[ID 3407] We thank reviewers for their efforts and giving us valuable comments. *Q.N*[-*M*] is a response to Reviewer N.

Q.1-1 It is unclear whether "depth" refers to L or T. In either case, the claim "test error bound monotonically decreasing with depth" is inappropriate. A. The depth (e.g., in L.42) meant T. We considered the dependence on T rather than L because we are interested in the over-smoothing caused by node aggregations. [Depth=L case] It is true that the bound can exponentially depend on L. Since this problem occurs in inductive MLPs, too, many studies derived generalization bounds that avoid this exponential dependency [Arora+ICML18; Nagarajan+ICLR19; Wei+NeuIPS19]. We can incorporate them to obtain tighter bounds (c.f., Remark 2). [Depth=T case] See Q.1-2. We admit that the expression L.46–47 was confusing because it is not clear that depth means T, and this bound needs not only w.l.c. but also the condition in L.213. We will address this problem in the updated version.

Q.1-2 Why did you think of a liner model as $\mathcal{G}^{(t)}$? What is the superiority of linear $\mathcal{G}^{(t)}$ over non-linear one? A. GNNs that consist of linear node aggregations and non-linear MLPs is one trend in the GNN research. For example, SGC [69], gfNN [NT+19(arXiv:1905.09550)], and APPNP [41] are such examples. Theoretically, [NT+19] and [51] claimed that non-linearity between aggregations is not essential for predictive performance. So, we believe that such models are worth investigating. [Superiority] Adding non-linearity changes two things. First, $\|P^{(t)}X\|_{\mathrm{F}}$ in (5) is replaced with $\prod_{s=2}^{t}\|\tilde{P}^{(s)}\|_{\mathrm{op}}\|X\|_{\mathrm{F}}$. It makes the interpretation of L.222–234 impossible, and the bound looser, essentially because the bound loses the information of eigenvectors. Second, the bound of Rademacher complexity of $\mathcal{F}^{(t)}$ is multiplied by 2^t (not 2^T). It changes the condition in L.213 to a stricter one $\alpha_t^{-1}2^tD^{(t)}\prod_{s=2}^{t}\|\tilde{P}^{(s)}\|_{\mathrm{op}} = O(\tilde{\varepsilon}^t)$. Nevertheless, we do not have a definitive answer whether "linear" GNNs are truly superior to "non-linear" ones. We may be able to use techniques similar to [51] for the first problem. Refined analyses could eliminate the 2^t term for the second problem.

Q.2 The paper is extremely hard to read because there is very little empirical evidence or clean argumentation that would motivate the reader to follows the suggested path. A. Sec.1 Par.2 and Sec.2 Par.2 correspond to the empirical superiority of multi-scale GNNs, and Sec.1 Par.3 and Sec.2 Par.3 to the motivation for using the boosting interpretation and learning theory. We are sorry that the paper has unclear points. However, we believe our paper is well-written, as other reviewers evaluated the clarity of our paper. We want the reviewer to reread it and reconsider the evaluation.

Q.3-1 The number of weights and hyperparameters could be large for large T and L [3-1]. A. For hyperparameters, we used the same hyperparameter set for every weak learner in experiments. The number of hyperparameters is indep. of T for this setting. The accuracy is as high as existing methods, even though such a simplification. For weights, $L = 1$ was enough for empirical performance. The number of weights in that model is comparable to a standard $3T$-layered MLP.

Q.3-2 The GB-GNN model does not outperform many competitors in experiments [3-2]. Besides, one of the models fails to run properly, even for the standard datasets [3-3]. A. We put importance on the consistency of theoretical and empirical behaviors, rather than achieving SOTA performances. Observing that many of the SOTA methods have little performance difference in benchmark datasets, we might almost reach the performance limit. Such quality issues of standard benchmarks are a major challenge in the GNN community and motivate recent benchmarking researches (e.g., Open Graph Benchmark, [Errica+ICLR20], and [Dwivedi+20(arXiv2003.00982)]). Considering the current situation and the theoretical nature of this paper, we think that accuracy comparable to existing methods is sufficient to guarantee our method's correctness. Regarding the failure, there are two reasons. First, our implementation naively processes all nodes at once. Second, since we train the fine-tuning model in an end-to-end manner, it uses memory proportional to T. These problems are not specific to our model but common to end-to-end deep GNN models. We can solve them by mini-batching. Also, we note that when training without fine-tuning, memory usage is constant w.r.t. T, since we do not have to retain intermediate weights and outputs. This memory-efficiency is an advantage of the boosting algorithm.

Q.3-3 The w.l.c. params are confusing. How do they influence the model performance, and how are they valued [8-1]? A. Algorithm 1 pre-determines w.l.c. params and train weak learners until the w.l.c. condition is satisfied. However, it is practically more convenient to train weak learners and determine w.l.c. params that they satisfy *a posteriori*. So, we did not evaluate how the change in w.l.c. params affect the performance empirically. Still, we can control w.l.c. params indirectly by the complexity of weak learners (e.g., width, layer size) and check the satisfiability of w.l.c. (c.f., Fig.2).

Q.3-4 In Equation 4, why could applying $\tilde{P}$ exclude neighbouring information (L.42) [8-2]? A. The *representation* in L.42 is the input for $b^{(t)}$ (i.e., not X but $P^{(t)}X$). What we intended was that by assuming the model in Sec.5, we could evaluate the model complexity using inductive models (c.f., Remark 2). We are sorry for the confusion.

Q.3-5 Do the experiment results indicate that a "weaker" learner is favourable than deep "stronger" ones [8-3]? A. We think both "weak" and "strong" learners can be problematic. If we use "strong" learners, we have a risk of over-fitting, especially when T is small. If we use "weak" learners, we can not reduce training loss and cannot make up for it in later iterations using over-smoothed representations. The present results suggest to balance between the two situations.

Q.3-6 Clarification of notations [5-1, 5-2]. A. $\tilde{B}^{(t)}$ is a user-defined constant in parallel to C_l's and L. $\tilde{C}^{(t)}$ is defined in L.201. In equation 4, different from the reviewer's comment, we use different weight matrices for each iteration.