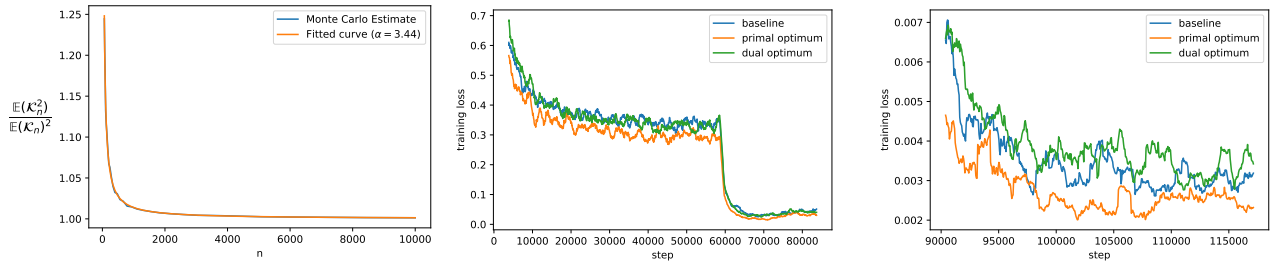


(A) (B) (C)

Figure I: (A) MLP $\alpha$ fitting. (B + C) Training loss curves at early and late stages for optimal primal/dual ResNext models (orange and green) and baseline (blue), trained on CIFAR100 (for details see section 4.2.1 in the main text)

**R1+R2+R3+R4:** We would like to thank the reviewers for their thoughtful comments, suggestions and constructive feedback. Typos and missing citations as pointed out by all reviewers will be addressed in future versions. The main point of contention articulated by all 4 reviewers is the validity of var(K) as a good proxy for generalization and the correctness of Proposition 1 in the general case, and so we will address these in detail. **var(K) as a proxy for generalization:** We agree with the reviewers that this is a strong assumption, and our main hypothesis is indeed to use var(K) as a proxy for *trainability*, as stated in the abstract (line 12) and introduction. As a direct evidence, we additionally show in Fig. I above that a lower training loss is achieved with the primal formulation, and the loss is roughly maintained in the dual formulation. In the ablation study and section 2, we empirically show how lower v(k) translates to better performance in the tested models under practical scenarios. We will eliminate the claims regarding v(k) and generalization in the main text, and instead leave the precise characterization of it as future work. Meanwhile, we mention [A] where a rigorous connection is made between v(k) and generalization (see section 3), advocating for ensembling as variance reducing alternatives. **Validity of Proposition 1:** Additional experiments on its validity (including Fig. I(A) above) will strengthen the paper, and will be added to the appendix. Note that in the context of section 2, its correctness in general is less important than its usefulness. In fact, any formula v(k) = f(n) that can be nicely fitted to Monte Carlo simulations in algorithm 1 can be used to derive optimal ensembles. We find that Proposition 1 holds nicely in general. **R1:Undefined notations.** $K_\infty$ is indeed the infinite width kernel. We will make this explicit in future versions. **R1:Gradient flow assumption.** We agree with the reviewer. However with some additional work the $\mathcal{O}((nm)^{-1})$ results can be derived in the infinite width limit for gradient descent with a small learning rate. We leave this extension to future work. **R2:Ablation study.** The ablation study in the paper is conducted on mnist, however figures 6,7 clearly demonstrate the effect of variance and capacity on larger datasets with modern architectures. Note that we do not make the claim that capacity in itself is harmful in the general case. **R2:Full Imagenet results.** We were unable to produce the full Imagenet results in time for the deadline of the main text, and so they were included in the appendix. In future versions they will be included in the main text. **R3: Connection of var(k) to generalization.** We do not see our results being at odds with previous work since we are not advocating for eliminating variance all together or using small learning rates. (thus eliminating the prospects of representation learning). Please see our above response for further details. **R3:Proposition 1 as an assumption.** We agree with the reviewer and will change its definition to an assumption in future versions. **R3:Lemma 1.** The proof of Lem 1 in the appendix mentions the finiteness of first moments (line 47), and we will add more details on that matter in future versions. **R3+R4:On Lemma 1 and Theorem 1.** The discussion of NTKs in the context of collegial ensemble requires us to expand the notion of the "kernel" regime, since width can be interchanged with multiplicity (number of models in the ensemble), to which section 1 is dedicated. The O(1/m) result in The 1 is non-trivial, implying that large collegial ensembles can approach the kernel regime with linear, instead of quadratic dependence on the number of weights even after training, where the independence assumption of the ensemble models is broken. The novelty here is both technical and in the final result, and we feel will be appreciated by the community. **R4:On $\alpha$ in proposition 1.** For a given architecture, $\alpha$ represents the rate in which v(k) changes with width. Computing its theoretical value for fully connected networks is a technical challenge that is beyond the scope of the paper (see [B]). Empirically, any architecture parameterized by its width can be fitted with $\alpha$ as described in Algorithm 1. **R1+R2+R3+R4:Additional experiments that reviewers may have missed** Our proposed method can be adapted to efficiency metrics other than parameter count. In section E of the appendix we adopt a FLOPS metric instead and show how for a given FLOPS budget, optimally performant ensembles can be derived.

REFERENCES

[A] M. geiger et al: Scaling description of generalization with number of parameters in deep learning. (Journal of Statistical Mechanics Theory and Experiment 2020). [B] B. Hanin et al: Finite Depth and Width Corrections to the Neural Tangent Kernel. (ICLR 2019).