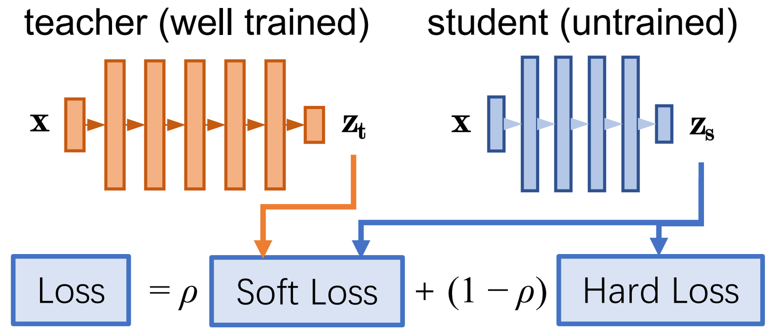
Knowledge Distillation in Wide Neural Networks: Risk Bound, Data Efficiency and Imperfect Teacher

Guangda Ji, Zhanxing Zhu ✉
Peking University

NEURAL INFORMATION PROCESSING SYSTEMS

Introduction

Knowledge distillation(KD)[1] is a model compression method that use a trained teacher network to train a smaller student network, so that the student can generalize better. However KD still lacks a satisfying explanation. In this work, we give theoretical analysis with recent neural tangent kernel[2,3] technique.



Contribution:

- We give an improved transfer risk bound that explains the fast convergence behavior of test error in pure soft label perfect distillation.
- We give a metric on the task's difficulty, called *data inefficiency*, and show that teacher's early stopping and higher soft ratio can reduce data inefficiency.
- In practical KD, the teacher is imperfect. We show that adding a little portion of hard label is necessary for better generalization.

Problem Setup

Problem: Binary classification problem with ground truth decision boundary $y_g = \mathbb{1}\{f_g(x) > 0\} \in \{0, 1\}$, and input distribution $P(x \in \mathbb{R}^d)$.

Method: Use gradient descent to train a student network $z_s = f(x; w)$, with w being its weights, and with loss of,

$$\mathcal{L}_{\mathrm{KD}} = \frac{1}{N}\sum_{n=1}^{N}\ell_n = \frac{1}{N}\sum_{n=1}^{N}\rho\overbrace{H(y_{t,n}, \sigma(\frac{z_{s,n}}{T}))}^{\text{soft loss}} + (1-\rho)\overbrace{H(y_{g,n}, \sigma(z_{s,n}))}^{\text{hard loss}},$$

- $H(p, q)$: binary cross-entropy loss,
- $z_{s,n} = f(x_n; w)$: student's logits,
- $\sigma(z)$: sigmoid function,
- ρ: soft ratio,
- $y_{t,n} = \sigma(z_{t,n}/T)$: teacher's soft labels,
- T: temperature.

