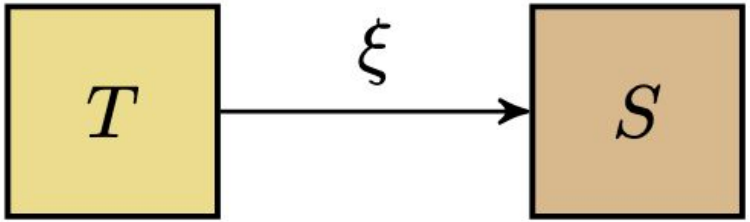
UNIVERSITY of WASHINGTON

NEURAL INFORMATION PROCESSING SYSTEMS

Understanding the Gains from Repeated Self-Distillation

Divyansh Pareek   Simon S. Du   Sewoong Oh

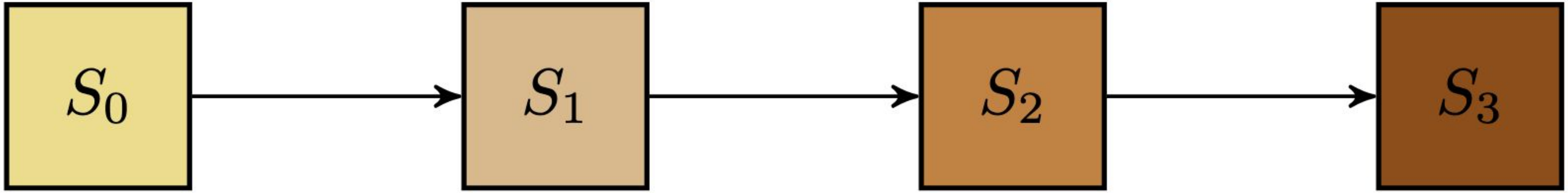
Self-distillation



$$\xi \cdot \boldsymbol{\ell}\big(\hat{y}_T, y_S(\theta)\big) + (1 - \xi) \cdot \boldsymbol{\ell}\big(y, y_S(\theta)\big)$$

- *same* architecture
- *same* training dataset
- only a different training objective

Self-distillation empirical gains



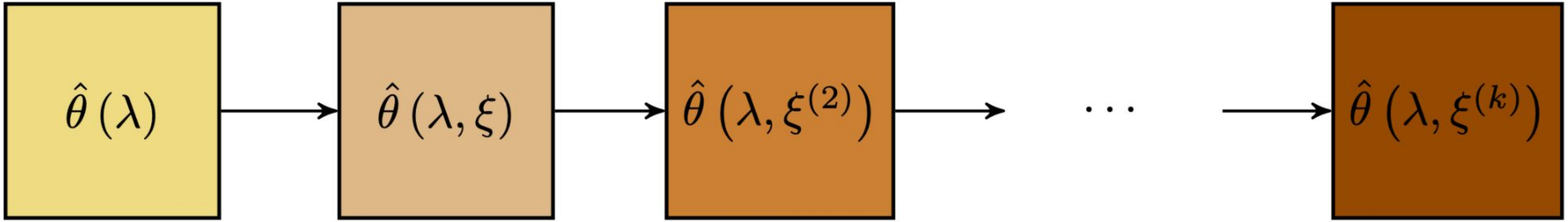
18.25 *17.61* *17.22* ***16.59***

- test error on CIFAR-100

Furlanello et al., "Born-Again Neural Networks". *(ICML 2018)*

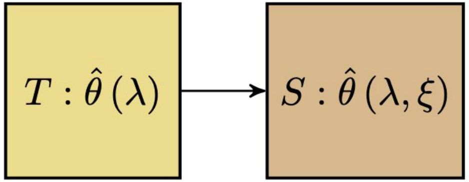
Question: *How much gain is possible by repeatedly applying self-distillation?*

Self-distillation under linear regression



- teacher is ridge with regularization λ.
- each step of self-distillation introduces a ξ param.

1-step result (previous work)



