UNDERSTANDING THE BENEFITS OF IMMATURE TEACHERS IN KNOWLEDGE DISTILLATION THROUGH ALIGNMENT AND REGULARIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Knowledge distillation (KD) trains a small student model on the outputs of a large teacher model. Its effect has mainly been explained from two perspectives: the transfer of inter-class similarities contained in teacher predictions and regularization by soft targets. However, because both effects arise from the same distillation loss, what determines their balance remains unclear. Moreover, reports that immature teachers taken during training can yield better students than mature teachers are difficult to explain from the knowledge-transfer view that more accurate teachers are more beneficial. We show that, in the high-temperature limit, this balance is determined by the teacher logit norm, and we examine the resulting hypothesis experimentally. In this limit, the distillation loss separates into a teacher-independent penalty on the student logit norm and an alignment term whose gradient is proportional to the teacher logit norm. On four image and text classification datasets, the test loss of students distilled from teachers of different maturities followed a nearly common U-shaped curve against the student logit norm, and teacher maturity changed the point that students reached on this curve. Fixing the direction of the mature teacher's logits and replacing only their norm with that of an immature teacher reduced the student test loss below that obtained with the mature teacher, without changing the teacher's predicted classes. At the same teacher logit norm, however, student performance differed with the direction. These results suggest that the balance between knowledge transfer and regularization shifts with the teacher logit norm, and that this balance partly underlies the effectiveness of immature teachers.

1 INTRODUCTION

Knowledge distillation (KD) trains a small student model on the outputs of a large teacher model (Buciluǎ et al., 2006; Hinton et al., 2015). Hinton et al. (2015) explained that teacher probabilities represent relationships between classes and that, by imitating them, the student inherits information unavailable from ground-truth labels alone. This explanation is supported by the gains from methods that better transfer intermediate representations or logits (Romero et al., 2015; Zagoruyko & Komodakis, 2017; Zhao et al., 2022) and by students inheriting teacher properties (Ojha et al., 2023). Under this explanation, students trained from better teachers are expected to perform better.

However, Cho & Hariharan (2019) showed that more accurate teachers do not necessarily yield better students and that stopping teacher training early can improve students. They attributed this partly to the capacity gap between the teacher and the student. Wang et al. (2022) likewise showed that teachers during training (hereafter, immature teachers) can yield better students than teachers after training (hereafter, mature teachers). They further showed that the intermediate representations of immature teachers retain more input information than those of mature teachers, and linked this property to their effectiveness. These explanations attribute the benefit of immature teachers to the information they hold and thus presuppose knowledge transfer from teacher to student.

In contrast, Yuan et al. (2020) explained the benefit of insufficiently trained teachers through regularization. They showed that even teachers with extremely low accuracy improve students over standard training, and that a similar improvement is obtained with an artificial distribution that as-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

signs a high probability to the correct class and uniform probabilities elsewhere, i.e., a target without class relationships. This result indicates that teacher-specific class relationships are not always necessary for the improvement by KD. On this basis, they regarded KD as label smoothing learned from teacher outputs and argued that its effect stems not only from the transfer of class relationships but equally or more from regularization by soft targets. Explanations that attribute the benefit of immature teachers solely to the information held by the teacher (Wang et al., 2022) therefore cannot fully account for the improvement by KD.

Thus, Yuan et al. (2020) acknowledged both effects in KD, and Tang et al. (2020) also decomposed the effect of KD into factors including label-smoothing regularization and the transfer of class relationships. However, these studies treat both effects as properties of the teacher's output distribution and do not show how the two separate within the distillation loss or how their balance varies across teachers. Consequently, the finding that teachers of intermediate maturity, i.e., at an intermediate stage of training, yield the best students (Cho & Hariharan, 2019; Wang et al., 2022) does not follow directly from these explanations. Our goal is to clarify why immature teachers can yield better students than mature teachers. To this end, we identify the factor that changes the balance between knowledge transfer and regularization across teachers under the same distillation loss and explain the effectiveness of immature teachers through it. This factor also helps to understand which teacher maturity is desirable and how teacher outputs should be modified to improve students.

We show that knowledge transfer and regularization appear as two terms of the distillation loss whose balance is set by the teacher logit norm. At sufficiently high temperature, the distillation loss becomes the squared distance between the teacher and student logits centered by subtracting the class mean (hereafter simply logits) (Hinton et al., 2015; Kim et al., 2021). It separates into a term that suppresses the student logit norm and a term that increases the inner product between the student and teacher logits (hereafter, alignment with the teacher, or teacher–student alignment; Figure 1). The former has the form of regularization on the squared logit norm (Pezeshki et al., 2021), is independent of the teacher, and corresponds to the regularization view. The latter corresponds to knowledge transfer, and its gradient scales with the teacher logit norm. This decomposition leads to the following hypothesis. Because immature teachers have small logit norms (Appendix D), the relative contribution of regularization increases, which improves student generalization. If the norm is too small, however, the gradient of alignment, which also aids generalization, becomes small. Student generalization should therefore peak at intermediate teacher logit norms. Under this hypothesis, the results of Yuan et al. (2020); Hinton et al. (2015); and Cho & Hariharan (2019) and Wang et al. (2022) can be interpreted as small-, large-, and intermediate-norm cases, respectively.

We test this hypothesis on image classification (CIFAR-10, CIFAR-100, Tiny ImageNet) and text classification (AG News). Our contributions are twofold. First, the test loss of students distilled from teachers of different maturities showed a U-shaped relationship with the student logit norm and with the inner product with the teacher logits, and teacher maturity changed the point that students reached on this curve. This relationship was observed on four datasets with different characteristics and, on CIFAR-10 and CIFAR-100, held across distillation temperatures. Second, because maturity changes both the norm and the direction (the logit vector divided by its norm) of the teacher logits, we replaced only the norm with that of an immature teacher while preserving the direction of the mature teacher's logits. This improved the student test loss over that obtained with the mature teacher. The norm-replaced teacher and the immature teacher providing its norm have identical norms for each input and differ only in direction, yet students distilled from them performed differently, and the maturity yielding the best student also differed. The former result suggests that the teacher norm changes the weight of regularization independently of the predictions, and the latter that the effect of alignment depends on the teacher logit direction. Hence, part of the effectiveness of immature teachers can be reproduced merely by adjusting a mature teacher's norm, and adjusting maturity can serve to change the balance between the two effects.

2 Related Work

2.1 Teacher quality and student performance

Beyond Cho & Hariharan (2019) and Wang et al. (2022), other studies have also reported that more accurate teachers do not necessarily yield better students, and two research directions address this issue. The first direction bridges the capacity gap between the teacher and the student. Mirzadeh

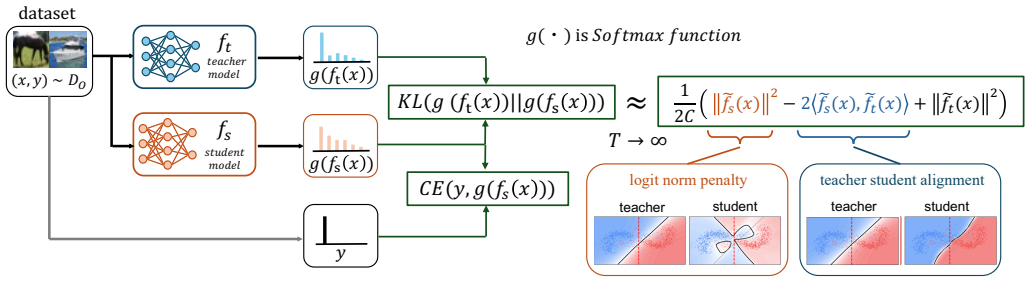

Figure 1: Overview of KD and the high-temperature decomposition of the distillation term. $f_t(x)$ and $f_s(x)$ are the teacher and student logits, and $\tilde{f}_t(x)$ and $\tilde{f}_s(x)$ are their respective centered logits. Under the high-temperature approximation, the distillation term decomposes into a logit norm penalty and teacher–student alignment, up to a student-independent constant. For simplicity, the figure omits the temperature of g and the factor αT^2; the approximation holds for $\mathcal{D}_T$ as $T \to \infty$ (Equations (5) and (6)). The bottom row shows toy decision boundaries when the teacher logit magnitude is changed at a fixed direction.

et al. (2020) mitigated the resulting degradation through staged distillation via an intermediate-sized model. The second direction revisits the distillation loss. Huang et al. (2022) argued that exact prediction matching via the KL divergence overly constrains students of strong teachers and proposed a looser matching criterion. These studies focus on teacher–student model differences or replace the loss. We instead keep the KL divergence and examine how logit norm changes during teacher training alter its behavior.

2.2 Knowledge transfer and regularization views of KD

The effect of KD has mainly been explained from two perspectives: regularization and knowledge transfer from the teacher to the student. Many studies besides Yuan et al. (2020) interpret KD as regularization. For example, self-distillation improves students even when the teacher and the student share the same architecture (Furlanello et al., 2018). Mobahi et al. (2020) theoretically explained this improvement as an amplification of regularization. Menon et al. (2021) showed that the variance of the student's empirical risk decreases as the teacher's predicted probabilities approach the Bayes class probabilities. From the knowledge-transfer perspective, KD has also been explained as making the student approach the teacher's function itself rather than merely receive its knowledge. Somepalli et al. (2022) showed that KD yields decision boundaries closer to the teacher's than standard training. Beyer et al. (2022) further regarded KD as function matching and substantially improved student accuracy through consistent teacher–student inputs and longer student training, even with conventional KD (Hinton et al., 2015). This paper provides a framework that unifies these two explanations.