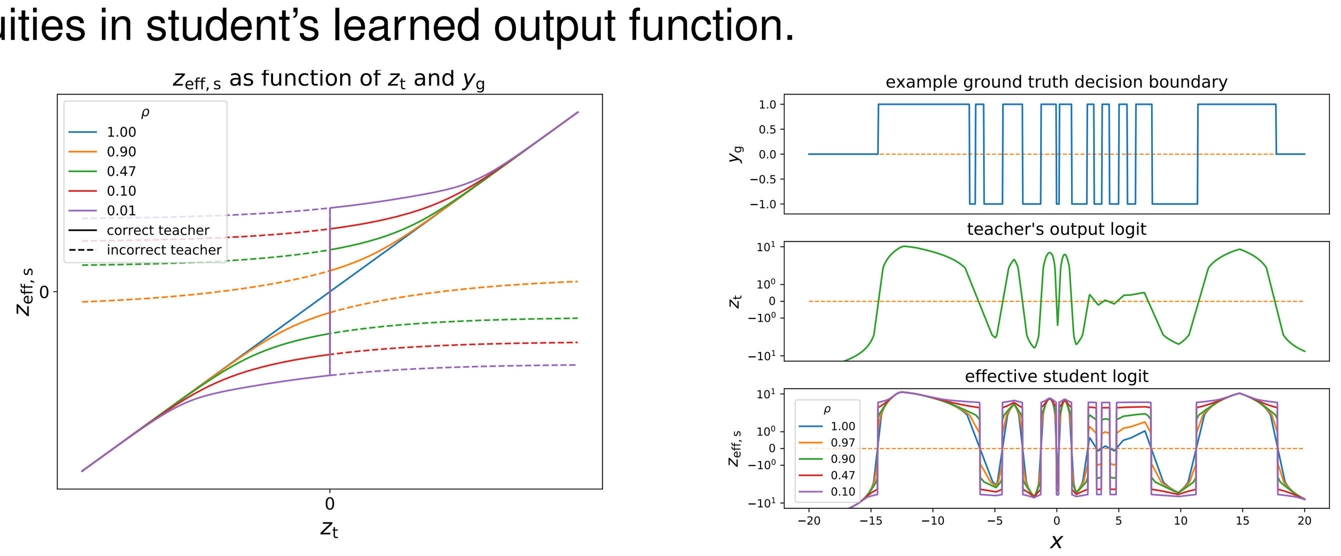
Assumption

The student is *wide* and *over-parameterized*.

Corollary 1: The student network converges to global minima[4] that minimize the loss at each sample point, so that,

$$\frac{d\ell}{d\hat{z}_s} = \frac{\rho}{T}(\sigma(\hat{z}_s/T) - \sigma(z_t/T)) + (1-\rho)(\sigma(\hat{z}_s) - y_g) = 0.$$

Solution to Eq. 1 defines an *effective student logit* $z_{s,\mathrm{eff}}(z_t, y_g)$. This results in discontinuities in student's learned output function.



Corollary 2: The student network can be approximated with its linear version,

$$f(x; w_{\mathrm{nlin}}) \approx f(x; w_0) + (w - w_0)^\top \partial_w f(x; w_0) = f(x; w_0) + \Delta_w^\top \phi(x),$$

- w_0: initial weight,
- Δ_w: weight change,
- $\phi(x)$: random feature,

and the converged student weight change is,

$$\Delta_{\hat{w}} = \phi(\mathbf{X})(\hat{\Theta}(\mathbf{X}, \mathbf{X}))^{-1}\Delta_{\mathbf{z}},$$

- $\hat{\Theta}(\mathbf{X}, \mathbf{X}) = \phi(\mathbf{X})^\top \phi(\mathbf{X})$: empirical tangent kernel, $\hat{\Theta}(\mathbf{X}, \mathbf{X}) \approx \Theta(\mathbf{X}, \mathbf{X})$,
- $\Theta(\mathbf{X}, \mathbf{X}) = \lim_{\mathrm{width}\to\infty} \hat{\Theta}(\mathbf{X}, \mathbf{X})$: neural tangent kernel(NTK).
- $\Delta_{\mathbf{z}} = \mathbf{z} - f(\mathbf{X}; w_0)$.

Result 1: Risk Bound

Assumption: Teacher network is perfect and can be represented by an oracle weight change Δ_{w_*} in random feature space.

Notation:

- Transfer risk $\mathcal{R} = \underset{x\sim P(x)}{\mathbb{P}}[z_t \cdot z_s < 0]$,
- $\bar{\alpha}(a, b) = \cos^{-1}(|a^\top b|/|a| \cdot |b|)$,
- Zero weight change Δ_{w_z}, $f(x; w_0) + \Delta_{w_z}^\top \phi(x) \approx 0$,
- Angle distribution $p(\beta) = \underset{x\sim P(x)}{\mathbb{P}}[\bar{\alpha}(\phi(x), \Delta_{w_*} - \Delta_{w_z}) > \beta]$, for $\beta \in [0, \pi/2]$.

