We thank three reviewers for admitting the importance of our work. We hope that the following explanation will help the reviewers to recognize the novelty and significance of our paper further.

To Reviewer 1: The minor points you provided are also of our concern and we will remark them in Discussion. Although the empirical Fisher is not necessarily representative of the geometry (especially, when the number of samples is very small), it always determines NTK's eigenvalues through a dual representation F^*. The NTK determines the optimization of sufficiently wide DNNs. We appreciate if you would read our responses to Reviewer 2 and 3. The experiment shown in the below will also respond to your concern on trained networks.

To Reviewers 2 & 3: Difference of this paper from Karakida, Akaho & Amari, AISTATS2019 [21]

We would like to emphasize that our work greatly differs from [21]. [21] has **not** shown any result on normalization methods. Moreover, our work is not just a simple re-application of calculation in [21]. Technically speaking, [21] evaluated the first-order term of F^* and neglected lower order term (i.e., the second term in (S.6)). In contrast, our paper enables researchers to evaluate this second term. It is essential because the batch normalization makes the first term comparable to the second term and requires careful evaluation of the second term. In particular, we found that a new quantity, i.e., the convergence rate q (or q^*), plays an essential role in the second term and newly developed a framework to evaluate it in Section B.1.3. This enlightened the new direction of theory and enabled us to give the novel insight into the use of normalization methods.

To Reviewer 2: Experiments on training DNNs and learning rates necessary for convergence

We agree that experiments will further increase our contribution. As Reviewer 2 recommended, we add an experimental result on the training with the steepest gradient descent argued in Section 5. We did it in the same setting as [21]; we trained DNNs with various widths by using various fixed learning rates, providing i.i.d. Gaussian input samples and labels generated by corresponding teacher networks. Our Fig. (a) is just a reproduction of Fig. 2 (left) in [21]. The theoretical value $\eta = 2/\lambda_{max}$ (Eq.27) computed on the FIM at random initialization predicted well the learning rate necessary for the gradient method to converge. An impressive result is Fig. (b). We confirmed that the batch normalization (mean subtraction) in the last layer allows larger learning rates for convergence and they are independent of width. This result coincides well with Reviewer 2 expectation. Technically speaking, we computed the red line in Fig. (b) by using the lower bound of λ_{max}, i.e., $\eta = 2/(\rho\alpha(\kappa_1 - \kappa_2))$.

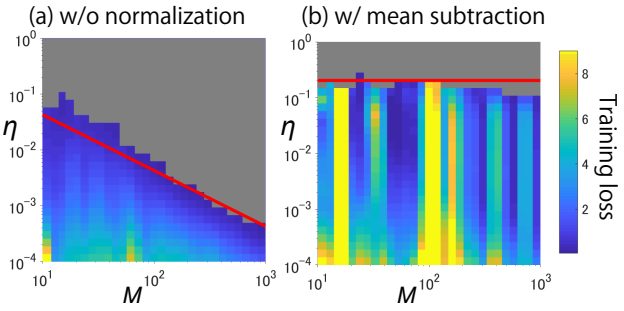

Figure 1: Exhaustively searched training losses depending on M(width) and η (learning rate). We trained deep ReLU networks for 1000 steps. Losses **exploded in gray area** (i.e., were larger than 10^3) and red lines show theoretical values of $2/\lambda_{max}$. Experimental setting: $\alpha_l = C = 1$, $L = 3$, $T = 1000$, $(\sigma_w^2, \sigma_b^2) = (4, 1)$.

One can also suppose many other experiments related to our theory, but they are too many to enclose in a single paper. Besides, experimental studies in more large-scale networks and datasets are not so easy task because of the computational cost of the huge FIM. So, we expect that our theory and the above experiment will encourage many researchers openly discuss and study the possibility of our results in follow-up works.

To Reviewer 3: The NTK (neural tangent kernel) determines optimization and loss landscape in wide DNNs

Answering to your concern, we would like to emphasize recent findings on NTK shown in Jacot et.al., NeurIPS2018 (cited as [19]) and Lee et al. arXiv2019 (cited as [20]). In particular, the work [20] clearly proved that the sufficiently wide DNNs works as a linear model expanded around random initialization θ_0:

$$f(x;\theta_t) = f(x;\theta_0) + \nabla_\theta f(x;\theta_0)^\top \omega_t, \tag{1}$$

where $\omega_t := \theta_t - \theta_0$ and t means the step of the gradient descent shown in lines 284-288. [20] proved a surprising fact that ω_t is sufficiently small for any $t > 0$ and the network can achieve a zero training error in the large M limit. This means that there is always a global minimum sufficiently close to random initialization. Therefore, the optimization of the wide DNN becomes a convex problem and the loss landscape becomes convex. This convexity is also proved in [19] and the FIM at random initialization determines the loss landscape through a quadratic form $\omega_t^\top F \omega_t$. Thus, its second derivative (Hessian) coincides with the FIM in the large M limit. We hope that the above additional explanations will further clarify our paper's significance, and we will appreciate if you could increase your score.