We thank all reviewers. It was very encouraging to read your opinion on the *strengths* of this work: importance of the problem, lack of prior work, novelty of the analysis, and consistency between our theory with empirical observation.

[R2] Positive about this work. Minor typos. We are very happy that you liked our contributions and found our results intriguing. We will surely fix typos in the final version.

[R3] In practice people use both predictions and original labels, but this paper only uses predictions. Thank for pointing out. In the literature, "self-distillation" refers to a host of related ideas. We have adopted the variant proposed by the well-cited work of [1], which only uses predictions. Incidentally, [1] compared training with pseudo labels generated merely from teacher's predictions versus its combination with the ground truth labels (respectively `BAN` and `BAN+L` in their Table 2), and observed that `BAN` outperforms `BAN+L`. What you are referring to is closer to `BAN+L` (unfortunately you did not specify a reference so that we could be more precise here). That being said, as self-distillation is a new area, there is not yet enough evidence to claim one variant is better than others. As such, we do not think the variant we studied is less qualified than others, if not more (due to `BAN+L` observation of [1]).

Regarding your interesting question of whether blending the original labels into the pseudo-labels can prevent collapse, it indeed does that, but at the cost of undoing some of the regularization benefits of self-distillation (consistent with `BAN+L` vs `BAN` observation of [1]). The emphasis on the original labels facilitates overfitting to those labels and diminishes the regularization effect. Note that the collapse is *not a practical concern*, as one should stop even earlier than that when over-regularization begins. The latter can be detected by measuring test performance on the validation set.

[1] T. Furlanello et al. "Born-Again Neural Networks", *ICML 2018*.

[R3] What is the purpose of lower bound on the number of rounds of self distillation. To be clear, this is the lower bound on the number of self-distillation rounds before the solution *collapses* to zero. This quantity is critical in determining the ultimate strength of regularization that self-distillation can achieve. The reason is, as proved in Sections 3.3 and 3.5, the sparsity level enhances with each self-distillation round (as long as collapse is not reached) [line 188-194]. Once the solution collapses, nothing interesting happens from that point on [line 154-156]. Thus, the highest achievable sparsity is right before the collapse, which is determined by the lower bound you mentioned.

[R3] What is the purpose of bounding $S_{\boldsymbol{B}}$ in Section 3.5? The goal of the paper is to understand the implicit regularization of self-distillation. Our analysis reveals that this regularization leads to a sparser matrix $\boldsymbol{B}$. $S_{\boldsymbol{B}}$ is simply a way to quantify such sparsity; it takes the diagonal matrix $\boldsymbol{B}$ and returns a single number. Since we showed the ratio between pairs of diagonals of $\boldsymbol{B}$ change monotonically in t [line 188-191], one can see that the sparsity index $S_{\boldsymbol{B}}$ increases in t, hence showing self-distillation enhancing the sparsity. We will clarify this in the revision.

In addition, $S_{\boldsymbol{B}}$ is used to analyze how ϵ affects the achievable sparsity level. Theorem 6 reveals that being near the interpolation regime can enhance the regularization effect of self-distillation (i.e. leading to sparser representation).

[R1] In practice it is unlikely that anybody would do self-distillation, as it is easier to just set the regularization parameter better in the first place. There seems to be a *misunderstanding* here, which has caused the result to seem trivial. While it is true that one can "just set the regularization parameter better in the first place", our results show that it is *not possible* to achieve the regularization effect due to self-distillation in this way. Increasing the ridge regularization coefficient c will scale all eigenvalues of the kernel by the same factor. Such uniform scaling does not change the sparsity pattern of the eigenvalues. In contrast, self-distillation exponentiates the eigenvalues by the number of steps t, which results in a non-uniform scaling. This allows to shrink some directions more than others. We add that increasing c without any self-distillation is similar to regularization by early stopping. We have discussed this in Section 3.4 and emphasized that these two regularization schemes behave very differently.

[R1] Is theoretical contribution of RKHS analysis alone interesting? Our results give insight into the behavior of self-distillation in the RKHS setting, which on its own is an interesting family of learning problems. The dynamics of self-distillation in RKHS setting follows a nonlinear recurrence for which there is no closed form solution, and this *significantly* complicates the analysis. We have been able to fully characterize the evolution of self-distillation in the RKHS setting and prove how the spectrum of the kernel evolves in an intriguing way that leads to sparsity. The fact that self-distillation promotes sparsity has *never* been noticed before, let alone proving it. We will clarify this in the revision.

[R1] There is a gap between theory (RKHS/NTK) and practical neural networks. We agree that a mathematically rigorous characterization of self-distillation for deep neural networks - comparable to our results in the RKHS setting - would be an exciting contribution. However, please note that as neural networks are highly nonlinear, almost *all* theoretical developments *require* some simplifying assumptions to keep the analysis tractable.