arxiv:2410.05464

Faster training with (progressive) distillation

Bingbin Liu
CMU → Simons → Kempner

Abhishek Panigrahi

Sadhika Malladi

Andrej Risteski

Surbhi Goel

Progress w/o massive compute?

Train small models faster, given big pretrained models?

Goal: Better efficiency.

- **Training**: fewer samples (statistical) / steps (computational).

System & training set	Train Frame Accuracy	Test Frame Accuracy
Baseline (100% of training set)	63.4%	58.9%
Baseline (3% of training set)	67.3%	44.5%
Soft Targets (3% of training set)	65.4%	57.0%

[Hinton et al. 15]

Progress w/o massive compute?

Train small models faster, given big pretrained models?

Goal: Better efficiency.

or: quantization, pruning.

- **Inference**: performant small models; "model compression".

Model	AIME 2024		MATH-500	GPQA Diamond	LiveCodeBench
	pass@1	cons@64	pass@1	pass@1	pass@1
QwQ-32B-Preview	50.0	60.0	90.6	54.5	41.9
DeepSeek-R1-Zero-Qwen-32B	47.0	60.0	91.6	55.0	40.2
DeepSeek-R1-Distill-Qwen-32B	**72.6**	**83.3**	**94.3**	**62.1**	**57.2**

[DeepSeek R1 report]

Distillation for faster training

arxiv:2410.05464

Background: what & how to distill.

- Theoretical explanation: generalization benefit.

Our work: Progressive distillation induces an implicit curriculum.

- Case study (sparse parity) + empirical confirmation

Future directions: better efficiency.

What is knowledge distillation?

Training a "student" model using a (trained) "teacher" model.

- Classification with cross-entropy loss: $f(x) \in \Delta^{k-1}, y \in [k]$.

Learn from data: $L_{CE}(f(x), y) = -\log[f(x)]_y = \mathrm{KL}(\delta_y \| f(x))$.

Distillation from f_T: $L_D(f(x), f_T(x)) = \mathrm{KL}(f_T(x) \| f(x))$.

reverse KL, l_2, etc.

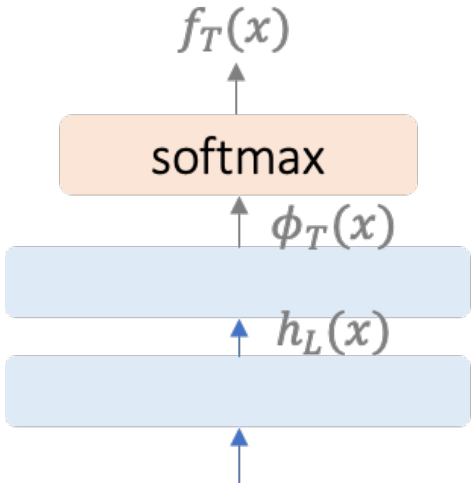
In practice, often use both: $\alpha L_{CE} + (1-\alpha) L_D$.

How to knowledge distillation?

Mimic the teacher's outputs or intermediate activations.

- Intermediate activation: need to match dimension.
- Post-softmax output $[f(x)]_i \propto \exp(\tau^{-1} \cdot [\phi(x)]_i)$.

(inverse) temperature



Various scenarios:

- Big/strong teacher → small/weak student. (today's focus)
- Same-sized teacher & student: self-distillation.
- Small/weak teacher → big/strong student: e.g. weak-to-strong generalization.
- An ensemble of teachers → a single student.

What makes distillation helpful?

Intuitively: "richer information" ... e.g. class relation, per-sample weighting.

- An ideal teacher: $f_T(x) = p^\star(y \mid x)$, i.e. providing the full label distribution.

more informative than the data
i.e. a single $y \sim p(\cdot \mid x)$.

System & training set	Train Frame Accuracy	Test Frame Accuracy
Baseline (100% of training set)	63.4%	58.9%
Baseline (3% of training set)	67.3%	44.5%
Soft Targets (3% of training set)	65.4%	57.0%

Better generalization

[Hinton et al. 15]

Theoretical work on better generalization

- Bayes distribution → smaller variance [Menon et al. 20]

$$R(f) \leq \hat{R}_*(f; S) + \mathcal{O}\left(\sqrt{\mathbb{V}_N^*(f) \cdot \frac{\log \frac{\mathcal{M}_N^*}{\delta}}{N}} + \frac{\log \frac{\mathcal{M}_N^*}{\delta}}{N}\right)$$

smallest for the Bayes predictor

- Imperfect teacher: bias-variance tradeoff.