2.3 Logit scaling in KD

Several distillation methods adjust the teacher's confidence or logit scale. Guo et al. (2020) normalized logits to remove the confidence gap between the teacher and the student. Li et al. (2022) addressed the excessive confidence of large teachers by applying different temperatures to correct and incorrect classes. NormKD (Chi et al., 2023) and Logit Standardization (Sun et al., 2024) adjust the logit scale per sample. The operation in Section 5 is likewise mathematically equivalent to changing the teacher temperature per input. We do not propose this operation as a distillation method but use it as an analysis to show that changes in the logit norm during teacher training alter the relative contributions of student logit regularization and alignment with the teacher. Our view also offers a perspective that interprets the effects of these methods as changes in the balance between regularization and alignment with the teacher.

3 Regularization and Alignment in the Distillation Loss

3.1 High-temperature decomposition

Consider C-class classification, and let $f_t(x), f_s(x) \in \mathbb{R}^C$ denote the output logits of the teacher and student models for an input x. We define the softmax function at temperature $T > 0$ as

$$g_T(z) = \mathrm{softmax}(z/T). \tag{1}$$

The standard KD loss of Hinton et al. (2015) is

$$\mathcal{L}_{\mathrm{KD}} = (1-\alpha)\,\mathrm{CE}(y, g_1(f_s(x))) + \mathcal{D}_T, \tag{2}$$

$$\mathcal{D}_T = \alpha T^2 D_{\mathrm{KL}}(g_T(f_t(x)) \,\|\, g_T(f_s(x))), \qquad 0 < \alpha < 1. \tag{3}$$

The teacher is fixed, and the student is trained to minimize $\mathcal{L}_{\mathrm{KD}}$ on the training data.

Because softmax is shift-invariant, the logits retain a degree of freedom that does not affect the output distribution. We remove it by subtracting the class mean and describe the high-temperature approximation with the resulting centered logits. With $\mathbf{1} \in \mathbb{R}^C$ the all-ones vector, we define

$$\tilde{f}_n(x) = f_n(x) - \frac{1}{C}\mathbf{1}\mathbf{1}^\top f_n(x), \qquad n \in \{s, t\}. \tag{4}$$

We omit the input x below for simplicity.

For fixed teacher and student logits, the distillation term approaches the squared distance between the centered logits as $T \to \infty$ (Hinton et al., 2015; Kim et al., 2021):

$$\mathcal{D}_T = Q + O(T^{-1}), \tag{5}$$

$$\begin{aligned} Q &= \frac{\alpha}{2C}\left\|\tilde{f}_s - \tilde{f}_t\right\|_2^2 \\ &= \underbrace{\frac{\alpha}{2C}\left\|\tilde{f}_s\right\|_2^2}_{\text{logit norm penalty}} - \underbrace{\frac{\alpha}{C}\left\langle \tilde{f}_s, \tilde{f}_t \right\rangle}_{\text{teacher–student alignment}} + \underbrace{\frac{\alpha}{2C}\left\|\tilde{f}_t\right\|_2^2}_{\text{constant w.r.t. student}}. \end{aligned} \tag{6}$$

The first term suppresses growth of the student logit norm. The second term has a negative sign, so the loss decreases as the inner product between the teacher and student logits increases; this term therefore aligns the student's outputs with the teacher's. The third term depends only on the teacher and does not affect the student.

3.2 Teacher logit norm sets the balance

In Equation (6), teacher–student alignment is proportional to the teacher logits. To make this explicit, we preserve the direction of the teacher logits and scale them to $\rho\tilde{f}_t$ with a per-input factor $\rho > 0$. The gradient with respect to the student logits is then

$$\nabla_{\tilde{f}_s} Q_\rho = \frac{\alpha}{C}\left(\tilde{f}_s - \rho\tilde{f}_t\right), \tag{7}$$

where the first term, which suppresses the norm, is independent of ρ, and only the second term, directed toward the teacher, is proportional to ρ. In the limit $\rho \to 0$, the distillation term has the same form as the L2 penalty in Spectral Decoupling (Pezeshki et al., 2021). Thus, under the high-temperature approximation, the magnitude of the teacher logits determines the relative contributions of alignment with the teacher and suppression of the student logits. At finite temperatures, the expression of Equation (6) as a squared norm penalty and a logit inner product is not exact. We use Equation (6) to derive our hypothesis and test its validity at finite temperatures, including $T = 1$.

3.3 Illustration on toy data

Before turning to real data, we visualize the relationship in Equation (6) on two-dimensional toy data. We use rotated two-moons data, whose nonlinear moon structure determines the labels, and add to the first coordinate x_1 a linear shift (shortcut) that mostly agrees with the labels (Geirhos

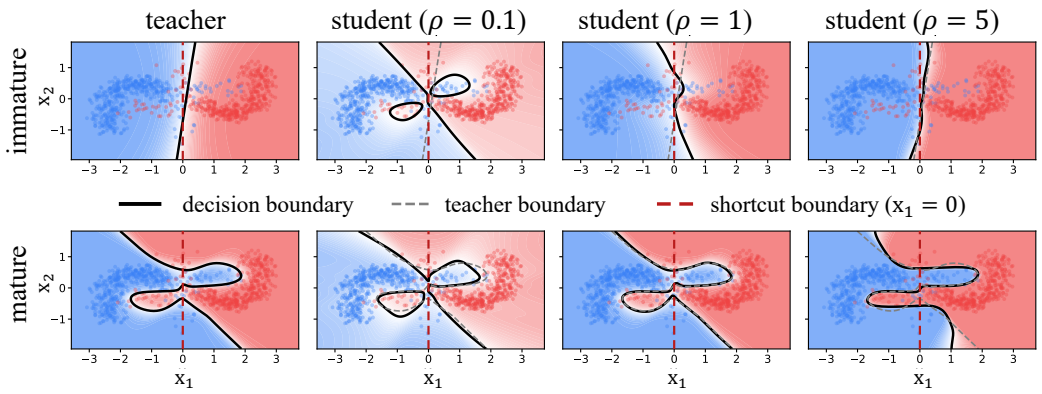

Figure 2: Decision boundaries of teachers and students on rotated two-moons data with a shortcut. The top and bottom rows show the immature and mature teachers, respectively, with their students. Each row shows the teacher, followed by students distilled with the teacher logits scaled by $\rho = 0.1, 1, 5$. Solid black, dashed gray, and dashed red lines denote each model's decision boundary, the teacher's decision boundary, and the shortcut-only decision boundary $x_1 = 0$, respectively.

et al., 2020; Shah et al., 2020). A classifier using only the shortcut has the linear decision boundary $x_1 = 0$, so a boundary closer to this line indicates greater reliance on the shortcut. We use MLP teachers trained for 2 epochs (immature) and 200 epochs (mature), and distill students at $T = 4$, where the high-temperature approximation is more applicable, from teacher logits scaled by a factor $\rho = 0.1, 1, 5$ shared across all inputs. Detailed settings are provided in Appendix B.

Figure 2 shows the results. The immature teacher in the top row learned a decision boundary along $x_1 = 0$ and relied on the shortcut, whereas the mature teacher in the bottom row learned a curved boundary that followed the moon structure. For students of the immature teacher, increasing ρ progressively moved the student boundary toward the teacher's boundary along the shortcut. This change agrees with the behavior expected from Equation (7). When ρ is small, the gradient toward the teacher is weak, and the distillation term acts mainly as regularization that suppresses the student logit norm, so the student does not inherit the teacher's shortcut. As ρ increases, alignment with the teacher contributes more, and the student follows the teacher's decision boundary. The same trend held for the mature teacher: as ρ increased, the student's decision boundary approached the teacher's. Thus, even at a fixed teacher logit direction, the norm changes how much of the teacher's decision boundary the student inherits. The following sections examine the relationship between the teacher logit norm and student generalization on real data.

4 Teacher Maturity and Student Generalization

4.1 Experimental setup

Students are trained with Hinton-style KD (Equation (2)), with $T = 1$ and $\alpha = 0.9$ in the main experiments. We use a ResNet-32 teacher and a ResNet-20 student (He et al., 2016) on CIFAR-10/100 (Krizhevsky, 2009), and a ResNet-34 teacher and a ResNet-18 student on Tiny ImageNet (Le & Yang, 2015). For text classification, we use AG News (Zhang et al., 2015) with a four-layer Transformer teacher and a two-layer Transformer student (Vaswani et al., 2017). Teacher maturity is measured by the number of teacher training epochs. We use teachers at 14 maturities from 1 to 400 epochs for CIFAR-10/100, 6 from 1 to 100 epochs for Tiny ImageNet, and 9 from 0.1 to 30 epochs for AG News. The learning rate is fixed to a small value without a scheduler on CIFAR-10/100 and AG News and decayed with a cosine schedule for teachers and students on Tiny ImageNet. Student data augmentation, which strongly affects distillation performance (Beyer et al., 2022; Stanton et al., 2021), is disabled in all conditions. Students are trained with three random seeds per teacher; figures report the mean unless otherwise stated. We evaluate final-epoch test accuracy and test cross-entropy at temperature 1 (hereafter, test loss). The student logit norm and the inner product with the teacher logits in the figures are averaged over test inputs.

