We thank all the reviewers for their detailed and helpful comments.

Response to Reviewers 1&4: Is the assumption $D_{\mathrm{KL}}(p_{\mathrm{true}}||p_0) < +\infty$ in Thm 4.5 attainable?

R: We would like to clarify that the definition of p_{true} is to satisfy $y = \int uh(\boldsymbol{\theta}, \mathbf{x})p_{\mathrm{true}}(\boldsymbol{\theta}, u)d\boldsymbol{\theta}du$, which is irrelevant to α. In comparison, p_t is the parameter distribution for the neural network defined in (3.1) with the scaling factor α. So the KL-divergence bound on p_t in Theorem 4.4 does not contradict with the existence of p_{true}. This assumption on p_{true} essentially assumes that the target function is in the very big function class $\mathcal{F} = \{f(\mathbf{x}) = \int uh(\boldsymbol{\theta}, \mathbf{x})p_{\mathrm{true}}(\boldsymbol{\theta}, u)d\boldsymbol{\theta}du, D_{\chi^2}(p_{\mathrm{true}}||p_0) < +\infty\}$, and therefore it is attainable. Note that this type of assumption on the target function is inevitable: Without any target function assumptions, the random label case is not excluded, and for random labels small test error is impossible. We will add more discussion in the camera ready.

Response to Reviewer 1:

Q1: The minimal eigenvalue of the NTK Gram matrix affacts trainable and generalization properties

R1: Thank you for pointing out the related work [S1]. We will cite this paper and add more discussion on the rate of the smallest eigenvalue of NTK Gram matrix in the camera ready. λ_0 indeed depends on n, which is consistent with existing NTK literature. However, our theorem assumptions are all still attainable in this setting.

Q2(a): (4.1) in Thm 4.4 cannot cover the mean filed case $\alpha = 1$. **(b):** The KL bound in Thm 4.4 increases with n.

R2: We believe both questions are caused by a misunderstanding on the scaling factor α. We would like to clarify that α is not $O(1)$. Instead, in all of our main results (Theorems 4.4 and 4.5), $\alpha = \mathrm{poly(n)}$. In other words, Theorem 4.4 is not supposed to cover the mean field case $\alpha = 1$ at all. Also, because of $\alpha = \mathrm{poly(n)}$, the KL bound in Theorem 4.4 will not increase with n. A simple way to parse and understand our results is to make an analogy between α and the square root of network width $\sqrt{m}$ in the standard NTK literature.

Response to Reviewer 2:

Q1: Explain how and why noisy gradient and regularizers can not be handled by standard NTK analysis

R1: Thank you for your suggestion. We will add a section to explain this claim in detail. Here we provide a short explanation. In noisy gradient descent, the weight decay regularizer pushes the weights towards zero, and gradient noises further push the weights towards a random direction. Therefore, they jointly make each weight fairly far away from initialization. However, the joint effect of weight decay and gradient noise does not push the distribution far away from initialization, as they together give a KL-divergence regularization in the energy functional.

Q2: The scaling factor should appear in the definition of tangent kernel

R2: Our definition of the NTK is correct and consistent with existing results. Our definition matches the definition in equation (16) in [25], where a similar large scaling factor is also considered.

Q3: Should specify in what sense does "parameters stay close to initialization".

R3: Here by "parameters stay close to initialization" we mean the "node-wise" ℓ_2-norm distance. Thanks for pointing out the related work. We will comment on it in the camera ready.

Q4: Do the results still hold, if the gradient noise and regularizer are controlled by different coefficient?

R4: When the gradient noise and regularizer scales are different, the corresponding regularizer on distribution is no longer on the KL-divergence towards initialization distribution p_0, but is towards some different Gaussian distribution $\widetilde{p}$. Therefore, this setting is likely different from the NTK regime, and our linear convergence result may no longer hold. Nevertheless, our generalization results can easily cover this setting by assuming $D_{\mathrm{KL}}(p_{\mathrm{true}}||\widetilde{p}) < +\infty$.

Q5: How does the scaling factor alpha affect the results? Why does scaling factor matter? From Eq. (4.1), it seems that for smaller alpha the theorem does not hold. A discussion of the order of scaling factor alpha is preferred

R5: As we discussed around line 94, α corresponds to the square root of network width in the standard NTK regime, and therefore condition (4.1) is the counterpart of the network width requirement $m \geqslant \mathrm{poly}(n)$ in standard NTK-type optimization results [2,14,15,35]. Therefore, requiring a large α is natural to ensure that the training of the network is in the NTK regime (or lazy training regime), and the setting with smaller α is not the focus of this paper. We will clarify it and provide the specific order of α in the camera ready.

Response to Reviewer 3:

Q1: There is no finite bound on the number of neurons.

R1: By using similar techniques as in [25,26], we are able to study how the training of a finitely wide network can be approximated by the PDE (3.4) in a bounded time interval $[0, T]$. We will add more discussion in the camera ready.

Response to Reviewer 4:

Q1: motivation of regularization, regularization might not be necessary for over-parameterized model

R1: Weight decay regularization is a widely used regularization in deep learning practice, and therefore we believe it is important to establish theoretical guarantees that cover weight decay. It is true that generalization bounds can be developed even without the use of regularizers. This is mainly due to the study of the implicit regularization induced by training algorithms. However, the study of explicit regularization is still an important problem, as explicit regularization can still affect generalization in a different way compared with implicit regularization. See e.g., [33].

Q2: Discuss on the relation to generalization bounds for SGD

R2: Our generalization bound is in the probability measure space. In comparison, existing generalization bounds for SGD are in parameter space, which is not applicable in our setting.