  → e.g. label smoothing: $p^t(x) = (1-\alpha)e_y + \frac{\alpha}{L}\mathbf{1}$: larger α → smaller var / bigger bias.

Theoretical work on better generalization

- Bayes distribution → smaller variance [Menon et al. 20]
- Label smoothing provides regularization: "teacher-free" [Yuan et al. 19]

Model	Baseline	Tf-KD$_{reg}$	Normal KD [Teacher]
MobileNetV2	68.38	70.88 (**+2.50**)	71.05 (**+2.67**) [ResNet18]
ShuffleNetV2	70.34	72.09 (**+1.75**)	72.05 (**+1.71**) [ResNet18]
ResNet18	75.87	77.36 (**+1.49**)	77.19 (**+1.32**) [ResNet50]
GoogLeNet	78.15	79.22 (**+1.07**)	78.84 (**+0.99**) [ResNeXt29]

Theoretical work on better generalization

- Bayes distribution → smaller variance [Menon et al. 20]
- Label smoothing provides regularization: "teacher-free" [Yuan et al. 19]

(Regression: denoise / amplify top eigen-dirs [Mobahi et al. 20, Nagarajan et al. 23, Pareek et al. 24])

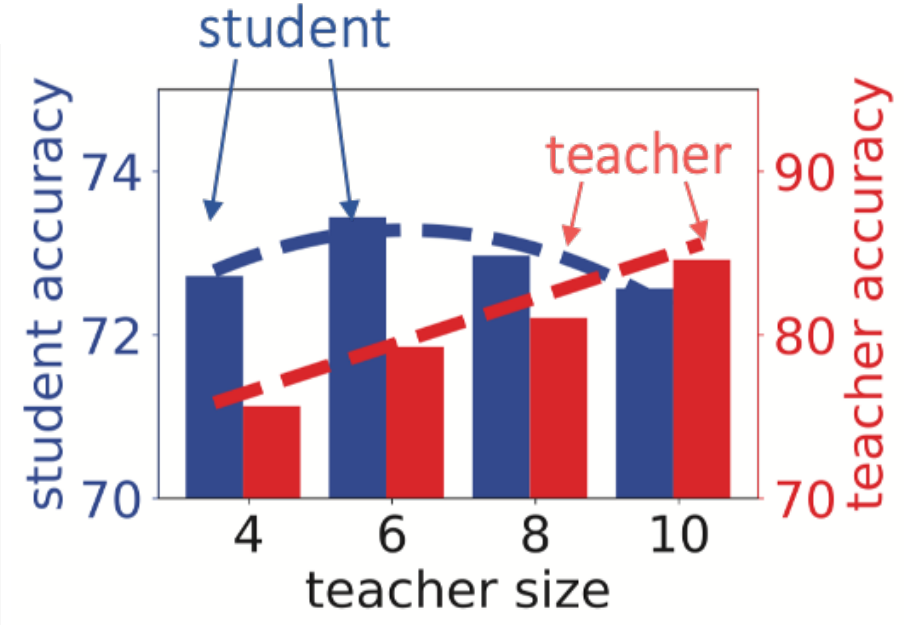
Not the full story: e.g. for sparse parity,

- Bayes distri. is one-hot, yet distillation helps.
- Label smoothing doesn't help.

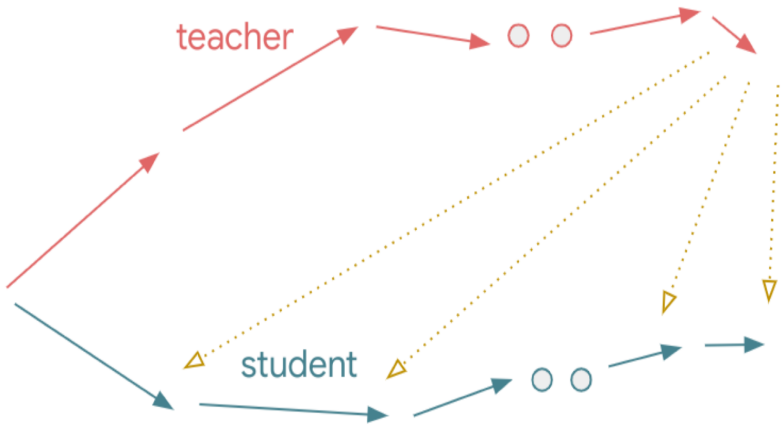
Better teacher → better student?