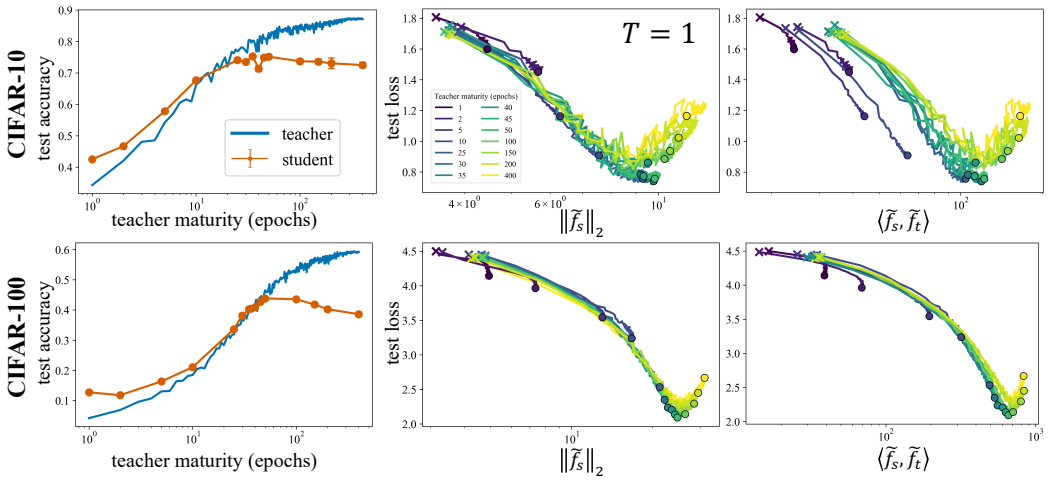

Figure 3: Teacher maturity and student performance on CIFAR-10 and CIFAR-100. Left: test accuracy of the teachers (blue) and final-epoch test accuracy of the students distilled from each teacher (orange; mean and standard deviation over three seeds) against teacher maturity. Center/right: test loss against the student logit norm/the inner product of the teacher and student logits. Colors indicate maturity, with lighter colors for more mature teachers. Curves are student training trajectories; $\times$ and circles mark epoch 1 and the final epoch.

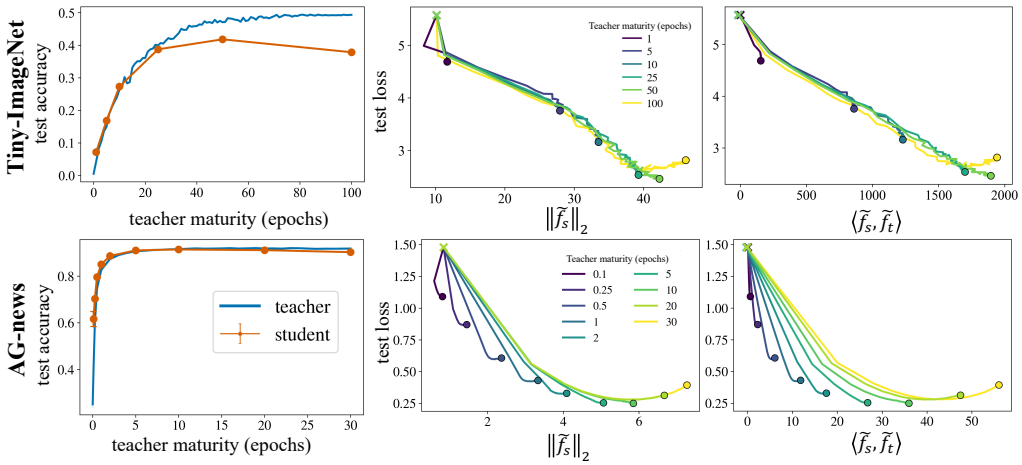

Figure 4: Results on Tiny ImageNet (top) and AG News (bottom). Left: teacher and student test accuracy against maturity. Center/right: test loss against the student logit norm/the inner product of the teacher and student logits. Notation follows Figure 3.

4.2 Immature teachers yield better students

The left columns of Figures 3 and 4 show the test accuracy of the teachers and of the students distilled from each teacher. Teacher accuracy increased with maturity, whereas student accuracy did not increase monotonically. On all datasets, students distilled from teachers of intermediate maturity outperformed those distilled from mature teachers, and on CIFAR-10 and CIFAR-100, the highest accuracy was also obtained with teachers of intermediate maturity. The improvement on AG News, however, was small.

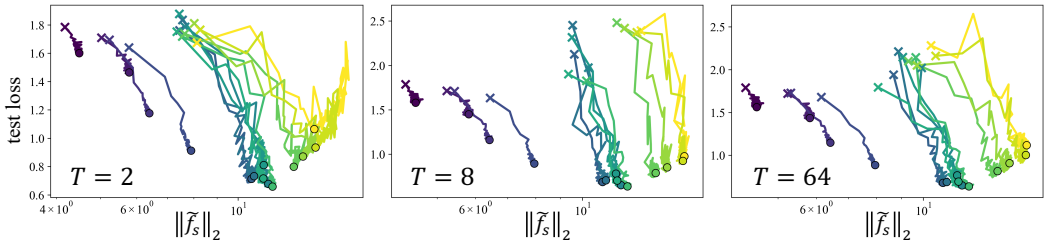

Figure 5: Test loss against the student logit norm on CIFAR-10 at $T = 2, 8, 64$ (left to right). Colors indicate maturity; $\times$ and circles mark epoch 1 and the final epoch. All training settings other than temperature and the random seeds are the same as in Figure 3.

4.3 Students share a common curve

The center and right columns of Figures 3 and 4 show test loss against the student logit norm and the inner product with the teacher logits. Each curve is the training trajectory of a student distilled from a teacher of a given maturity. On CIFAR-10, CIFAR-100, and Tiny ImageNet, the trajectories nearly overlapped on a single curve regardless of the teacher. This curve was U-shaped: test loss decreased as the norm and inner product increased, reached a minimum, and then increased. Students distilled from teachers of low maturity ended training on the left side of the U, those from teachers of intermediate maturity near the minimum, and those from mature teachers beyond it. Under our decomposition, the left side of the U corresponds to weak alignment with the teacher and dominant regularization, and the right side to dominant alignment. On Tiny ImageNet, the increase on the right side was small. On AG News, test loss was also U-shaped with an intermediate minimum.

On CIFAR-10 and CIFAR-100, students distilled from mature teachers exhibited epoch-wise double descent (Nakkiran et al., 2021): test loss decreased, increased, and then decreased again. These changes also remained on the common curve and appeared as an oscillation along the right side of the U. However, the minimum along the trajectories of students distilled from mature teachers exceeded the final-epoch loss of students distilled from teachers of intermediate maturity. Thus, the effect of immature teachers is not equivalent to early stopping of students distilled from mature teachers, and differences between teachers appear near the minimum of the curve. Our decomposition-based interpretation of this phenomenon is given in Appendix E.

4.4 Robustness across distillation temperatures

Figure 5 shows the results on CIFAR-10 at distillation temperatures $T = 2, 8, 64$. At $T = 1$, student trajectories moved from small toward large norms and inner products (Figure 3). At $T \geq 2$, in contrast, each student's norm remained nearly constant during training, test loss did not increase along the trajectories, and the U-shaped relationship appeared across the final points of different teacher conditions rather than along trajectories. This relationship therefore arises from differences between teachers rather than from overfitting over training time. Moreover, the appearance of the same relationship at high temperatures, where the approximation in Equation (6) is more applicable, supports the hypothesis that the teacher logit norm governs student generalization.

5 Intervening on the Teacher Logit Norm

5.1 Norm replacement

Because the direction and norm of teacher logits change simultaneously across maturities, the comparison in Section 4 alone cannot separate their effects. We therefore fix the direction of the mature teacher's logits for each input and distill students after replacing only their norm with that of an immature teacher. Let $\tilde{f}_t^{(M)}(x)$ and $\tilde{f}_t^{(r)}(x)$ denote the logits of the mature teacher and of a teacher

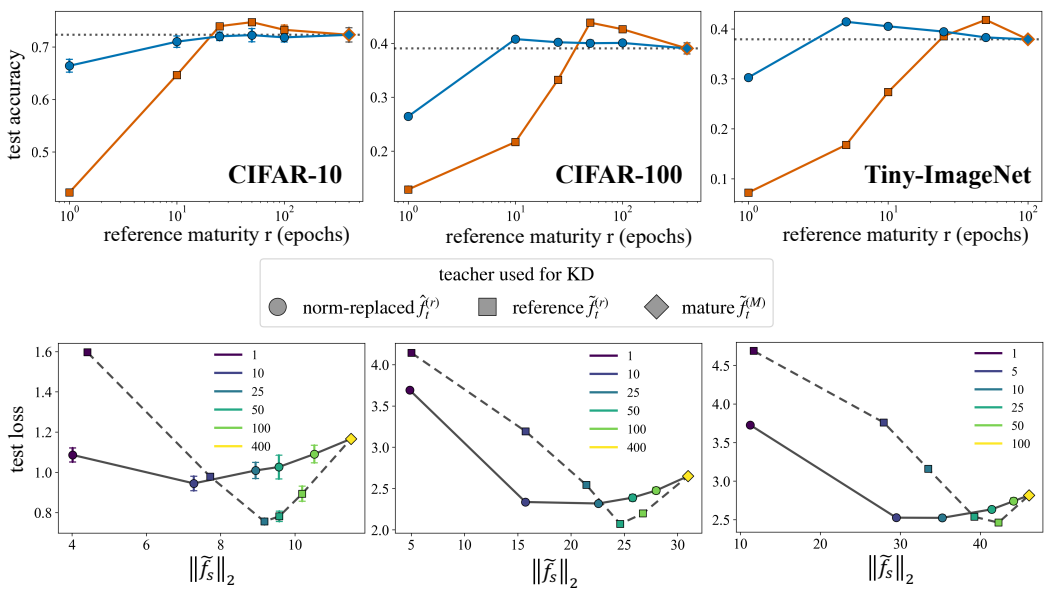

Figure 6: Student performance with the direction of the mature teacher's logits preserved and only their norm replaced (CIFAR-10, CIFAR-100, and Tiny ImageNet, left to right). Top: test accuracy against reference maturity r. Bottom: final-epoch test loss against the student logit norm. Circles, squares, and diamonds denote students distilled from norm-replaced teachers $\hat{f}_t^{(r)}$, teachers at reference maturity r ($\tilde{f}_t^{(r)}$), and mature teachers ($\tilde{f}_t^{(M)}$), respectively. Top: blue/orange lines connect circles/squares. Bottom: solid/dashed lines connect circles/squares in order of reference maturity; colors indicate reference maturity. Gray dotted lines show the values for students distilled from mature teachers. AG News results are provided in Appendix F.

