We thank all the reviewers for their detailed comments. Please find our response to the major comments below (we will fix all typos/minor comments in the final version).

Response to Reviewer 1:

`Related work:` We will include comparison with [Fiat et al., 2019] in the main section.

`Feature Learning:` It is true that NPF learning happens when we train standard DNN with ReLU and it is only natural that it closes the gap. However, NPF learning (interpretation) is different from the standard interpretation of feature learning, where, the hidden features are learnt in the penultimate layer and the final layer learns a linear classifier on these features. To see the difference, consider S1={FLNPF, DNPFL, and ReLU-DNN} vs S2 = FRNPF vs S3 = Infinite width CNTK. S1, S2, S3, all of them generalise (S2 is the least with 67%, yet, on a 10 class task like CIFAR-10, is still way better than random classification accuracy of 10%). Both S1 and S2 are finite width, so standard feature learning happens in both S1 and S2, but, S1 with NPF learning is better (78% or above in CIFAR-10) than S2 (67% in CIFAR-10) with no NPF learning. Thus neither finite width, nor the standard feature learning is useful to explain the difference between S1 and S2. S2 and S3, both have no NPF learning, yet, S3 generalises better, which can be attributed to the fact that infinite width ensures better averaging and hence results in a well formed kernel. NPF learning also happens in the DNPFL setting, which is different from ReLU-DNN. To conclude, *NPF learning is a measure which discriminates/describes the different regimes (i.e., S1, S2, S3) better than the standard feature learning explanation*.

Response to Reviewer 2:

`Assumption 5.1 and Goal of the paper:` Most analysis of DNNs with ReLU is on what happens at initialisation. In a DNN with ReLU, NPV and NPF are not statistically independent at initialisation, i.e., Assumption 5.1 does not hold. However, in the current state-of-the-art analysis, in $w \to \infty$ regime, activations change only at rate of $\sqrt{\frac{1}{w}}$ ([Jacot et al., 2019]), i.e., activations/NPFs do not change during training. Hence, though Assumption 5.1 may not hold exactly, it is *not a strong assumption* to fix the NPFs for the purpose of analysis. Thus, statistically decoupling the NPV from NPFs is only natural, and furthermore it adds strength: the NTK = const* NPK is *interpretable* in terms of the active sub-networks (NTK is defined in terms of gradients with no interpretation), which also shows that the active sub-networks are *fundamental entities arising naturally in the NTK framework*.

Fundamental role of gates is further accomplished by the FLNPF experiments, where we show that by copying gating information alone, and resetting and training NPV (i.e., Θ^{v}) from scratch (Assumption 5.1 holds in this case), we can recover the performance of the DNN with ReLU. Further, in the DNPFL setting, (which is not hypothetical) Assumption 5.1 holds, and the DNPFL does generalise well in the experiments.

`Prior experimental work:` The goal of this paper is not to study the utility of active sub-networks as representations, but to directly look at the generalisation capability. However, we will move relevant work (example [Srivastava et al., 2014]) in the appendix to the main body.

`Regularisation:` Similar to [Arora et al., 2019], data-augmentation, batch norm, residual connections, dropout and other forms of regularisations are avoided. However, studying these in our framework is future work.

Response to Reviewer 3:

MNIST and CIFAR-10 are used as standard datasets in most analytical works such as ours, see [Arora et al., 2019] for example.

Response to Reviewer 4:

`Analytical part in supplementary:` The main contribution of the paper is analytical in nature, and aims at providing an understanding into the internal workings of a DNN. We believe that it is critical that the setup and subsequent explanations belong to the main body. While we give details of the experimental setup in the appendix, we will be happy to move the important points in the main body to improve clarity.

`Explanation of generalisation:` By generalisation we mean performance on test data. Agreed that on MNIST, all the cases namely FRNPF, DNPFL, FLNPF and ReLU-DNN have marginal performance difference. However, on CIFAR-10 the difference between FRNPF (67%) and FLNPF, DNPFL, ReLU-DNN (all above 78%) is more than 10%. The crucial insight from this work is that mere gating/masking property is enough to give us 67% (on CIFAR-10, this is non-trivial because a random classifier will only have 10% accuracy), and in addition if the gates also *adapt* during training (as in standard ReLU-DNN) gives the rest 10%. Further, once we have the *learnt* gates, we can reset and learn NPV from scratch without loss in performance. Thus, the experiments were designed to test the role of gating and we believe we have extensive experiments to support our claims.

`Expanded conclusion section:` We will make the summary of main contributions more clearer in the conclusion.