Better teacher ↛ better student

"capacity gap"
(due to differences in size / training steps)



[Mirzadeh et al. 19]



[Harutyunyan et al. 23]

Better teacher ↛ better student

"capacity gap"
(due to differences in size / training steps)

Model	Dataset	BLKD	TAKD
CNN	CIFAR-10	72.57	73.51
	CIFAR-100	44.57	44.92
ResNet	CIFAR-10	88.65	88.98
	CIFAR-100	61.41	61.82
ResNet	ImageNet	66.60	67.36

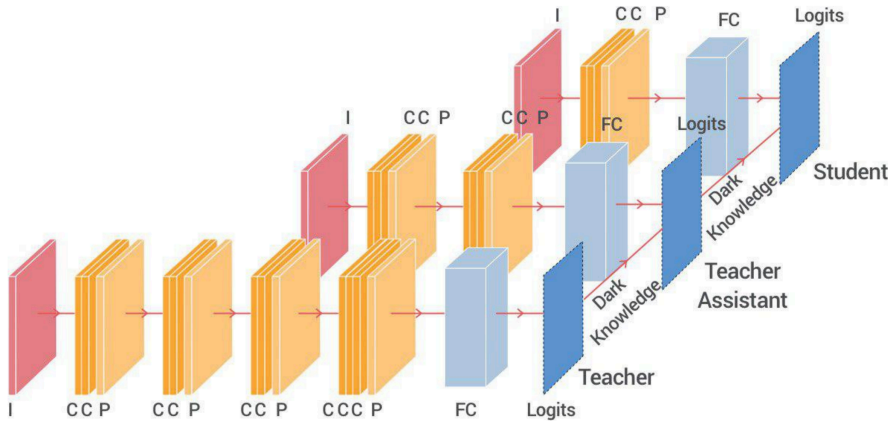
[Mirzadeh et al. 19]

	CIFAR-100	
ResNet-56 → LeNet-5x8	50.1 ± 0.4	61.9 ± 0.2
ResNet-56 → ResNet-20	68.2 ± 0.3	69.6 ± 0.3
ResNet-110 → LeNet-5x8	48.6 ± 0.8	60.8 ± 0.2
ResNet-110 → ResNet-20	67.8 ± 0.2	69.0 ± 0.3

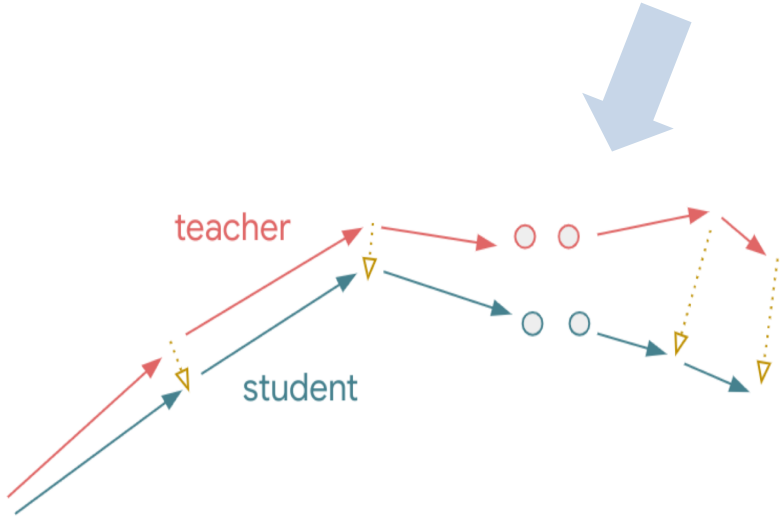
[Harutyunyan et al. 23]

Better teacher ↛ better student

"capacity gap"
(due to differences in size / training steps)



[Mirzadeh et al. 19]



[Harutyunyan et al. 23]

Why intermediate teachers help?

Prior work: smaller gap → better generalization (upper) bounds.