of maturity r (hereafter, the teacher at reference maturity r), respectively. We use the teacher output

$$\hat{f}_t^{(r)}(x) = \frac{\|\tilde{f}_t^{(r)}(x)\|_2}{\|\tilde{f}_t^{(M)}(x)\|_2} \tilde{f}_t^{(M)}(x), \tag{8}$$

which corresponds to setting ρ in Section 3.2 to $\rho(x) = \|\tilde{f}_t^{(r)}(x)\|_2 / \|\tilde{f}_t^{(M)}(x)\|_2$ for each input. Because this operation preserves the direction of the mature teacher's logits, the ranking of the classes predicted by the teacher remains unchanged. In this experiment, teachers and students are retrained with a cosine learning rate schedule; the other student settings follow Section 4.

5.2 The norm alone changes student performance

Figure 6 shows student performance against reference maturity r. Lower reference maturities, i.e., smaller teacher logit norms, yielded smaller student logit norms. On all datasets, students using norms from intermediate reference maturities achieved lower test loss than students distilled from mature teachers. Test accuracy exceeded that obtained with mature teachers on CIFAR-100 and Tiny ImageNet but barely increased on CIFAR-10. Because test accuracy depends only on the predicted classes, whereas test loss also depends on the logit magnitude, the effect of norm replacement, which mainly acts on the student logit magnitude, likely appeared more clearly in test loss. Even with the direction fixed, test loss showed a U-shaped relationship with the student norm. Thus, changing only the norm, without changing the teacher's predicted classes, alters student performance.

5.3 The direction matters at a fixed teacher norm

At the same reference maturity r, the norm-replaced teacher and the teacher at reference maturity r have identical norms for each input and differ only in direction (circles and squares in Figure 6). Note that the bottom row of Figure 6 plots the student logit norm, so circles and squares at the same

reference maturity generally differ in horizontal position. At low reference maturities, students distilled directly from the teacher at reference maturity r performed much worse than those of the mature teacher, whereas those using the mature teacher's direction degraded less. At intermediate reference maturities, however, students using the direction of the teacher at reference maturity r performed better. Consequently, the loss-minimizing reference maturity was lower with the mature teacher's direction and higher with direct distillation. In addition, directly distilled students followed a common curve, as in Section 4.3, whereas norm-replaced students deviated from it, with lower loss than the curve at small student norms and higher loss near its minimum. Thus, even at the same teacher norm, student performance depends on the logit direction. By Equation (7), the norm sets the balance between regularization and alignment, and the direction sets where alignment pulls the student, as discussed in Section 6.

6 Discussion

Why immature teachers can be better. Maturity changes both the norm and the direction of the teacher logits (Section 5). In direct distillation, the norm grows with maturity and strengthens alignment while the direction becomes a more useful alignment target, likely improving students up to intermediate maturities. For sufficiently mature teachers, however, the logit norm is large, and the outputs are sharp distributions that nearly reproduce the training labels. Stronger alignment thus fits the student to these sharp targets and increases the student logit norm (the right side of the U), which the relatively weakened regularization cannot suppress. The student's predictions consequently become overly sharp, and the larger loss on misclassified inputs likely increased test loss. Indeed, reducing only the norm at the mature teacher's direction improved the loss (Figure 6), which shows that the mature teacher's logit norm contributes to this degradation. Intermediate-maturity teachers are thus likely effective because their norm strengthens alignment and their direction is a useful target, without overly weakening regularization.

Relation to knowledge transfer and regularization. Even when only the norm was changed, test loss showed a U-shaped relationship with the student logit norm (Figure 6). This relationship therefore arises from the teacher norm alone, without changing the predictions. On the toy data (Figure 2), students followed the teacher's outputs more closely as the norm increased. These results suggest that the U shape in the intervention reflects a change in the balance between regularization and alignment independent of the predictions. In contrast, the reference maturity that minimized test loss was lower with the mature teacher's direction and higher with direct distillation (Section 5.3). Thus, a factor other than the norm, namely the logit direction, affected the students. Hence, the weight of regularization likely depends mainly on the teacher norm, and the usefulness of alignment on the logit direction. In this view, knowledge transfer and regularization are not competing explanations but two effects of the same distillation term whose balance shifts with the teacher norm.

Limitations. To observe differences in teacher outputs, we disabled student data augmentation and assigned a large weight to the distillation term ($\alpha = 0.9$). Whether the same relationship holds under standard settings remains to be verified. Moreover, our main metric, test loss, also depends on the logit magnitude, so verification by evaluation after temperature calibration is needed.

7 Conclusion

We explained why immature teachers can yield better students than mature teachers from the perspective that the teacher logit norm sets the balance between knowledge transfer and regularization. Under the high-temperature approximation, the distillation term separates into a teacher-independent student logit norm penalty and an alignment term whose gradient scales with the teacher logit norm. Empirically, student test loss followed a nearly common U-shaped curve against the student logit norm and the teacher–student inner product, and maturity changed where students ended. Replacing only the norm of the mature teacher with that of an immature teacher reduced student test loss, whereas at the same norm, performance differed with the direction. That is, the norm sets the balance between regularization and alignment, and the direction sets where alignment pulls the student. These results suggest that immature teachers are effective partly because the teacher logit norm changes this balance.

AI USE STATEMENT

We used generative AI tools (OpenAI Codex and Anthropic Claude) to revise, organize, and translate the manuscript; to draft parts of the text, mainly the descriptions in the appendix, from the authors' notes and results; to search for related literature; and to implement and debug the analysis and visualization code. The research questions, the experimental design, and the training experiments were conceived and conducted by the authors. The authors verified the literature suggested by AI against the original sources and checked the behavior and outputs of AI-generated code. The authors take responsibility for the content of this paper, including text, claims, and artifacts produced with the assistance of generative AI.

REPRODUCIBILITY STATEMENT

Section 4 and Appendix C describe the training settings for all experiments, and Section 3.3 and Appendix B describe the toy data generation procedure and models. The approximation in Section 3 is based on the derivations of Hinton et al. (2015) and Kim et al. (2021), and is derived in Appendix A.

REFERENCES

Lucas Beyer, Xiaohua Zhai, Amélie Royer, Larisa Markeeva, Rohan Anil, and Alexander Kolesnikov. Knowledge distillation: A good teacher is patient and consistent. In *CVPR*, pp. 10925–10934, 2022.

Cristian Buciluǎ, Rich Caruana, and Alexandru Niculescu-Mizil. Model compression. In *KDD*, pp. 535–541, 2006.

Zhihao Chi, Tu Zheng, Hengjia Li, Zheng Yang, Boxi Wu, Binbin Lin, and Deng Cai. NormKD: Normalized logits for knowledge distillation. *arXiv preprint arXiv:2308.00520*, 2023.

Jang Hyun Cho and Bharath Hariharan. On the efficacy of knowledge distillation. In *ICCV*, pp. 4794–4802, 2019.

Tommaso Furlanello, Zachary Lipton, Michael Tschannen, Laurent Itti, and Anima Anandkumar. Born again neural networks. In *ICML*, pp. 1607–1616, 2018.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Jia Guo, Minghao Chen, Yao Hu, Chen Zhu, Xiaofei He, and Deng Cai. Reducing the teacher-student gap via spherical knowledge distillation. *arXiv preprint arXiv:2010.07485*, 2020.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pp. 770–778, 2016.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Tao Huang, Shan You, Fei Wang, Chen Qian, and Chang Xu. Knowledge distillation from a stronger teacher. In *NeurIPS*, 2022.

Taehyeon Kim, Jaehoon Oh, Nak Yil Kim, Sangwook Cho, and Se-Young Yun. Comparing Kullback-Leibler divergence and mean squared error loss in knowledge distillation. In *IJCAI*, pp. 2628–2635, 2021.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2015.

Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Ya Le and Xuan Yang. Tiny ImageNet visual recognition challenge. Technical report, Stanford University, 2015.

Xin-Chun Li, Wen-Shu Fan, Shaoming Song, Yinchuan Li, Bingshuai Li, Yunfeng Shao, and De-Chuan Zhan. Asymmetric temperature scaling makes larger networks teach well again. In *NeurIPS*, 2022.

Aditya K Menon, Ankit Singh Rawat, Sashank Reddi, Seungyeon Kim, and Sanjiv Kumar. A statistical perspective on distillation. In *ICML*, pp. 7632–7642, 2021.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. Improved knowledge distillation via teacher assistant. In *AAAI*, pp. 5191–5198, 2020.