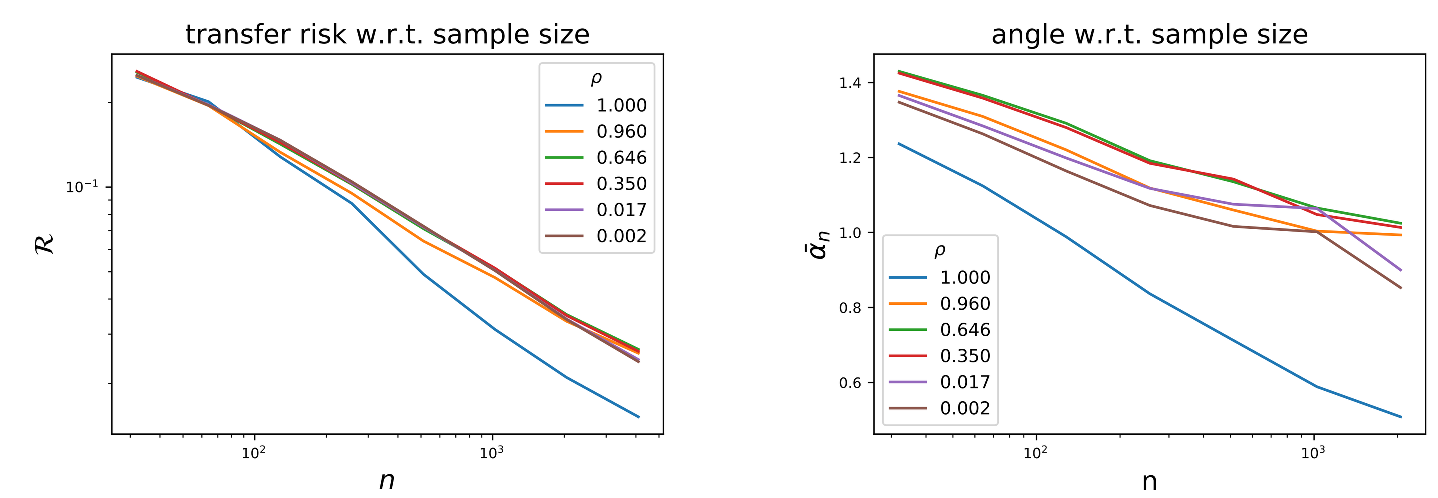
Theorem (Risk Bound)

Given n training samples $\mathbf{X} = [x_1, \cdots, x_n]$, denote $\bar{\alpha}_n = \bar{\alpha}(\Delta_{w_*} - \Delta_{w_z}, \Delta_{\hat{w}} - \Delta_{w_z})$, then the transfer risk is bounded by,

$$\mathcal{R}_n \le p(\pi/2 - \bar{\alpha}_n).$$

Key idea: The student learns a projection, $\Delta_{\hat{w}} = \phi(\mathbf{X})\Theta_n^{-1}\phi(\mathbf{X})^\top \Delta_{w_*} = \mathbf{P}_\Phi \Delta_{w_*}$, so that $\bar{\alpha}_n$ decreases with n and the wrong prediction area decrease. Smaller $\bar{\alpha}_n$ means better generalization.

Finding: Experimentally observed smaller generalization error on pure soft distillation can be explained by faster converging speed of $\bar{\alpha}_n$ w.r.t. n.