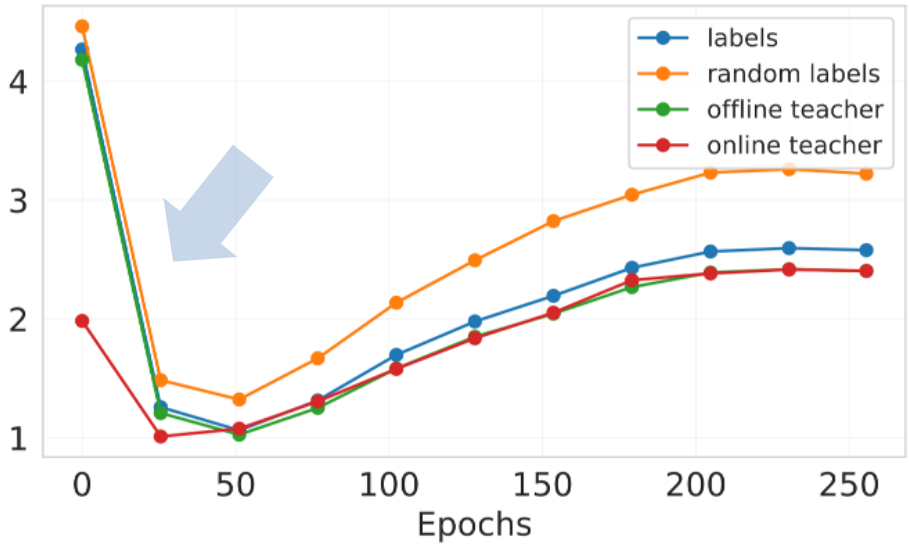
- Supervision complexity [Harutyunyan et al. 23]:

  For a kernel classifier,

$$f^* \in \operatorname*{argmin}_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i) + \frac{\lambda}{2} \|f\|_{\mathcal{H}}^2$$

$$\Rightarrow \quad \|f^*\|_{\mathcal{H}}^2 \leq \boldsymbol{Y}^\top K^{-1} \boldsymbol{Y}.$$

student teacher



Why intermediate teachers help?

Prior work: smaller gap → better generalization (upper) bounds.

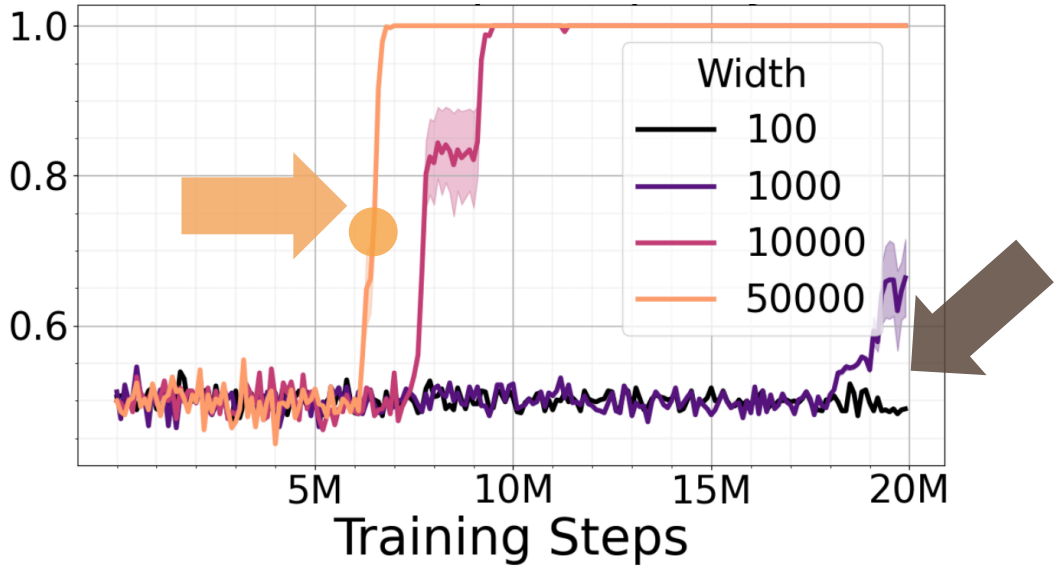
Our work: an **optimization** perspective.

- Intuition: using teacher's trajectory to guide the student's optimization.
- Case study: **sparse parity** … prior theory fails to explain the gain.
- Empirical validation on more realistic settings (PCFG and natural languages).

Case study: sparse parity

$$x = 1\ \underbrace{-1\ -1\ 1}_{S}\ -1\ 1\ 1\ 1\ -1\ 1 \quad \rightarrow y = \prod_{i \in S} x_i = 1$$

- Bigger model trains faster. [Edelman et al. 23]
  - SQ lower bound d^k [Kearns 98]
- Smaller models train as fast,
  when using intermediate checkpoints.