Hossein Mobahi, Mehrdad Farajtabar, and Peter Bartlett. Self-distillation amplifies regularization in Hilbert space. In *NeurIPS*, pp. 3351–3361, 2020.

Preetum Nakkiran, Gal Kaplun, Yamini Bansal, Tristan Yang, Boaz Barak, and Ilya Sutskever. Deep double descent: Where bigger models and more data hurt. *Journal of Statistical Mechanics: Theory and Experiment*, 2021(12):124003, 2021.

Utkarsh Ojha, Yuheng Li, Anirudh Sundara Rajan, Yingyu Liang, and Yong Jae Lee. What knowledge gets distilled in knowledge distillation? In *NeurIPS*, pp. 11037–11048, 2023.

Mohammad Pezeshki, Oumar Kaba, Yoshua Bengio, Aaron C Courville, Doina Precup, and Guillaume Lajoie. Gradient starvation: A learning proclivity in neural networks. In *NeurIPS*, pp. 1256–1272, 2021.

Adriana Romero, Nicolas Ballas, Samira Ebrahimi Kahou, Antoine Chassang, Carlo Gatta, and Yoshua Bengio. FitNets: Hints for thin deep nets. In *ICLR*, 2015.

Harshay Shah, Kaustav Tamuly, Aditi Raghunathan, Prateek Jain, and Praneeth Netrapalli. The pitfalls of simplicity bias in neural networks. In *NeurIPS*, pp. 9573–9585, 2020.

Gowthami Somepalli, Liam Fowl, Arpit Bansal, Ping Yeh-Chiang, Yehuda Dar, Richard Baraniuk, Micah Goldblum, and Tom Goldstein. Can neural nets learn the same model twice? Investigating reproducibility and double descent from the decision boundary perspective. In *CVPR*, pp. 13699–13708, 2022.

Samuel Stanton, Pavel Izmailov, Polina Kirichenko, Alexander A Alemi, and Andrew G Wilson. Does knowledge distillation really work? In *NeurIPS*, pp. 6906–6919, 2021.

Shangquan Sun, Wenqi Ren, Jingzhi Li, Rui Wang, and Xiaochun Cao. Logit standardization in knowledge distillation. In *CVPR*, pp. 15731–15740, 2024.

Jiaxi Tang, Rakesh Shivanna, Zhe Zhao, Dong Lin, Anima Singh, Ed H. Chi, and Sagar Jain. Understanding and improving knowledge distillation. *arXiv preprint arXiv:2002.03532*, 2020.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, pp. 5998–6008, 2017.

Chaofei Wang, Qisen Yang, Rui Huang, Shiji Song, and Gao Huang. Efficient knowledge distillation from model checkpoints. In *NeurIPS*, 2022.

Li Yuan, Francis EH Tay, Guilin Li, Tao Wang, and Jiashi Feng. Revisiting knowledge distillation via label smoothing regularization. In *CVPR*, pp. 3903–3911, 2020.

Sergey Zagoruyko and Nikos Komodakis. Wide residual networks. In *BMVC*, 2016.

Sergey Zagoruyko and Nikos Komodakis. Paying more attention to attention: Improving the performance of convolutional neural networks via attention transfer. In *ICLR*, 2017.

Xiang Zhang, Junbo Zhao, and Yann LeCun. Character-level convolutional networks for text classification. In *NeurIPS*, pp. 649–657, 2015.

Borui Zhao, Quan Cui, Renjie Song, Yiyu Qiu, and Jiajun Liang. Decoupled knowledge distillation. In *CVPR*, pp. 11953–11962, 2022.

APPENDIX CONTENTS

A High-temperature expansion 13
B Toy experiment details 14
C Training details 14
D Teacher logit norms across maturities 15
E Learning curves 17
F Norm replacement on AG News 19
G Additional results at different distillation temperatures 20
H Results with other architectures 21

A HIGH-TEMPERATURE EXPANSION

This section derives Equation (5) in the main text, following the argument of Hinton et al. (2015). We first derive the gradient of the distillation term and its high-temperature limit (Appendix A.2), and then show that the distillation term itself reduces to the squared distance between the centered logits (Appendix A.3). Expanding this distance gives Equation (6), and replacing $\tilde{f}_t$ with $\rho\tilde{f}_t$ and differentiating gives Equation (7). Throughout, the teacher and student logits are fixed and bounded, and the asymptotic notation refers to $T \to \infty$.

A.1 CENTERED LOGITS

Softmax is invariant to adding the same constant to all logits. Since $\tilde{f}_n = f_n - \frac{1}{C}\mathbf{1}\mathbf{1}^\top f_n$ differs from f_n only by such a constant, we have $g_T(f_n) = g_T(\tilde{f}_n)$ for $n \in \{s, t\}$, and the centered logits satisfy

$$\sum_{i=1}^{C} \tilde{f}_{n,i} = 0. \tag{A.1}$$

We write the teacher and student distributions as $p = g_T(f_t)$ and $q = g_T(f_s)$.

A.2 GRADIENT OF THE DISTILLATION TERM

Since the teacher entropy does not depend on the student, $D_{\mathrm{KL}}(p \,\|\, q) = -\sum_k p_k \log q_k + \text{const}$. By the chain rule,

$$\frac{\partial D_{\mathrm{KL}}(p \,\|\, q)}{\partial f_{s,i}} = \sum_{k=1}^{C} \left(-\frac{p_k}{q_k}\right) \frac{\partial q_k}{\partial f_{s,i}}, \qquad \frac{\partial q_k}{\partial f_{s,i}} = \frac{1}{T}\, q_k \left(\delta_{ki} - q_i\right), \tag{A.2}$$

where δ_{ki} is the Kronecker delta. Using $\sum_k p_k \delta_{ki} = p_i$ and $\sum_k p_k = 1$,

$$\frac{\partial D_{\mathrm{KL}}(p \,\|\, q)}{\partial f_{s,i}} = -\frac{1}{T} \sum_{k=1}^{C} p_k \left(\delta_{ki} - q_i\right) = \frac{1}{T}\left(q_i - p_i\right), \qquad \nabla_{f_s} \mathcal{D}_T = \alpha T \left(q - p\right). \tag{A.3}$$

This expression is exact at every temperature.

When the temperature is large compared with the logits, $\exp(\tilde{f}_{n,i}/T) = 1 + \tilde{f}_{n,i}/T + O(T^{-2})$. With Equation (A.1), the denominator of the softmax is $C + O(T^{-2})$, so that

$$p_i = \frac{1}{C} + \frac{\tilde{f}_{t,i}}{CT} + O(T^{-2}), \qquad q_i = \frac{1}{C} + \frac{\tilde{f}_{s,i}}{CT} + O(T^{-2}). \tag{A.4}$$

Both distributions approach the uniform distribution, and their deviations from it are proportional to the centered logits. Substituting Equation (A.4) into Equation (A.3),

$$q - p = \frac{\tilde{f}_s - \tilde{f}_t}{CT} + O(T^{-2}), \qquad \nabla_{f_s} \mathcal{D}_T = \frac{\alpha}{C}\left(\tilde{f}_s - \tilde{f}_t\right) + O(T^{-1}). \tag{A.5}$$

Thus, at high temperatures, distillation moves the student's centered logits toward the teacher's, as shown by Hinton et al. (2015). The gradient of the KL divergence alone is of order T^{-2}; the factor T^2 in $\mathcal{D}_T$ keeps the magnitude of the distillation gradient independent of the temperature.

A.3 THE DISTILLATION TERM AS A SQUARED DISTANCE

We next show that the distillation term itself reduces to the squared distance whose gradient is Equation (A.5). With $x_k = (q_k - p_k)/p_k$, the KL divergence is

$$D_{\mathrm{KL}}(p \,\|\, q) = \sum_{k=1}^{C} p_k \log \frac{p_k}{q_k} = -\sum_{k=1}^{C} p_k \log\left(1 + x_k\right). \tag{A.6}$$

Since $x_k = O(T^{-1})$, we use $\log(1+x) = x - x^2/2 + O(x^3)$. The first-order term vanishes because $\sum_k p_k x_k = \sum_k (q_k - p_k) = 0$, and therefore

$$D_{\mathrm{KL}}(p \,\|\, q) = \frac{1}{2} \sum_{k=1}^{C} \frac{(q_k - p_k)^2}{p_k} + O(T^{-3}). \tag{A.7}$$

Substituting $p_k = 1/C + O(T^{-1})$ in the denominator and $q - p$ from Equation (A.5),

$$D_{\mathrm{KL}}(p \,\|\, q) = \frac{C}{2} \sum_{k=1}^{C} (q_k - p_k)^2 + O(T^{-3}) = \frac{1}{2CT^2} \left\| \tilde{f}_s - \tilde{f}_t \right\|_2^2 + O(T^{-3}). \tag{A.8}$$

Multiplying by αT^2 yields

$$\mathcal{D}_T = \frac{\alpha}{2C} \left\| \tilde{f}_s - \tilde{f}_t \right\|_2^2 + O(T^{-1}) = Q + O(T^{-1}), \tag{A.9}$$

which is Equation (5). Hinton et al. (2015) derived this limit under the assumption that the logits have zero mean for each input; the centering in Equation (4) makes this assumption hold by construction.

A.4 Validity at finite temperatures

The approximations above neglect terms of higher order in $\tilde{f}_t/T$ and $\tilde{f}_s/T$. They are therefore accurate when the centered logit norms are small relative to the temperature. At $T = 1$, this condition is generally not satisfied, which is why we use Equation (6) only to derive our hypothesis and test it experimentally at finite temperatures, including $T = 1$ (Sections 4 and 5).