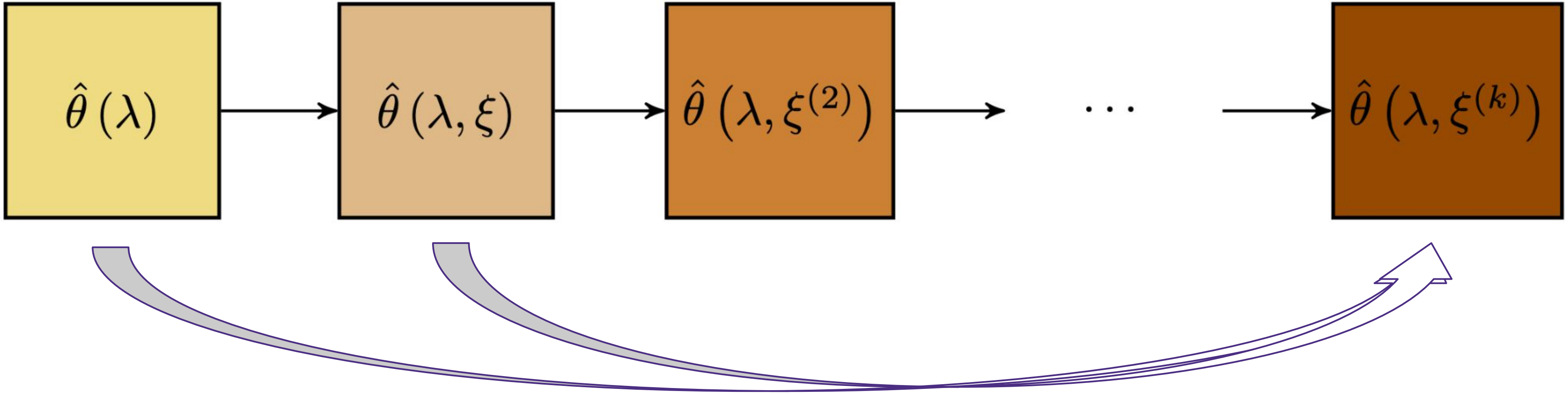
- **Result**: *optimal* S can have a strictly lower excess risk than *optimal* T.

$$\min_{\lambda>0} \operatorname{ExcessRisk}\left(\hat{\theta}(\lambda)\right) > \min_{\lambda>0,\xi\in\mathbb{R}} \operatorname{ExcessRisk}\left(\hat{\theta}(\lambda,\xi)\right)$$

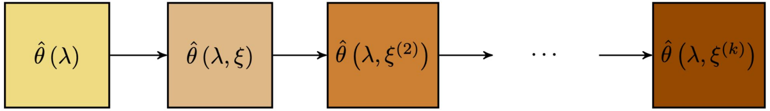
- The parameter ξ controls a bias-variance tradeoff. Increasing ξ reduces the variance term in the excess risk.

Das and Sanghavi, "Understanding Self-Distillation in the Presence of Label Noise". ***(ICML 2023)***

Question: Gains from multi-step SD?



What is the gain from running *k-1* additional steps of self-distillation?



General result (ours)

- **Main Theorem** *(informal)*: There exist a family of linear regression problem instances such that,

$$
\begin{aligned}
\text{there exist } \lambda > 0, \xi^{(r)} \in \mathbb{R}^r, &\quad \text{ExcessRisk}\left(\hat{\theta}(\lambda, \xi^{(r)})\right) \leq \frac{\gamma^2}{n} \;, \\
\text{for all } \lambda > 0, \xi \in \mathbb{R}, &\quad \text{ExcessRisk}\left(\hat{\theta}(\lambda, \xi)\right) \geq c_1 \cdot \frac{r\gamma^2}{n} \;, \\
\text{for all } \lambda > 0, &\quad \text{ExcessRisk}\left(\hat{\theta}(\lambda)\right) \geq c_0 \cdot \frac{r\gamma^2}{n} \;.
\end{aligned}
$$

- *r* denotes the rank of the input (design matrix **X**).
- *n* denotes the number of samples, γ^2 denotes the noise variance.

Discussion

- Mainly two conditions define the regime of separation.

1. θ^* is highly-aligned with one of the eigenvectors of $\mathbf{XX^T}$.
2. The ratio of eigenvalues $\lambda_1(\mathbf{XX^T}) / \lambda_r(\mathbf{XX^T})$ is $\Theta(1)$.

- We also show the necessity of these assumptions.

- Here $\mathbf{X}$ denotes the input design matrix (with rank r), and θ^* denotes the ground-truth parameter vector.

Experiments

- Regression tasks from the UCI repository.

Table 1: Test set MSE for optimal ridge and 1, 2-step SD.

Dataset	Optimal ridge	Optimal 1-step SD	Optimal 2-step SD
Air Quality	2.01	1.99	**1.06**
Airfoil	1.34	1.22	**1.19**
AEP	**0.62**	0.62	0.63

- We also verify that the AEP dataset does not satisfy the assumptions of the theorem.

Conclusion

- Optimal multi-step self-distillation can outperform optimal 1-step (or 0-step) self-distillation by a factor of upto $\Omega(d)$.

UNIVERSITY of WASHINGTON

Thank you

Poster at NeurIPS 2024: **#94147** in session 5, Friday, Dec 13 at 11am.