that Σ exists inside the matrix inverse (S.72) and it may cause instability as the damping term gets close to zero. This instability of (S.72) was empirically confirmed in the singular tri-diagonal case ($L = 3s + 2$). For *general* singular Σ, this instability seems essentially unavoidable. In contrast, exact NGD has $\Sigma = 11^\top$ and this Σ works as a *special* singular matrix in (S.72). We can make Σ inside of the inverse disappear and avoid the instability! That is, we have $S_0^\top(\Sigma \otimes I_{CN})S_0 = J_0^\top J_0$ and it makes (S.72) the pseudo-inverse of the exact NGD (9) as follows:

$$(S_0^\top(\Sigma \otimes I_{CN})S_0/N + \rho I)^{-1} J_0^\top (f - y)/N = J_0^\top (J_0 J_0^\top/N + \rho I)^{-1}(f - y)/N \tag{C.1}$$

We can take the zero damping limit without any instability. Note that the transformation (C.1) holds for any J_0. Potentially, there may exist a combination of a certain singular Σ and a certain J_0 (e.g. certain network architecture) which can avoid the instability of (S.72). Finding such an exceptional case may be an interesting topic, although it is out of the scope of the current work. To avoid the misunderstanding of the specialty of $\Sigma = 11^\top$, we will add the above explanation in the revised manuscript.

(ii) On the mini-norm solution. "when $\lambda \to 0$, it seems that G_0 doesn't matter anymore". G_0 is essential and explicitly appears when $\lambda \to 0$. The point is that we consider the limit of $\lambda \to 0$ after taking $\operatorname{argmin}_\theta$ in the derivation of the mini-norm solution. In other words, the operation $\lim_{\lambda\to 0} \operatorname{argmin}_\theta$ is not necessarily equal to $\operatorname{argmin}_\theta \lim_{\lambda\to 0}$. Let us denote $E_\lambda(\theta) := \frac{1}{2N}\|y - J_0\theta\|_2^2 + \frac{\lambda}{2}\theta^\top G_0 \theta$. Since we consider an overparameterized model, we have many global minima satisfying $E_0(\theta) = 0$ and $\operatorname{argmin}_\theta E_0(\theta)$ is not unique. In contrast, $\theta_\lambda^* := \operatorname{argmin}_\theta E_{\lambda>0}(\theta)$ is unique. After a straight-forward linear algebra, $\nabla_\theta E_{\lambda>0}(\theta) = 0$ leads to

$$\theta_\lambda^* = (\lambda G_0 + J_0^\top J_0/N)^{-1} J_0^\top y/N = G_0^{-1} J_0^\top (\lambda I + J_0 G_0^{-1} J_0^\top/N)^{-1} y/N = \frac{1}{\lambda + \alpha} G_0^{-1} J_0^\top y/N \tag{C.2}$$

where we used a matrix formula $(A + BB^\top)^{-1}B = A^{-1}B(I + B^\top A^{-1}B)^{-1}$ (Eq.(162) in [K. B. Petersen, & M. S. Pedersen, The matrix cookbook. (2012)]) and the isotropic condition $J_0 G_0^{-1} J_0^\top/N = \alpha I$. After all, $\lim_{\lambda\to 0}\theta_\lambda^*$ is equivalent to the NGD solutions θ_∞ (Line 254) and G_0 explicitly appears. Each NGD dynamics converges to different weights depending on G_0. To avoid misunderstanding, we will add the above derivation of the ridge-less limit in the revised manuscript.

Reviewer 5 also gave us a short comment that he/she was unsure whether our work would bring "a huge impact to the research area". This comment seems too general to answer, but we would like to emphasize that our work gives many strengths as other reviewers highly evaluated in their reviews. Finally, we appreciate your constructive questions and hope that our answers will resolve your confusion and lead to your correct judgment.

To Reviewer 1: Thank you for your positive feedbacks. They are very helpful in enriching our paper. We agree that we should more explicitly discuss the justification of the gradient independence assumption. We will move the discussion on it (Line 679-686) to the main body, and remark that this assumption has been justified in some limiting cases, and such justification may be applicable to our case. We will also add minor additional information and modification corresponding to all of your comments.

To Reviewer 2: Thank you for your positive feedbacks and constructive suggestions! We agree that extending our work to finite width will be an exciting direction. We expect that follow-up works will explore more intensive research on the finite width by leveraging the current study. Related to your interest in the inductive bias, our reply to Reviewer 5 (ii) may be informative.

To Reviewer 4: Thank you for your positive feedbacks and for greatly acknowledging the significance of our work. As you recommend, we will make our Python codes used to produce all of the experimental results available. We agree that it will be exciting to invent NGDs with novel FIM approximation satisfying the isotropic condition. We hope that our paper will encourage many researchers to openly discuss and study such algorithms in follow-up works. In particular, it may be interesting to divide each weight vector of units and use corresponding smaller blocks. We will also add more discussion on our assumptions. For example, we move the validity of the gradient independence assumption remarked in Line 679-686 to the main text. The NTK theory requires $||x_n||_2 = 1$, but it is very realistic because one can easily achieve this just by normalizing each sample.