B Toy experiment details

We center two-moons data with noise standard deviation 0.12 and rotate it by 35°. In this data, the nonlinear moon structure determines the labels. We then add a shift of ± 1.2 and Gaussian noise with standard deviation 0.25 to the first coordinate x_1. The training set contains 900 points, and the sign of the shift agrees with the label for 90% of them. This shortcut (Geirhos et al., 2020) makes a simple linear feature correlated with the labels coexist with a nonlinear one (Shah et al., 2020). Teachers and students are MLPs with two hidden layers of width 128 and tanh activations. Both are trained with Adam (Kingma & Ba, 2015) at a learning rate of 10^{-3} and a batch size of 128. Teachers are trained with the cross-entropy loss for 200 epochs, and we use the checkpoints at 2 epochs (immature) and 200 epochs (mature). Students are trained for 300 epochs with $T = 4$ and $\alpha = 0.7$, using teacher centered logits scaled by a factor $\rho \in \{0.1, 1, 5\}$ shared across all inputs.

C Training details

Table C.1 summarizes the settings. Students are trained with Hinton-style KD (Equation (2)) with $T = 1$ and $\alpha = 0.9$ unless otherwise stated. In the temperature experiments, only T is changed and all other settings and random seeds are kept fixed. For each teacher, we fix the teacher model and train students with three random seeds, except for the WRN experiments, for which we report a single seed (Appendix H). On AG News, each of the three student seeds is instead paired with a teacher trained with its own seed, so the variation across seeds also includes that of the teacher. Teachers are trained with the cross-entropy loss only. On CIFAR-10/100 (Section 4 and Appendices D, E, and G) and on AG News, the learning rate is constant without a scheduler. On Tiny ImageNet, and for norm replacement (Section 5) and the WRN students (Appendix H) on CIFAR-10/100, teachers and students are trained with a cosine learning rate schedule over all training epochs. We use a small initial learning rate because, with a large one, teachers converge within a few epochs and differences between their outputs become small. Data augmentation is applied only to teacher training; student data augmentation is disabled in all conditions (Beyer et al., 2022; Stanton et al., 2021). On CIFAR-10/100, inputs are normalized with per-channel means and standard deviations, and teacher augmentation consists of random 32×32 crops with 4-pixel padding and horizontal flips. Teacher

logits used for distillation and for the analysis in Appendix D are computed on training inputs without augmentation. ResNet-32 and ResNet-20 are CIFAR ResNets (He et al., 2016) with $6n + 2$ layers ($n = 5$ and 3), stage widths 16, 32, and 64, and 1×1 convolutional projection shortcuts. In Appendix H, the student is replaced with WRN-16-2, the pre-activation WideResNet (Zagoruyko & Komodakis, 2016) with depth 16 and widening factor 2, without dropout. For AG News, the teacher and student Transformers have hidden dimensions 256 and 128, feed-forward dimensions 1024 and 512, and four attention heads. Maturities below one epoch (e.g., 0.1 epochs) correspond to teacher checkpoints saved after the corresponding number of optimization steps. The TextCNN student in Appendix H applies convolutions with kernel sizes 3, 4, and 5 and 128 channels each to 128-dimensional token embeddings, without dropout. AG News has four classes with 120,000 training and 7,600 test examples; inputs are tokenized with a vocabulary of 32,000 tokens and truncated to at most 256 tokens.

Table C.1: Training settings. Maturities are teacher training epochs. r denotes the reference maturities used for norm replacement (Section 5), and M the mature teacher.

	CIFAR-10 / CIFAR-100	Tiny ImageNet	AG News
Teacher / student	ResNet-32 / ResNet-20	ResNet-34 / ResNet-18	4-layer / 2-layer Transformer
Student in App. H	WRN-16-2 (1 seed)	WRN-16-2 (1 seed)	TextCNN
Optimizer	SGD, momentum 0.9	SGD, momentum 0.9	AdamW
Initial learning rate	10^{-3}	10^{-3}	10^{-5}
Learning rate schedule	constant; cosine for Section 5 and App. H	cosine	constant
Weight decay	5×10^{-4}	5×10^{-4}	0.01
Batch size	256	128	256
Training epochs	400	100	30
Teacher augmentation	random crop (pad 4), horizontal flip	random crop, horizontal flip	token masking ($p = 0.2$)
Student augmentation	none	none	none
Teacher maturities	1, 2, 5, 10, 25, 30, 35, 40, 45, 50, 100, 150, 200, 400	1, 5, 10, 25, 50, 100	0.1, 0.25, 0.5, 1, 2, 5, 10, 20, 30
Reference maturities r	1, 10, 25, 50, 100	1, 5, 10, 25, 50	0.1, 0.5, 2, 5, 20
Mature teacher M	400	100	30

D TEACHER LOGIT NORMS ACROSS MATURITIES

For each teacher checkpoint, we compute the centered logit norm $\|\tilde{f}_t(x)\|_2$ for every training input without augmentation, which are the inputs used for distillation. Figure D.1 shows its distribution against teacher maturity (top) and its histogram at selected maturities (bottom). On both datasets, the median and the percentile bands rise with maturity, and the histograms of more mature teachers lie at larger norms. On CIFAR-10, the distributions of immature teachers early in training are broad with long tails; as maturity increases, the mode of the norm shifts toward larger values and the peak appears sharper. On CIFAR-100, the mode likewise shifts toward larger values as maturity increases, whereas the distributions of immature teachers early in training already show a sharp peak.

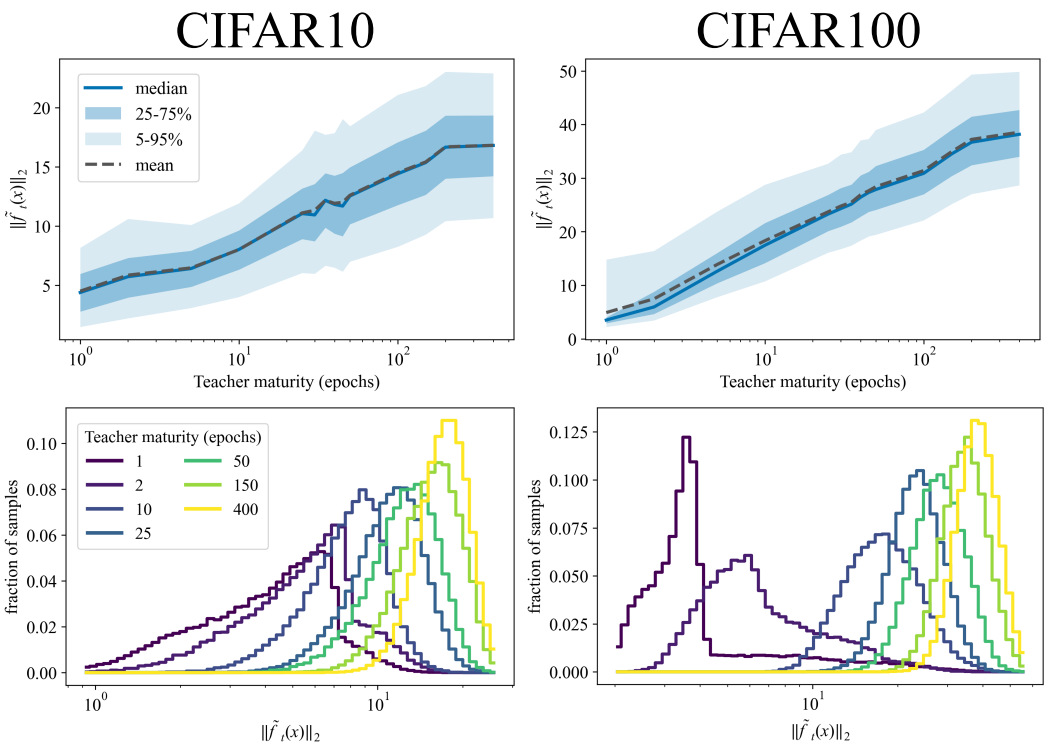

Figure D.1: Teacher centered logit norms on training inputs without augmentation on CIFAR-10 (left) and CIFAR-100 (right). Top: norm against teacher maturity; the line shows the median, the dashed line the mean, and the shaded bands the 25–75th and 5–95th percentiles. Bottom: histograms of the norm at selected maturities.

E LEARNING CURVES

Figure E.1 shows the learning curves of the teachers on CIFAR-10 and CIFAR-100. On both datasets, the training and test curves separate as training proceeds: training accuracy keeps rising above test accuracy, and training loss keeps decreasing below test loss, which levels off. On CIFAR-10, the training accuracy of the mature teacher nearly saturates, so its outputs nearly reproduce the training labels, as stated in Section 6.

Figure E.2 shows the test accuracy and test loss of students against training epoch; brighter colors indicate students distilled from more mature teachers. The curves of students distilled from the least mature teachers (purple) lie below the others in accuracy and above them in loss. The curves of students distilled from the most mature teachers (yellow) overlap with the others early in training, but their accuracy then plateaus. Moreover, unlike the other curves, their test loss shows double descent, and this tendency becomes gradually stronger as teacher maturity increases. The same tendency is observed on CIFAR-100.