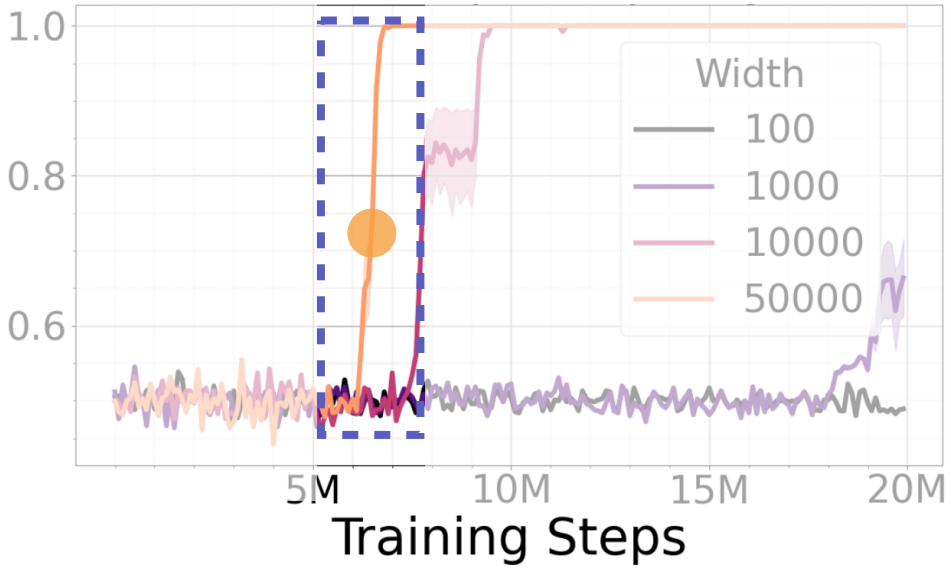


Case study: sparse parity

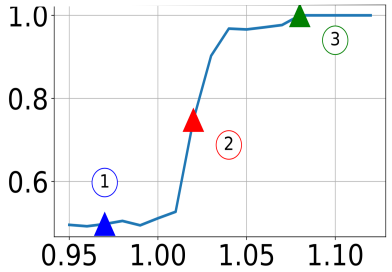
$$x = 1 \underbrace{\text{-1 -1 1}}_{S} \text{ -1 1 1 1 -1 1} \quad \rightarrow y = \prod_{i \in S} x_i = 1$$

- Bigger model trains faster. [Edelman et al. 23]
  - SQ lower bound d^k [Kearns 98]
- Smaller models train as fast,
  when using intermediate checkpoints.

extra training signals



Signals from intermediate teachers

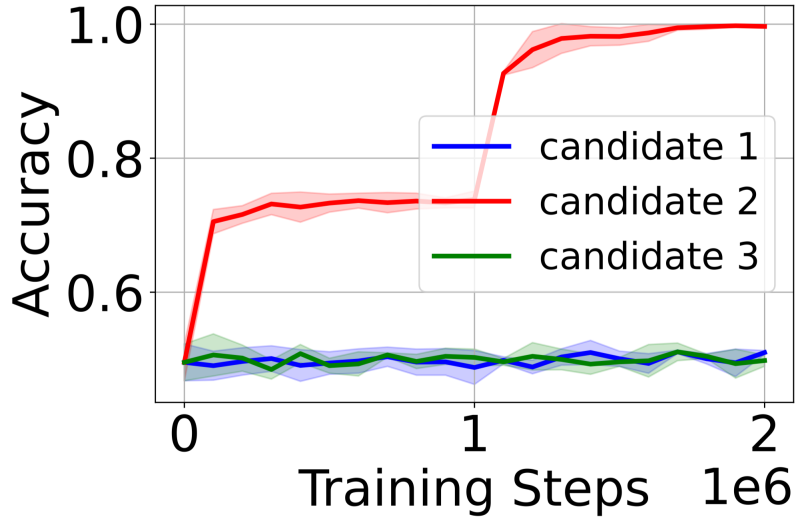


Setup: learning from 1 intermediate teacher + the final teacher.

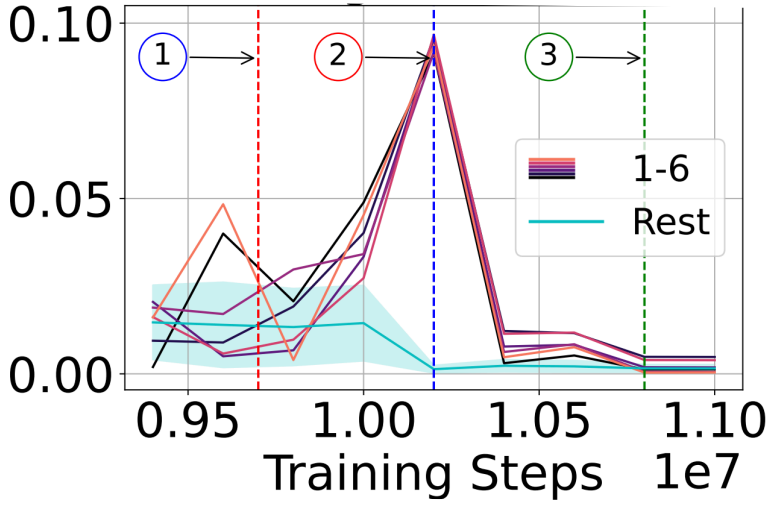
Compare 3 teacher checkpoints: before / during / after the phase transition.

