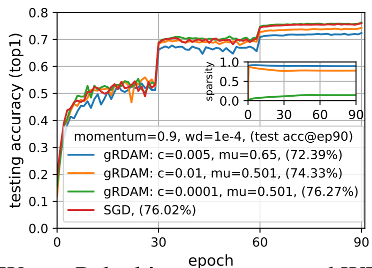
We thank all reviewers for their comments, which are in *italic* below. The number(s) in each item is the reviewer(s) being addressed to. **CR** is the shorthand for camera-ready. Citation number here aligns with the references of the paper.

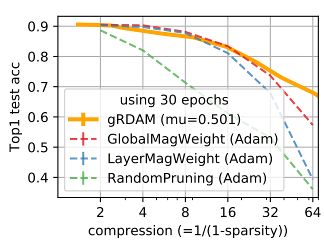
1. *Hyperparameters $c, \mu$ still require tuning (***1,3,4***).* The expert knowledge in our paper refers to knowing a good sparsity level specific to the task and network. Tuning gRDA can be easier than prescribing the level of sparsity (which gRDA does not require) when dealing with a new dataset/network with few prior research to base on. We develop a rule of thumb to select $c, \mu$ (not in the original gRDA paper [8]). Our empirical results and theory suggest $\mu \in \{0.501, 0.51, 0.55\}$ generally performs well regardless of the task and network used. For a given $\mu$, we search for the greatest $c$ (starting with e.g. $10^{-4}$) such that gRDA yields a comparable test acc. as SGD using $1-5$ epochs.




2. *gRDA does not perform well under the PyTorch learning rate (lr) schedule; adding momentum and weight decay (WD) (***2,3,4***).* The PyTorch lr schedule for the ImageNet is usually applied jointly with the SGD with Polyak's momentum. Without it, SGD only yields a test accuracy of 68.76% for ImageNet-ResNet50 under this lr schedule. The current gRDA has no momentum, so its performance in the upper panel of Table 1 seems a little disappointing despite the extremely high sparsity (will revise Sect. 4.1 to include this). The absence of momentum could also be a reason for the underperformance at the middle to high sparsity level in Figure 4 (R3,R4). R4 suggests to include momentum and WD in gRDA. We try Polyak's momentum and WD with the original $g(n,\gamma)$, termed gRDAM, on ResNet50 with ImageNet (figure on left). While the sparsity reaches 88% and the test acc. improves, the test acc. of gRDAM still generally jumps less than SGD with momentum when the lr drops. However, gRDAM with $c = 10^{-4}$ yields a higher test acc. than SGD but with only 14.25% sparsity. Indeed, there is still much room for improvement which calls for a new theory to find a proper $g(n,\gamma)$. Other momentum, e.g. Nesterov and QHM (Ma and Yarat, 2019, ICLR) can also be considered.

3. *Contributions (***1,2***).* We provide the first systematic study on the effectiveness of gRDA for pruning modern DNNs on large-scale tasks, while the original gRDA paper [8] has not focused on deep learning. Inspired by the good empirical performance of gRDA, we theoretically study gRDA and discover that it asymptotically performs the directional pruning (DP, see below for a comparison to OBS), which is empirically verified by the connectivity (Sect. 4.2) and subspace restriction (Sect. 4.3). This justifies a unified view of gRDA and DP. See 1. for a response on the expert knowledge.

4. *The shape of $\mathcal{P}_0$ and a comparison with the "optimal brain surgeon"(OBS)(***2***).* We will revise Sect. 1.1 to define $\mathcal{P}_0$ as the eigenspace corresponding to the zero eigenvalues of the Hessian $H(\mathbf{w}(\infty)) = \nabla^2 \ell(\mathbf{w}(\infty))$ where $\mathbf{w}(t)$ is the gradient flow and $\mathbf{w}(\infty) = \lim_{t\to\infty} \mathbf{w}(t)$ that achieves the minimum (under weak conditions) where flat directions exist under overparameterization. This $\mathcal{P}_0$ is the eigenspace of zero eigenvalues of $\bar{H}$ as in (A3). Perturbation from $\mathbf{w}^{SGD}$ along $\mathcal{P}_0$ causes little changes to loss $\ell$ if $\mathbf{w}^{SGD}$ reaches the same minimum valley as $\mathbf{w}(\infty)$, which holds under a small learning rate. An analytic map between DP and OBS is interesting for future study, and we believe the two are generally nonequivalent. Particularly, DP perturbs from $\mathbf{w}^{SGD}$ continuously in $\lambda$ like a restricted $\ell_1$ weight decay on $\mathcal{P}_0$ (Remark 1.2), while OBS yields a discontinuous perturbation like a hard thresholding (see OBS, p.165 [29]).




5. *Re-training and pre-training (***2,3,4***).* We agree with R2 and R3 and will revise the tone about re-training in the CR. For R3's question, directional pruning (DP) does not require pre-training with SGD, as gRDA achieves that in one shot training from scratch (shown in Eq. (8) in Thm 1). Although generally not recommended, gRDA can be implemented on the pre-trained models (R4). For an illustration, we re-train ResNet20 on CIFAR-10 using the gRDAM on a pre-trained model (ShrinkBench by Blalock et al., arXiv:2003.03033) and compare with their Fig. 11 using their codes and setting. The results (on the left) show that gRDAM outperforms several magnitude pruning based methods under high compression level. To further improve, we should modify $g(n,\gamma)$; see 2.

6. *Previous work of gRDA and incorrect/missing references (***1,2,3***).* We will be more careful on referring gRDA; e.g, in Line 97, 234 (R1) and 109 (R2). Sect. 1 will be revised to discuss the gRDA (R2,R3). The [46] in Line 221 should have been Papyan (2018, arXiv:1811.07062) (R1); Izmailov et al. (2018, UAI) will be cited in Line 210 in the CR (R2).

7. *Notational confusions and $c = 0$ in gRDA (***3***).* Note $\ell(w) = N^{-1}\sum_{i=1}^{N} \mathcal{L}(h(X_i;\mathbf{w}), Y_i)$, where $(X_i, Y_i)$'s are the training data, so $\ell(\mathbf{w})$ and $\mathcal{L}$ are different. As $c = 0$, gRDA is unpenalized, so it reduces to SGD (Eq. (8) in Thm 1).

8. *Principle to stop training (***4***).* We stop training when the test acc stabilizes. While the sparsity usually also stabilizes at a high level with the test acc like in CIFAR-10/100, sometimes it can slowly decrease like in ImageNet-RN50 in Figure 3. However, given that the level of sparsity in Figure 3 is still high, the concern should be minor.

9. *Additional comments of R3 (***3***).* The main challenge in implementation is tuning $c, \mu$ as addressed in point 1. For a small network like ResNet20 on CIFAR-10, gRDA with $(c,\mu) = (0.01, 0.7)$ achieves test acc/sparsity 90.19/90.46%, while SGD (w/o momentum & WD) achieves a test acc. of 89.12%, so it is still overparameterized. We leave a decent comparison of pruning methods to future study due to its independent interest (Blalock et al., 2020, arXiv:2003.03033).