Our bound is improved compared to previous work.[5]

Result 2: Data Inefficiency

Definition

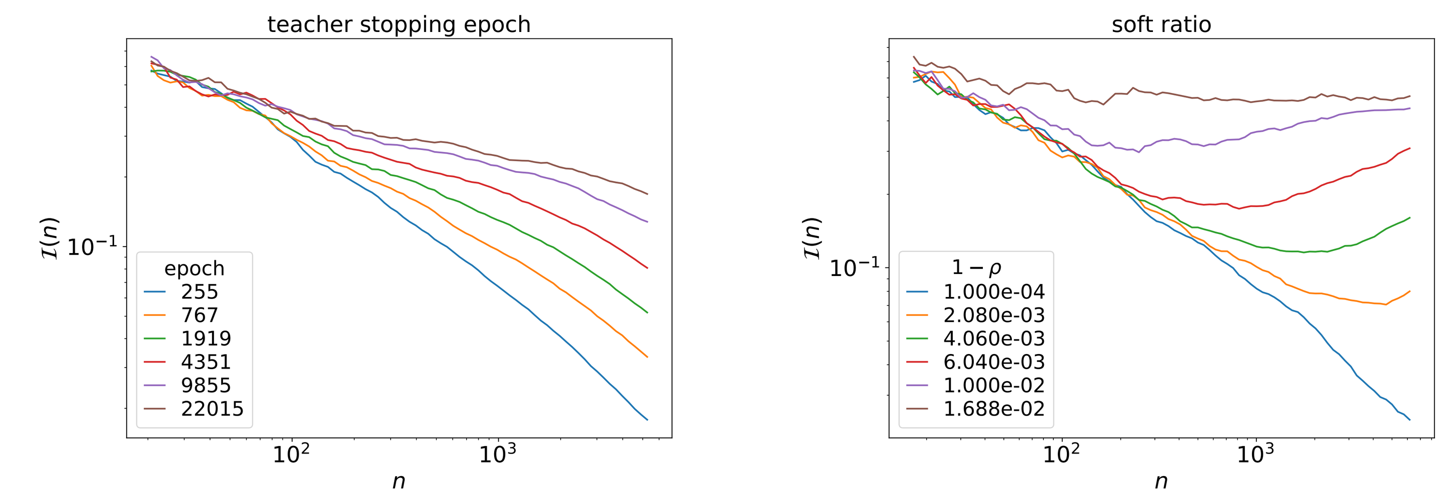
Data inefficiency is a discrete form of $\partial \ln ||\Delta_{\hat{w},n}||_2 / \partial \ln n$,

$$\mathcal{I}(n) = n\left[\ln \mathbb{E}||\Delta_{\hat{w},n+1}||_2 - \ln \mathbb{E}||\Delta_{\hat{w},n}||_2\right]$$

where $||\Delta_{\hat{w},n}||_2 = \sqrt{\Delta_{\mathbf{z}_n}^\top \Theta_n^{-1} \Delta_{\mathbf{z}_n}}$ is the norm of student's converged weight change trained by n samples.

Findings:

- Data inefficiency reveals the difficulty of weight recovery, which measures how well the student recovery the oracle weight with given amount of data.
- We use difficulty control experiments and demonstrate that data inefficiency is positively correlated to the difficulty of given task.
- In KD, two factors can reduce data inefficiency, both factors have smoothing effect on student's output function,
  - Early stopping of teacher,
  - Higher soft ratio ρ.



