To Reviewers 1, 3 and 4 for Results on **Normal Training**

Sec. 3.2 has stated $\|\mathbf{A}\|_* = \|\boldsymbol{\sigma} \circ \boldsymbol{\sigma} - \mathbf{1}\|_1$ allows larger singular values for dominant singular vectors of $\mathbf{W} \in \mathbb{R}^{m \times n}$. This advantage holds when $\mathbf{W}$ in $\mathbf{A}$ is initialized with $\mathbf{W}$'s singular values $\boldsymbol{\sigma}$ being larger than 1. In the originally submitted paper and Github code, only WideResNet28-10 initializes W with scaled Gaussian values $\mathcal{N}(0, \sqrt{2/m})$ to make $\mathbf{W}^T\mathbf{W} \approx 2\mathbf{I}_n$ so that $\boldsymbol{\sigma} \approx \sqrt{2} > 1$ in the beginning of the training. However, the other two networks still initialize $\mathbf{W}$ by standard Gaussian values, i.e., “Kaiming's normal”, with $\mathbf{W}^T\mathbf{W} \approx \mathbf{I}_n$, and we had no enough time to change such weak initialization before the paper submission deadline. Now, we have **only** improved their initialization from the default “Kaiming's normal” to $\mathcal{N}(0, \sqrt{2/m})$, making $\mathbf{W}^T\mathbf{W} \approx 2\mathbf{I}_n$. The final performance becomes better than or comparable with the counterparts.

Table 1. Comparisons of top-1 error rate (%) with ResNet 110, Wide ResNet 28-10, and ResNext 29-8-64 on CIFAR10 and CIFAR100 under the normal training process.

Model	Regu.	CIFAR10	CIFAR100
ResNet-110	None	7.04	25.42
	FO	6.78	25.01
	MC	6.97	25.43
	SRIP	6.55	25.14
	Ours	**6.50**	**24.98**
Wide ResNet 28-10	None	4.16	20.55
	FO	3.76	18.56
	MC	3.68	18.90
	SRIP	3.60	18.19
	Ours	**3.47**	**18.17**
ResNext 29-8-64	None	3.70	18.53
	FO	3.58	17.59
	MC	3.65	17.62
	SRIP	**3.48**	16.99
	Ours	3.49	**16.95**