Double descent in KD from mature teachers. Epoch-wise double descent (Nakkiran et al., 2021) was clearly observed in students distilled from mature teachers on CIFAR-10 and CIFAR-100 (Section 4.3). Our two-effect decomposition suggests a possible interpretation. As the teacher logit norm grows, alignment with the teacher's outputs becomes stronger relative to the suppression of the student logit norm. The student is therefore trained not only to predict the ground-truth labels but also to reproduce the teacher's output on each training example more closely. In standard double descent, test performance is known to degrade temporarily near the interpolation threshold, at which the model becomes able to fit the training data (Nakkiran et al., 2021). Applied to KD, the interpolation threshold may be regarded as the point at which the student becomes able to reproduce the teacher's outputs on the training data. Under this view, stronger alignment brings the student closer to this threshold, and the observed double descent may arise in this process. The capacity gap of Cho & Hariharan (2019), in which the student lacks the capacity to reproduce the mature teacher's outputs, may likewise be reconsidered not only as a teacher–student performance gap but also in terms of where the student lies relative to this threshold. Verifying this interpretation requires directly relating how closely the student reproduces the teacher's outputs to the onset of double descent.

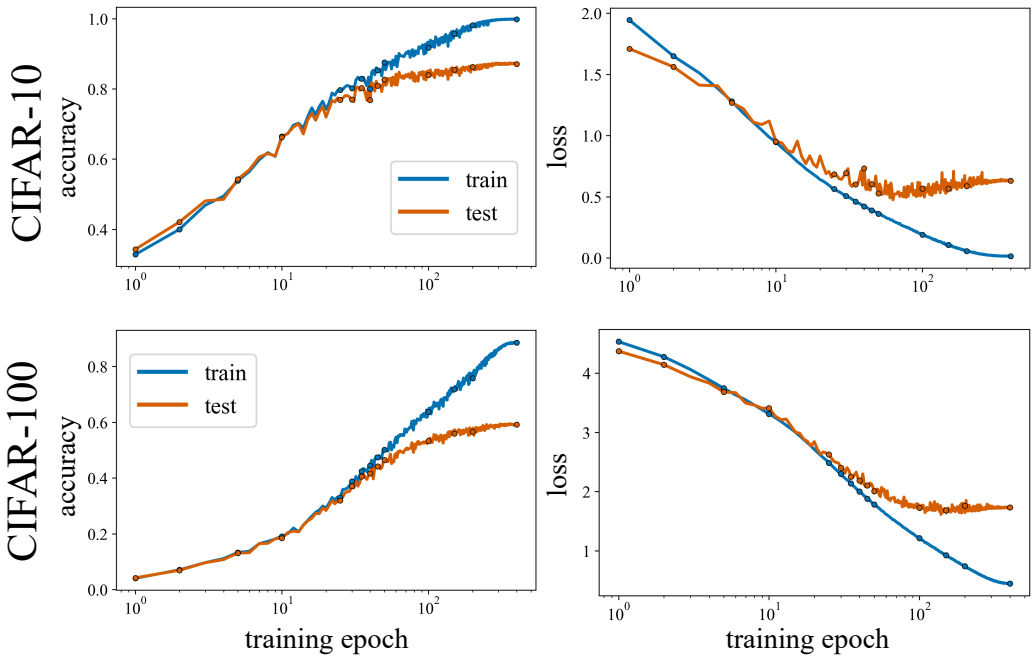

Figure E.1: Teacher learning curves on CIFAR-10 (top) and CIFAR-100 (bottom): accuracy (left) and loss (right) on the training and test sets against training epoch. Markers indicate the maturities used for distillation.

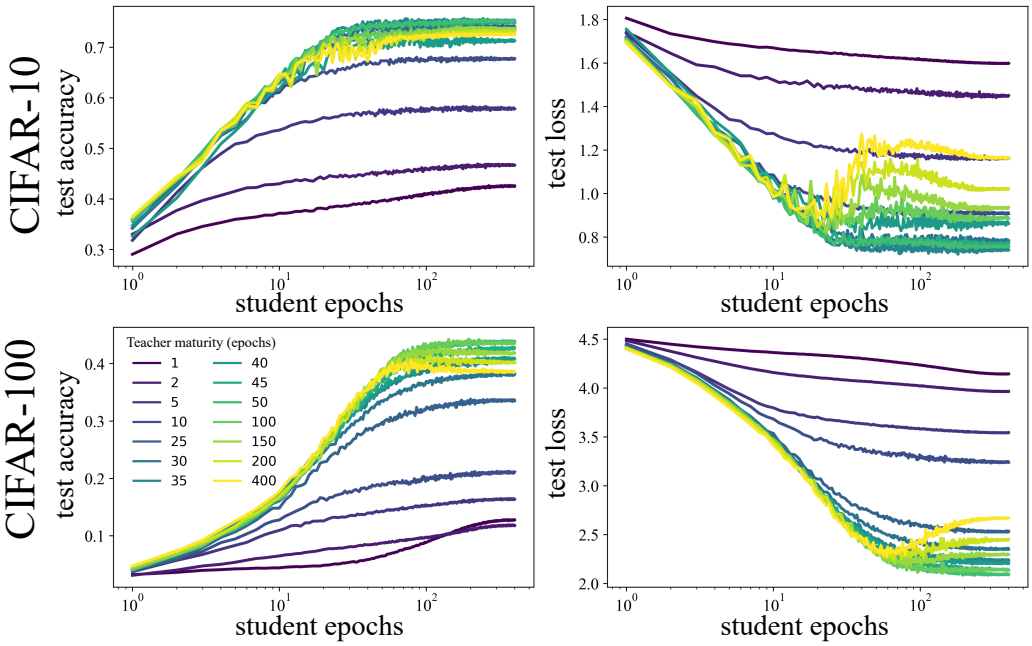

Figure E.2: Student learning curves on CIFAR-10 (top) and CIFAR-100 (bottom): test accuracy (left) and test loss (right) against training epoch. Colors indicate teacher maturity, from dark (immature) to yellow (mature), as in Figure 3.

F NORM REPLACEMENT ON AG NEWS

Figure F.1 shows the counterpart of Figure 6 on AG News, using the teacher at $M = 30$ epochs as the mature teacher. All other settings follow Section 5.1, except that the learning rate is constant, as in the other AG News experiments (Table C.1). In the top panel, the accuracy of students distilled from norm-replaced teachers is nearly the same at every reference maturity, with a slight improvement for the teacher whose norm is replaced with that of the 0.1-epoch teacher. For students distilled directly from reference teachers, in contrast, accuracy is far lower for the least mature teachers, and only the students of the 5- and 20-epoch teachers reach higher accuracy than the student of the most mature teacher. In the bottom panel, the test loss of students distilled from norm-replaced teachers (solid line) forms a U shape against the student logit norm, and its lowest point lies below the mature teacher. At small reference maturities, the dashed line lies above the solid line. Since the norm-replaced teacher and the reference teacher at the same r have the same norm and differ only in direction, this difference shows that the direction also affects the students. These results agree with those on CIFAR in Section 5.

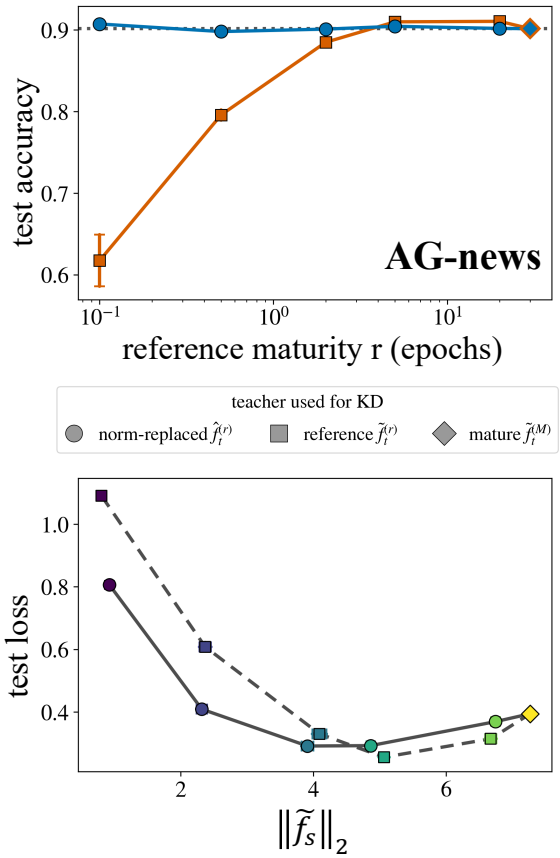

Figure F.1: Norm replacement on AG News. Top: test accuracy against reference maturity r for students distilled from norm-replaced teachers $\hat{f}_t^{(r)}$, reference teachers $\tilde{f}_t^{(r)}$, and the mature teacher $\tilde{f}_t^{(M)}$. Bottom: test loss against the student logit norm at the final epoch. Notation follows Figure 6.

G ADDITIONAL RESULTS AT DIFFERENT DISTILLATION TEMPERATURES

Figure G.1 shows the counterpart of Figure 5 on CIFAR-100, and Figure G.2 shows test loss against the teacher–student centered logit inner product on CIFAR-10 and CIFAR-100 at the same temperatures. All settings other than T and all random seeds are identical to those in Figure 3. Unlike on CIFAR-10, the student trajectories on CIFAR-100 move toward larger norms during training at $T \geq 2$ as well. Along each trajectory, however, test loss decreases, and the U-shaped relationship appears across the final points of different teacher conditions: the final points of teachers of intermediate maturity lie lowest, and those of the most mature teachers lie above them. The relationship with the inner product shows the same pattern on both CIFAR-10 and CIFAR-100 (Figure G.2).