Implicit curriculum

1. Serving as better supervision.



2. Providing "extra signals".

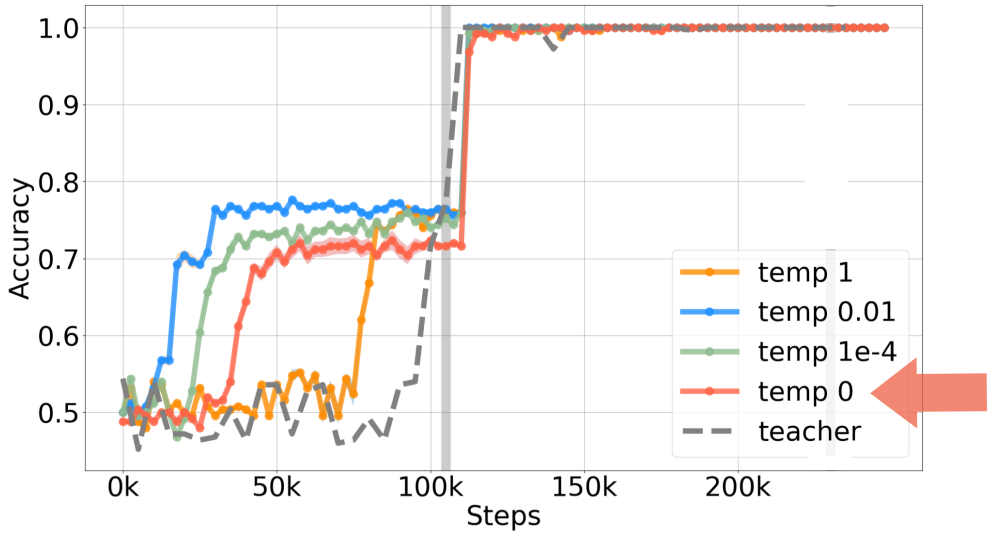


Implicit curriculum helps with optimization

- Case study: sparse parity
  - Note: can't be explained by soft logits: success at a low/zero temperature.

Recall: $[f(x)]_i \propto \exp(\tau^{-1} \cdot [\phi(x)]_i)$.

Smaller $\tau \rightarrow$ sharper logits.



Implicit curriculum helps with optimization

- Case study: sparse parity: speedup from "extra training signals."
  - What are the signals? … Fourier coefficients.
  - Why are they helpful? … sample complexity.
  - How do they emerge in the teacher? … initial population gradient.
- Progressive distillation & empirical validation

Training setup

Target: (d, k)-sparse parity: $y = \prod_{i \in S} x_i, x \in \{\pm 1\}^d, |S| = k.$

Model: 2-layer MLP: $f(x) = \sum_{j \in [m]} a_j \cdot \mathrm{ReLU}(\langle w_j, x \rangle + b_j).$

Training with the $\ell(f(x), y) = -f(x) \cdot y$ or $f_T(x)$ for the student.

- Teacher: 2-phase: initial large batch, followed by online SGD.
- Student: 2-shot distillation, from the end of each phase.

Signals: Fourier coefficients on x_i, $x \in [S]$

Fourier basis: monomials $\chi_{\tilde{S}}(x) := \prod_{i \in \tilde{S}} x_i$, for $\tilde{S} \subset [d]$.

- Natural for sparse parity: $y_S = \chi_S$.

- Fourier coefficients = projections onto the basis:

$$\hat{f}_{\tilde{S}}(f) = \langle \chi_{\tilde{S}}, f \rangle = \mathbb{E}_x[\chi_{\tilde{S}}(x) \cdot f(x)].$$

Signals: Fourier coefficients on x_i, $x \in [S]$

Our focus: $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in [d]$).

Learning from $y = \chi_S(x) \rightarrow \Omega(d^k)$ samples (recall: $|S| = k$).

Learning from $\sum_i \chi_{\{i\}} \rightarrow \Omega(d)$ samples.

$\tilde{\Theta}_{k,\epsilon}(d^2)$ for student's 2-shot distillation.

- Why: sample complexity to learn $\chi_{\tilde{S}}$: $\Omega(d^{|\tilde{S}|})$ (SQ lower bound).

