Response for paper ID 8808

We thank all reviewers for their thoughtful feedback. Please find detailed responses to your comments below.

Rev1:

Thank you very much for carefully reading our paper and your supportive comments.

- **Unenglish sentences that could be cleaned up:** We will make the manuscript proofread by a professional English editing service, which we believe resolves grammatical issues.

Rev2:

Thank you very much for carefully reading our paper and your supportive comments.

- **Theorem 2 is a lot to unpack. Asumptions 1, 2 should be made easier:** The analysis tends to be complicated because the infinite dimensional Langevin dynamics requires a few involved conditions such as smoothness on the objective functions. We are doing our best to make the results presented as intuitively as possible. We will add more digested expositions to the assumptions and convergence rate analysis.
- **Application to non-residual deep network. It is important to discuss this:** Thank you very much for pointing our an important point. Indeed, our approach *can* be applied to non-residual deep network. The reason why we presented ResNet is just due to space limitation and the fact that ResNet has a continuous-depth representation and such a continuous depth representation is also an interesting application of our analysis. Since we had this application in our mind, we have presented only ResNet in the main text. We will add some more comments about this point in the final version.
- **I would encourage the authors to improve the exposition:** Thank you for your suggestive comment. We tried to keep technical details as much as possible so that there does not occur confusion and misunderstanding. However, as you pointed out, we would like to use more spaces for intuitive expositions and move some details to the appendix.

Rev3:

- **Is it possible to derive a high probability bound on the training loss of one training trajectory?:** Yes, the most direct way is to apply the Markov's inequality. Moreover, the expectation with respect to the training sample observation is derived from a exponential tail probability bound and thus we can derive a high probability bound with respect to sample observation. As for the training trajectory, the mixing time of the dynamics is fast (exponential with respect to the iteration) and thus as the iteration number increases, the probability in which the trajectory does not contain a "nice" solution satisfying the risk bound decreases exponentially to 0. Since the high probability bound makes the statements complicated (the technical contents are already a bit involved), we have shown the expectation bound for simplicity. We will add more comments on the high probability bound in the final version.
- **How is ρ_0 reflected in your results on excessive risk?:** As the algorithm progresses, the solution "forgets" the initial solution exponentially fast. In that sense, it does not affect so much on the excess risk. On the other hand, the concentration function is characterized by the relative location between the optimal solution and $\mathcal{H}_K$, the geometry of $L_2(\rho_0)$ indirectly affects the excess risk through the shape of $\mathcal{H}_K$. However, it is highly problem dependent.
- **Assumption 1 (ii), (iii) :** The two layer neural network model presented in the excess risk bound satisfies these conditions (Eq.(8) with bounded input $\|x\| \leq D$ (a.s.) and smooth loss). More specifically, under the setting of Theorem 2, Assumption 1 is satisfied.
- **ResNet in line 188-192 is strange:** We would like to remark that this is a standard definition where the residual blocks are two layer neural networks. Each layer ℓ receives an output from the previous layer as x_ℓ and it outputs $x_\ell + g_\ell(x_\ell) = (\mathrm{I} + g_\ell(\cdot))x_\ell$ to the next layer where g_ℓ is a two layer neural network given as $g_\ell(x) = \int a_{w,\ell}\sigma(W(w,\ell)^\top x)\mathrm{d}\rho_0(w)$. This formulation is standard in theoretical analyses of ResNet (e.g., [1, 2]).

Rev4:

- **Eigen decay may be a strong assumption since it directly helps to erase the dependence of the dimension for the bound. With regularizer, it is always convenient to obtain certain tight generalization bounds:** Indeed, as you pointed out, regularization is the most essential ingredient to obtain a width free generalization bound. Conversely, we can not expect nice generalization without any regularization. Although a global optimal solution for non-convex loss does not directly indicate good generalization, our analysis connects generalization and algorithmic convergence, which we believe is an interesting point. In a real deep learning, we consider that such a regularization is imposed through several explicit/implicit regularizations.

References

[1] Y. Lu, C. Ma, Y. Lu, J. Lu, and L. Ying. A mean-field analysis of deep ResNet and beyond: Towards provable optimization via overparameterization from depth. In *Proceedings of ICML2020*, pages 137–147, 2020.

[2] E. Weinan, J. Han, and Q. Li. A mean-field optimal control formulation of deep learning. *Research in the Mathematical Sciences*, 6(1):10, 2019.