The authors sincerely thank all the reviewers for their enormous effort and constructive comments. We will thoroughly review all related works, avoid typos, and carefully address all concerns in the final version.

To Reviewer #1:

Q1: *The work "Don't Decay the Learning Rate, Increase the Batch Size" suggested opposite theories.*

A1: We respectfully argue that the mentioned paper exactly supports our theory. This paper empirically demonstrates that when the rate of batch size to learning rate remains the same, the generalization is invariant, which is theoretically validated by our work.

Q2: *More varieties in model/data are needed.*

A2: We conducted more experiments of MLP on MNIST, AlexNet on CIFAR-10/100, and ResNet on ImageNet, during the rebuttal stage. All obtained results are consistent with the proposed theory and will be added to the final version.

Q3: *Actual accuracy after changing protocols. A reduced generalization gap does not grant a better test result.*

A3: We respectfully argue that the reported results are based on the test performance. Also, the highest accuracy is the actual accuracy of the models after changing protocols.

Q4: *The effect of disabling momentum and constant batch size and learning rate.*

A4: The effects of momentum and adaptive batch size and learning rate have been sufficiently studied (although only from the empirical view). By contrast, our paper studies the relationship between the generalization ability and the rate of learning rate to batch size from both theoretical and empirical aspects. It will be interesting to develop theories for momentum and adaptive batch size and learning rate. However, this is beyond the scope of this paper.

To Reviewer #2:

Q5: *The assumptions made to obtain this bound are not convincing.*

A5: Justifications for the assumptions are given below: (1) the gradients on one single data point, a mini-batch, and the whole training set are assumed to be unbiased estimations of the expected gradient, because all data points are independently drawn from the same distribution; (2) applying the large number theorem, it is rigorously correct that the gradient of one single data point is Gaussian distributed when the training set size is sufficiently large; and (3) existing experiments have demonstrated that the loss surface around the local minima is two-order smooth (cf. Appendix A.1).

Q6: *The relationship between generalization and the rate seems to be more complex than the authors suggest.*

A6: There are two terms in our bound which have positive and negative correlations with the rate of batch size to learning rate, respectively. However, we can prove that the positive correlation is dominant, when the parameter size d is sufficiently large ($d > (|S|/\eta)/2\text{Tr}(A)$). The condition has been empirically validated in [22, 25, 26] for deep neural networks.

Q7: *The Hessian matrix changes drastically during learning, and it's not obvious how this fact changes the conclusions.*

A7: We respectfully argue that the Hessian matrix A of the loss surface around global minima is invariant during learning. It only relies on the neural network architecture and the data distribution.

Q8: *The experimental result is not very surprising.*

A8: We acknowledge that some existing results suggested our finding, which are however only from the empirical aspect. Theoretically, the understanding is still elusive. Our contributions validate the intuitions theoretically and empirically. Moreover, our empirical studies are more comprehensive.

Q9: *PCC is only for linear models and the p-values are overkill.*

A9: We agree that PCC is designed for linear models, while it can still measure correlations. We further perform Spearman correlation test on the collected data, which is not based on any model. The results are in full agreements with our theory and will be added to the final version. Further, we respectfully argue that p-value is still important and even irreplaceable to measure the likelihood of an argument. The articles in *Nature* and *The American Statistician* concern the abuse of p-value that strictly uses $p = 0.05$ as the threshold of "statistically significant", especially when the sample size is small (e.g., smaller than 100). By contrast, the p-values in our paper are far below 0.05 and the sample size is sufficiently large. Specifically, the p-values for the rate are smaller than 10^{-88} and the sample size is $1,600$.

To Reviewer #3:

Thank you very much for your positive support to our work. We will thoroughly revise our paper to review all related works and to avoid typos.