$\rightarrow$ Fewer samples for learning lower-degree monomials [Edelman et al. 22, Abbe et al. 23].

Signals: Fourier coefficients on $x_i, x \in [S]$

Our focus: $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in [d]$).

- How: population gradient at initialization [Edelman et al. 22].

  Consider a single neuron $w \in \mathbb{R}^d$:

$$f(x) = \sigma(w^\top x + b)$$
$$l(y, y') = -yy'$$

$$-\widehat{\text{LTF}}_{S'} \leftarrow g_i := (\nabla_w \mathbb{E}_x[l(y, f(x; w))])_i = -\nabla_w \mathbb{E}_x[1[w^\top x + b \geq 0] \cdot yx_i]$$

$$= -\mathbb{E}_x[\underbrace{1[w^\top x + b \geq 0]}_{\text{LTF}} \cdot (\prod_{j \in S} x_j) \cdot x_i]$$

Fact: $|\widehat{\text{LTF}}_{S_1}| > |\widehat{\text{LTF}}_{S_2}|$

for odd $|S_1|, |S_2|$ s.t. $|S_1| < |S_2|$.

$\chi_{S'}$,

$S' = S \backslash \{i\}$ (if $i \in S$) or $S \cup \{i\}$ (if $i \notin S$)

Signals: Fourier coefficients on $x_i, x \in [S]$

Our focus: $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in [d]$).

- How: population gradient at initialization [Edelman et al. 22].

  Consider a single neuron $w \in \mathbb{R}^d$:

$$f(x) = \sigma(w^\top x + b)$$
$$l(y, y') = -yy'$$

$$-\widehat{\text{LTF}}_{S'} \leftarrow g_i := (\nabla_w \mathbb{E}_x[l(y, f(x; w))])_i = -\nabla_w \mathbb{E}_x[1[w^\top x + b \geq 0] \cdot yx_i]$$

$$= -\mathbb{E}_x[1[w^\top x + b \geq 0] \cdot (\prod_{j \in S} x_j) \cdot x_i \,] \Rightarrow |g_i| \geq |g_j| + \gamma_k, i \in S, j \notin S\,.$$

(Fourier gap)

large gradients → support

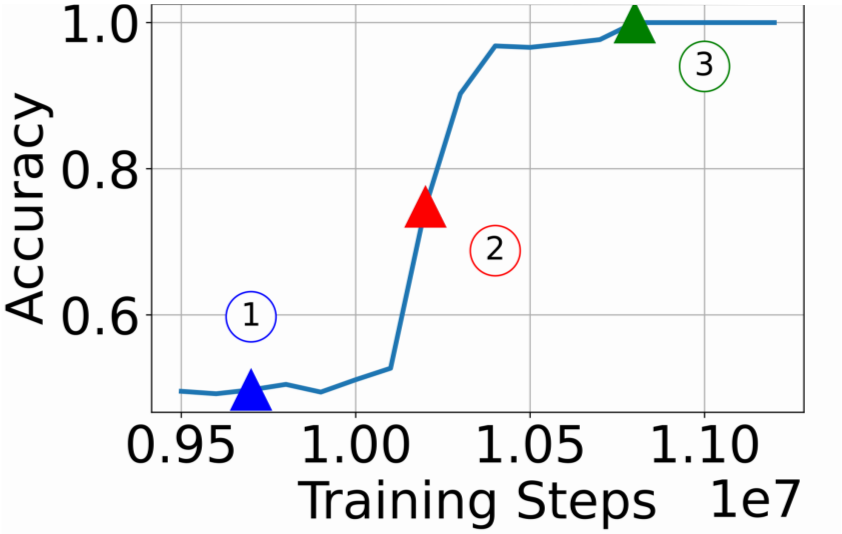
Fact: $|\widehat{\text{LTF}}_{S_1}| > |\widehat{\text{LTF}}_{S_2}|$ for odd $|S_1|, |S_2|$ s.t. $|S_1| < |S_2|$.

