We thank reviewers for their thorough reading. We will fix the typos and clarify the unclear points in the next version of our paper.

On the Batch Size. Reviewer #2 and #4 concern about the batch size. Batch size has been an important component of past analyses. For nets with BN the direct relationship between noise and batch size mentioned by Reviewer #2 does not hold (Though there were attempts to make that theory fit using Ghost BN). In our theory the noise is captured in covariance matrix (in eqn 9 and 10) which could depend on the batch size. When the nets are without BN, e.g. with LN or GN, the magnitude and trace of the covariance scales inversely proportional to the batch size B. Thus for SDE, intrinsic LR λ_e and the batch size B can be further grouped into λ_e/B, which alone controls eqn 9 and 10. However, this analysis doesn't hold for the general case where BN is allowed and thus we treat batch size as a fixed hyper-parameter like width and depth. We will clarify the connection to batch size and point out that our theory could automatically include the effect of batch size when applicable, such as when Layer Norm and Group Norm is used.

On the benefits of BN. Regarding Comment 3,4 of Reviewer #2 and Comment 1 of Reviewer #5, we want to clarify we do not claim that our theory covers all the benefits of BN. The fast equilibrium conjecture only partially explains the benefits of BN. Besides this conjecture, there are many other benefits, e.g., BN affects the signal propagation at initialization, BN induces noise when calculating the batch statistics, and our theory clearly does not capture these benefits.

To Reviewer #1 @ Experimental verification for the two predictions: 1. In Figure 7(b), the dotted red line (1/effective lr) increases by 10 times and then drops by approximately $\sqrt{10}$ slowly in 40 epochs. If we make the second phase longer, one should expect the ratio becomes closer to $\sqrt{10}$. 2. Figure 10 gives a more clear and direct justification for the claim that reaching equilibrium takes $O(1/\lambda_e)$ steps.

@ **Q1: What if the noise is heavy-tailed?** If the gradient has unbounded second moment, then we should see the weight norm has a sudden huge increase in practice, though with small probability. However, this is not observed in any of our settings, so it's not clear to us whether the heavy tail assumption holds for our setting. Moreover, the recent work (`https://arxiv.org/abs/1907.03215`) suggests that when the learning rate is sufficiently small (smaller than some practical constant), the performance of neural nets only depends on the continuous-time covariance.

@ **Q2: Do our experiments still hold if large depth causes gradient explosion at initialization?** Our theory still holds. The gradient explosion at initialization leads to huge weight norm, which makes effective lr tiny and the nets untrainable in Yang et al., since they don't have WD, the weight norm can only become even larger. However, in our experiments, the huge weight norm led by gradient explosion could be fixed by WD and the curves of 1/effective LR again matches our theoretic prediction, i.e. dropping slowly after the sudden increase. We ran the experiments but due to the space limit, we cannot include the figure in the rebuttal.

@ **Q3: What happens if we drop the LR before equilibrium?** The answer depends on whether or not this is the final LR decay. In the earlier phases, decaying too soon (i.e. before equilibrium) will increase convergence time in next phase, but similar test error is reached. For final LR decay, if it happens too soon then the test error is hurt. A good example is the blue curve in Figure 1,(a)-(c). If the lr decay happens within normal training budget, then the blue curve will get much lower test acc. That's why people usually think tiny learning rate generalizes poorly.

@ **Q4: Is small LR or large LR preferable?** Small LR is preferable under the practically feasible regime. See Figure 1(d). It's not clear when the lr goes to 0 since training time becomes too large.

@ **Q5: How about continuously-varying LR schedule?** Our theory doesn't analyse continuously-varying lr schedules and we leave this for future work. We suspect equilibrium does not exist in such settings and thus the analysis should be more complicated.

@ **Q6: What is the "number of trials" in Figure 2(b)?** We repeat the training process 500 times in Figure 2, and we call each run as a "trial".

@ **Q7: How about Momentum?** Yes. When discussing the relation between our analysis and the practice, we switch from SGD to Momentum in the experiments (Figure 1). We also discuss the extension of this conjecture to momentum formally in Section C.1. (BTW we don't quite claim λ_e determines everything but that it explains a lot.)

@ **Q8: Why do weight norms diverge from each other in Fig 4(b)?** According to our theoretical predictions, the two training processes in Figure 4 should have the same effective learning rate ($\eta/\|\mathbf{w}_t\|^2$). Since they have different learning rates (η), their norms (the solid lines) should diverge from each other.

To Reviewer #2 @ Comment 5: Time rescaling? It does not make a difference. Note that the phenomenon is described jointly by eqn 9 and 10. The different scaling suggested above would still capture the same phenomenon but would not highlight at a glance that the evolution depends only on intrinsic LR.

@ **Comment 6: We seemed to complain about gaussian noise assumptions but used it anyway.** We show mathematically the noise is not an isotropic Gaussian: in fact it's position-dependent and perpendicular to the current weight. It wasn't meant as criticism of use of Gaussians per se in modeling.

@ **Comment 7: Explain more about GD $\neq$ GF.** Our point is that GD $\neq$ GF is unavoidable for scale-invariant networks with WD. This chaotic behavior is not limited to the toy example. See Figure 5(b),(c) in the appendix for the experiments on small CIFAR.

@ **Comment 8: Explain more about SWA.** Reviewer is referring to standard SWA which fixes initialization and seems to suggest that at the end SGD is fluctuating around a local minimum (loosely speaking). Our experiment around line 164-165 is trying to answer the question: "Is this equilibrium independent of initialization?" The fact that SWA over solutions obtained from different initializations harms the test accuracy suggests that the answer is no. We will make this more clear in the future version.

@ **Comment 9: Changing initial LR = changing initialization scale?** Lemma 2.4 in (Li & Arora, 2020) [28] formally proves that changing LR is equivalent to changing the scale of the norm for SGD and scale invariant objective. In other words, shrinking norm will not change the network in function space, and it changes the future networks after SGD updates in the same way as increasing LR does. Related experiments could also be found in (Li & Arora, 2020) [28], such as training neural nets with exponential increasing LR schedule could lead to the same trajectory of parameters in direction, while the norm of parameters is exponentially larger.

@ **Comment 11: Gradient flow vs finite LR?** We think this refers to text around 252. We are not proposing a 2-step schedule there. The schedule could have many decay steps; the only point being made there is that at the end training one needs a *few* epochs of very small learning rate at the end to reach best accuracy. The final stage does not go to equilibrium and it is customary to turn off WD. In the SDE view, turning off WD is like zeroing the effective LR. That is the sense in which it is like gradient flow. (Another evidence is that the norm does not change much during these few epochs; see Lemma 5.2 For relation between norm growth and effective LR.)

To Reviewer #4 @ About the storyline. We will revise our paper to improve the presentation of our main storyline and figures. Our storyline is cohesive: we use a suitable SDE to reconcile the modern neural networks with BN + WD and traditional optimization analysis. This yields a new parameter intrinsic LR, a challenge to the benefit of initial large LR, and the fast equilibrium conjecture.

To Reviewer #5 @ Comment 3: What is the range of WD that leads to no mixing in parameter space? As long as WD is non-zero, the norm is not diverging, and thus the training process is likely to mix in parameter space. However, it may take exponential time. The conjecture in our paper concerns the ability of SGD to mix in functional space in polynomial time.