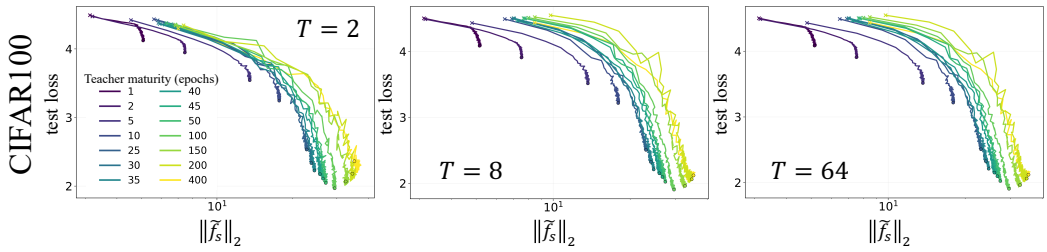

Figure G.1: Test loss against the student logit norm on CIFAR-100 at $T = 2, 8, 64$ (left to right). Crosses mark the first epoch and circles the final epoch. Notation follows Figure 5.

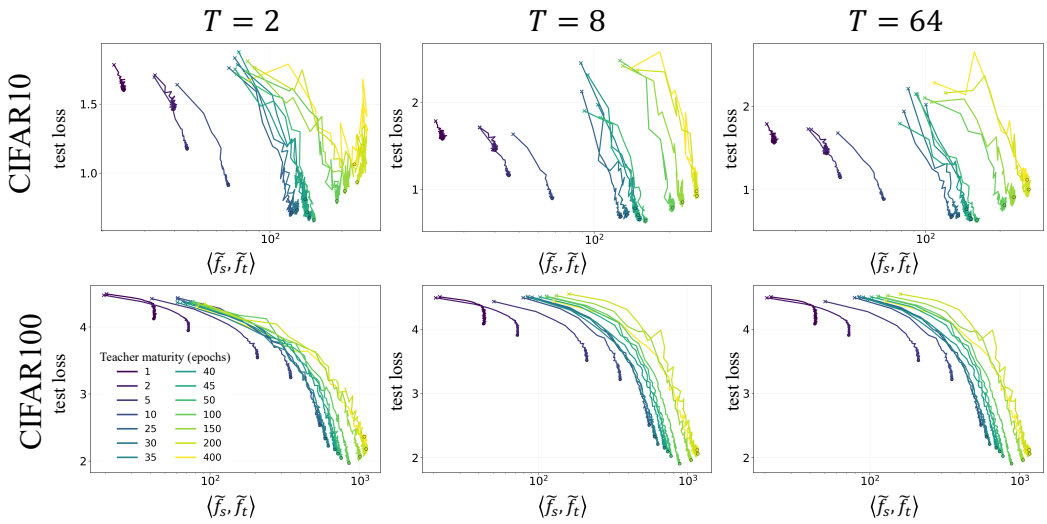

Figure G.2: Test loss against the teacher–student centered logit inner product on CIFAR-10 (top) and CIFAR-100 (bottom) at $T = 2, 8, 64$ (left to right). Notation follows Figure 5.

H RESULTS WITH OTHER ARCHITECTURES

To examine whether the results depend on the student architecture, we replace the student with WRN-16-2 on CIFAR-10, CIFAR-100, and Tiny ImageNet and with TextCNN on AG News, keeping the teachers unchanged; on CIFAR-10/100, these are the teachers trained with the cosine schedule of Section 5 (Table C.1). We repeat the experiments of Section 4; all other settings are unchanged.

Figure H.1 shows the TextCNN results on AG News, averaged over three seeds. As with the Transformer student, the trajectories of students distilled from teachers of different maturity nearly overlap on a single curve. Within the range of norms reached by the TextCNN students, however, this curve keeps decreasing, and the rise on the right side of the U does not appear. Moreover, the student accuracy curve levels off from early maturities and does not fall for mature teachers.

For WRN, we report a single seed (seed 0) on all three datasets (Figure H.2). On CIFAR-10 and CIFAR-100, the student accuracy curve peaks at intermediate maturity and falls for the mature teacher. In addition, the test loss trajectories nearly overlap on a U-shaped curve; these results are the same as with the ResNet students (Figure 3). On Tiny ImageNet, the trajectories also overlap, but the curve keeps decreasing up to the most mature teacher, and the U shape does not appear.

One possible explanation for the absence of the U shape on AG News with TextCNN and on Tiny ImageNet with WRN is the difficulty of the task for the student. Under our decomposition, immature teachers are beneficial when the strong alignment of the mature teacher moves the student beyond the minimum of the U. Whether the student reaches this point may depend on how difficult the task is for the student. On AG News, the task is easy compared with the other datasets. The accuracy curves of both teachers and students lie higher than on the other datasets and level off from early maturities. The degradation on the right side of the U arises from overly sharp predictions, which increase the loss on misclassified inputs (Section 6). On an easy task, few test inputs are misclassified, so this degradation is small, and there is little room for teachers of different maturity to differ in student accuracy. This may also explain the small improvement in accuracy on AG News with the Transformer student (Section 4.2). On Tiny ImageNet, in contrast, the task is difficult for the student. The student accuracy curve lies lower than on the other datasets, and test loss keeps decreasing up to the most mature teacher. We hypothesize that the students therefore remain on the left side of the U, where alignment is weak relative to regularization, and that stronger alignment with more mature teachers continues to improve them.

Previous studies have reported benefits of immature teachers also on complex tasks such as ImageNet (Cho & Hariharan, 2019; Wang et al., 2022). These studies trained students with data augmentation, which we disabled (Section 4.1). Data augmentation may raise the accuracy that the student reaches and thereby lower the effective difficulty of the task for the student. Students trained with it may therefore move beyond the minimum of the U even on complex tasks, where immature teachers become beneficial.

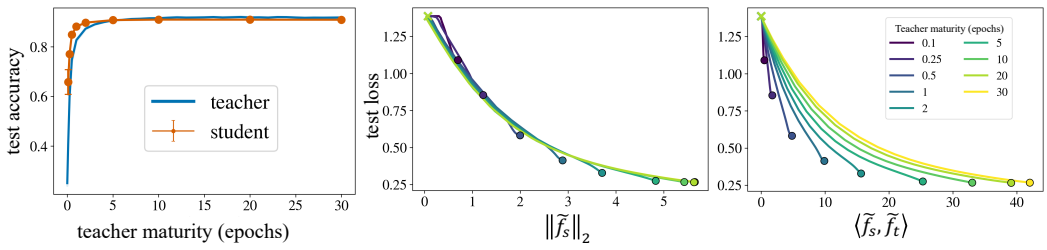

Figure H.1: Results with the TextCNN student on AG News (three seeds). Left: test accuracy of the teacher and the student against teacher maturity. Center and right: test loss against the student logit norm and the teacher–student inner product. Notation follows Figure 3.

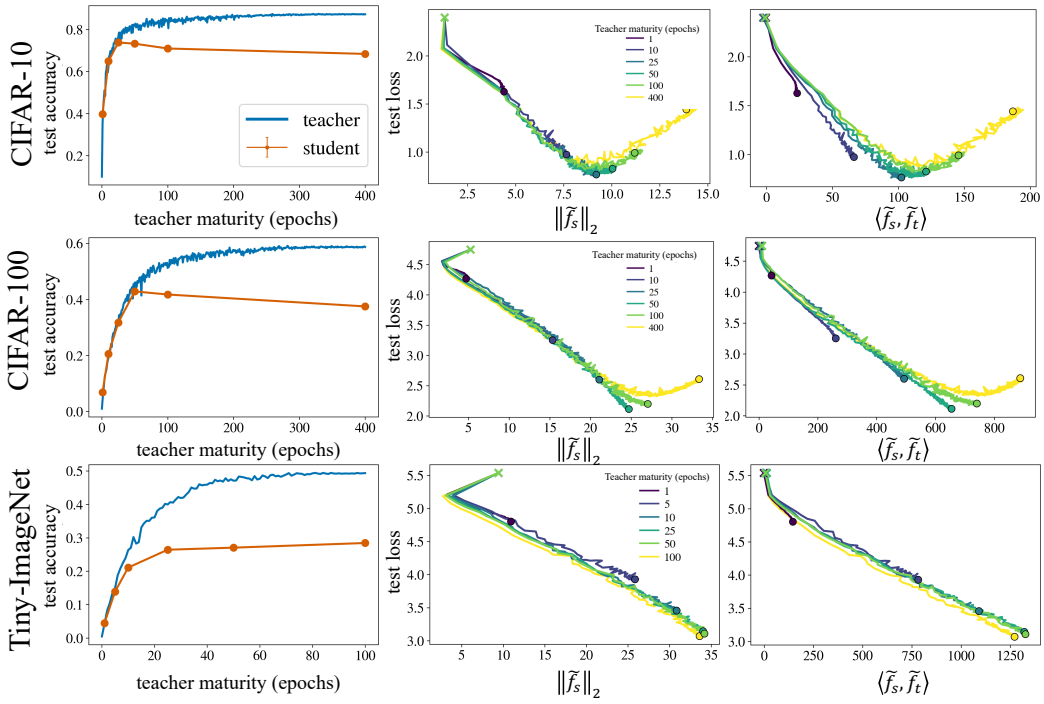

Figure H.2: Results with the WRN-16-2 student on CIFAR-10, CIFAR-100, and Tiny ImageNet (top to bottom), from a single seed (seed 0), so no error bars are shown. Columns follow Figure H.1.