We thank all the reviewers for their precious suggestions. Here are our response.

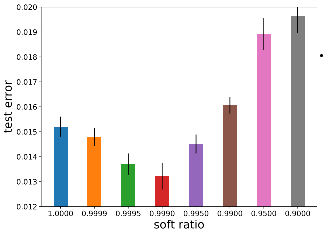
Reviewer 1
Q1: Advantages over bound in [18] under small sample size. As our discussion in Sec 3.2 shows, [18]'s bound only uses α'_1, which is the angle between teacher and student's weight change when student is trained with *only one* sample. Even in the regime of small sample size, α'_n in our bound (*considering all the training data*) is still much smaller than α'_1, therefore our bound still shows advantage.

Q2: Reasons for considering over-parameterized student network. It is well known that theoretical analysis on practical distillation (i.e. small student network) is still a challenging open problem. The over-parameterization assumption in this work is for the concern of theoretical convenience, and we acknowledge it might be a little impractical at the current stage. In this work our focus is to establish an explicit and global relation between labels and network parameters. NTK technique is one of the few choices so far. We believe our result is the first and an important step for later conducting in-depth theoretical analysis on practical distillation.

Q3: Small training sample size. We fix the training sample size of teacher in our definition of data inefficiency, in order to fix the difficulty of the task. Indeed, if teacher and student are trained by the same small sample set, the analysis would be extremely complex. We will try to attack this problem as future work.

Reviewer 2
Q1: No ImageNet experiments. Training wide networks on ImageNet is extremely. We have validated our findings across synthetic and medium-scale(CIFAR10) datasets, and we believe our findings still are true on large scale datasets.

Reviewer 3
Q1: No small scale synthetic experiments. We perform imperfect teacher distillation on synthetic dataset. The right figure shows a trade-off behavior between soft and hard labels, which is similar to that of CIFAR10 dataset(Fig.5 left). We will add the code and more details to future version of this paper.
Q2: Practical impact of data inefficiency metric. We already use data inefficiency to show that early stopping and higher soft ratio may benefit distillation. In practice we can further design new loss that has lower inefficiency to improve distillation. More interestingly, data inefficiency can be a measure of the difficulty of certain task.



Imperfect teacher distillation on synthetic dataset. For teacher we use the training strategy in Fig.5 right. It is early stopped at $e = 5113$ and at a test error of 1.06% to make this phenomenon obvious. The sample size for student is 2^{14}. This figure shows a clear trade-off on soft ratio.

Reviewer 4
Q1.1: Definition of neural network f. It is stated in Supplementary Material Sec.S3. We will add it into the main text of our paper.
Q1.2: Reasons for over-parameterization of student. We answer a similar question in **Q2** of ***Reviewer 1***.
Q1.3: Convergence of distillation loss. Even though [7] only proves the convergence for only L2 loss, we believe this also holds for our distillation loss, with a little bit of modification. The key idea of [7] is that, convergence is guaranteed by the near constancy of NTK matrix $\hat{\Theta}(\mathbf{X}, \mathbf{X})$. Then we can use the following equation to prove convergence, $\dot{\mathbf{z}}_{\mathrm{s}} = -\eta\hat{\Theta}(\mathbf{X}, \mathbf{X})(\mathbf{z}_{\mathrm{s}} - \mathbf{z}_{\mathrm{eff}})$. As proved in the original paper of NTK([11]), the near constancy of $\hat{\Theta}(\mathbf{X}, \mathbf{X})$ has no requirement on the type of loss, so this is also true for our distillation loss. Then, the difference only lies in the second term $\mathbf{z}_{\mathrm{s}} - \mathbf{z}_{\mathrm{eff}}$, which in the case of distillation, is substituted with $\partial_{\mathbf{z}_{\mathrm{s}}}\mathcal{L}(\mathbf{z}_{\mathrm{s}}, \mathbf{z}_{\mathrm{eff}})$ (same as eq.6 in [13]). Due to the fact of finite training data and the convexity of $\mathcal{L}$ (w.r.t $\mathbf{z}_{\mathrm{s}}$), the gradient can be lower bounded by another L2 loss, so $|\partial_{\mathbf{z}_{\mathrm{s}}}\mathcal{L}(\mathbf{z}_{\mathrm{s}}, \mathbf{z}_{\mathrm{eff}})| \geq \mu|\mathbf{z}_{\mathrm{s}} - \mathbf{z}_{\mathrm{eff}}|$, then the convergence of distillation loss can be guaranteed. A similar proof of convergence is used in Theorem A.3 of [18] for linear distillation. We will add this point to the new version of the paper.

Q2: Definition of b in line 140. Fig.4 middle and right suggest an empirical power law relation of data inefficiency $\mathcal{I}(n)$ w.r.t. sample size n, $\mathcal{I}(n) \sim n^{-b}$. b is the parameter to describe this relation. This observation help us to get the asymptotic behavior of $||\Delta_{\hat{w}}||_2$ w.r.t. n, which leads to the asymptotic estimate of transfer risk bound. However, at line 140 we haven't introduce $\mathcal{I}(n)$, therefore we give an equivalent definition of $\partial_n \ln ||\Delta_{\hat{w}}||_2 \sim n^{-b-1}$. We are not certain whether $b > 1$ or not because b depends on various hyper-parameters. However, we do find b to be bigger when teacher's stopping epoch is small (i.e. the task is easy), which might be better than classic bound.

Q3: Lack of theorems in Sec.4, 5. Thank you for the suggestion. We will summarize the core results and formulate them into formal theorems.

Q4: Organization of Sec.3, 4, 5. Actually the three sections are not disjoint since each later section serves as a complementary of previous one. In Sec.3 we use $p(\beta)$ as an estimate of transfer risk bound, but this measure needs much effort to calculate and cannot show the obvious advantage of soft label. Therefore in Sec.4 we propose a simpler measure of data inefficiency. Sec.5 serves as complementary of both Sec.3 and 4, for that the previous two sections consider only perfect teacher.