Result 3: Imperfect Teacher

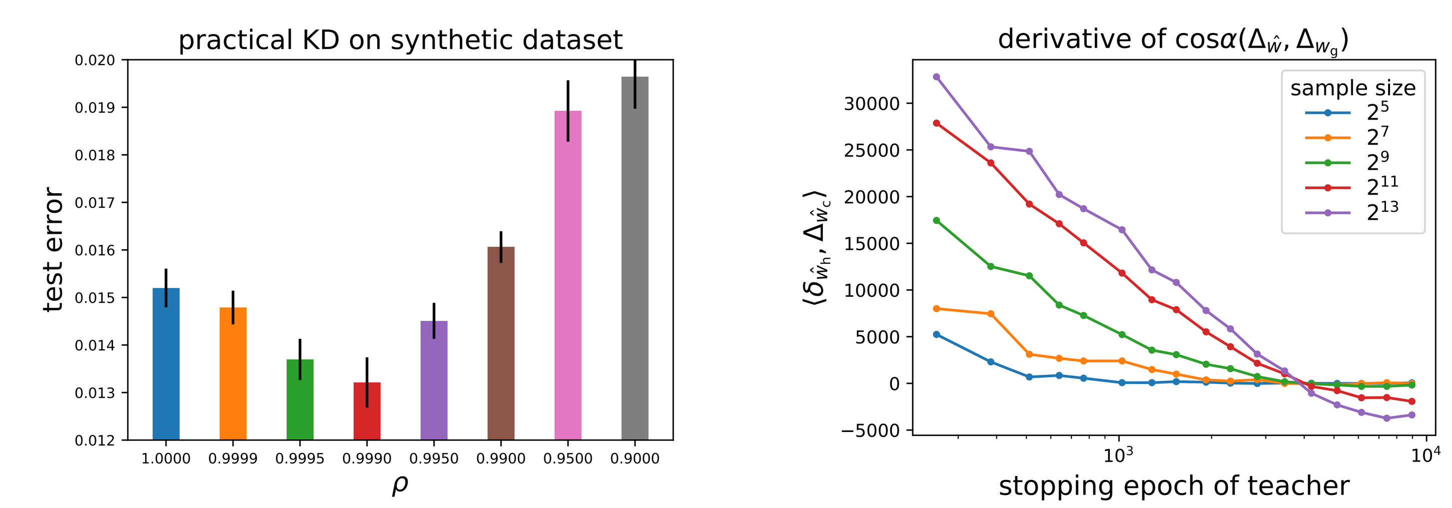
Findings: In practical KD, teacher is not perfect. Both real and synthetic experiments show that $\rho = 1$ is not optimal, and a little portion of hard label is needed.

Explanations:

- Locally: Effective student logits $z_{s,\mathrm{eff}}(z_t, y_g)$ has the function of moving the student logits closer to correct prediction region.
- Globally: Adding hard labels can reduce the angle $\alpha(\Delta_{\hat{w}}, \Delta_{w_g})$ between the weights of oracle and student,

$$\left.\frac{\partial \cos\alpha(\Delta_{\hat{w}}, \Delta_{w_g})}{\partial(1-\rho)}\right|_{\rho=1} \propto \left(\langle \Delta_{\mathbf{z}_g}, \delta\mathbf{z}_h\rangle_{\Theta_n} - \frac{\langle \Delta_{\mathbf{z}_g}, \Delta_{\mathbf{z}_t}\rangle_{\Theta_n}}{\langle \Delta_{\mathbf{z}_t}, \Delta_{\mathbf{z}_t}\rangle_{\Theta_n}}\langle \Delta_{\mathbf{z}_t}, \delta\mathbf{z}_h\rangle_{\Theta_n}\right) = \langle \delta_{\hat{w}_h}, \Delta_{\hat{w}_c}\rangle,$$

$\langle \delta_{\hat{w}_h}, \Delta_{\hat{w}_c}\rangle > 0$ when teacher makes more mistake than the best of hard label training student.



References

[1] Geoffrey Hinton et al. (2015). "Distilling the knowledge in a neural network". In: arXiv preprint arXiv:150302531 .

[2] Arthur Jacot et al. (2018). "Neural tangent kernel: Convergence and generalization in neural networks". In: Advances in neural information processing systems, pp. 8571–8580.

[3] Jaehoon Lee et al. (2019). "Wide neural networks of any depth evolve as linear models under gradient descent". In: Advances in neural information processing systems, pp. 8572–8583.

[4] Simon Du et al. (2019). "Gradient Descent Finds Global Minima of Deep Neural Networks". In: International Conference on Machine Learning, pp. 1675–1685.

[5] Mary Phuong et al. (2019). "Towards understanding knowledge distillation". In: International Conference on Machine Learning, pp. 5142–5151.