$\chi_{S'}$, $S' = S \backslash \{i\}$ (if $i \in S$) or $S \cup \{i\}$ (if $i \notin S$)

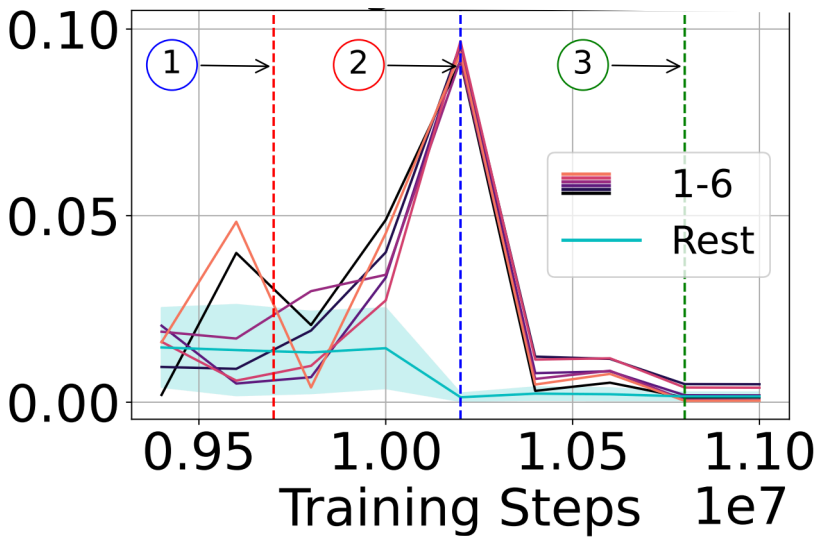
Signals: Fourier coefficients on $x_i, x \in [S]$

Our focus: $\hat{f}_{\tilde{S}}$, for singleton $\tilde{S}$ (i.e. $\{i\}, i \in [d]$).

- How: population gradient at initialization exhibits a **Fourier gap** [Edelman et al. 22].



larger $|\chi_{\{i\}}|$ for $i \in S$



Implicit curriculum helps with optimization

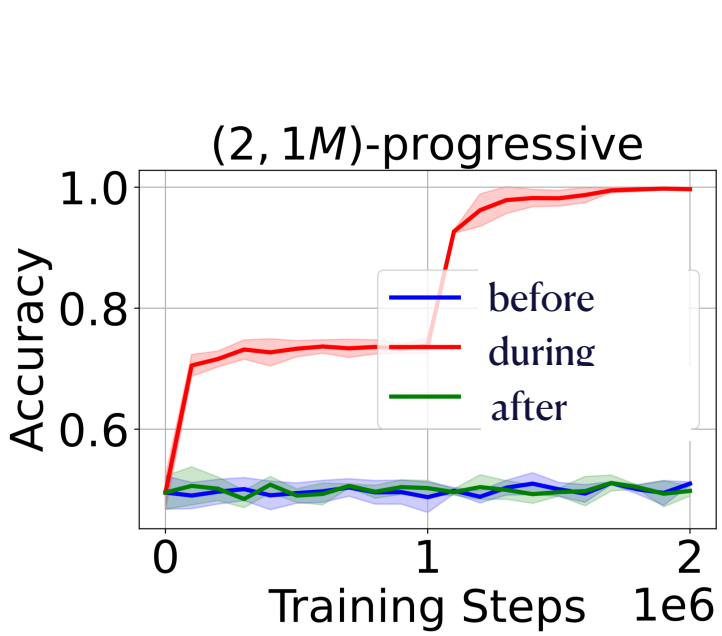
- Sparse parity: faster feature learning from "extra training signals."
  - **What** the curriculum are: Larger Fourier coeffs on x_i, $i \in [S]$.
  - **Why** they are helpful: sample complexity $\Omega(d^k) \to \tilde{\Theta}(2^k d^2/\epsilon^2)$.
  - **How** they emerge: Initial population gradient reveals the support.

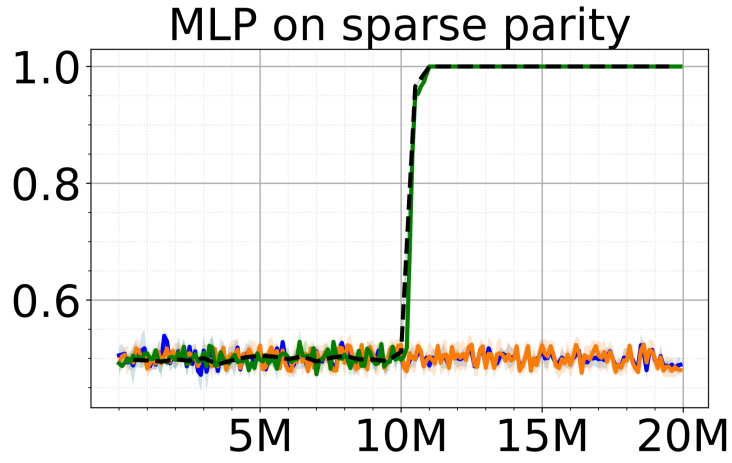
Implicit curriculum: a helpful decomposition.

Next: progressive distillation & empirical validation

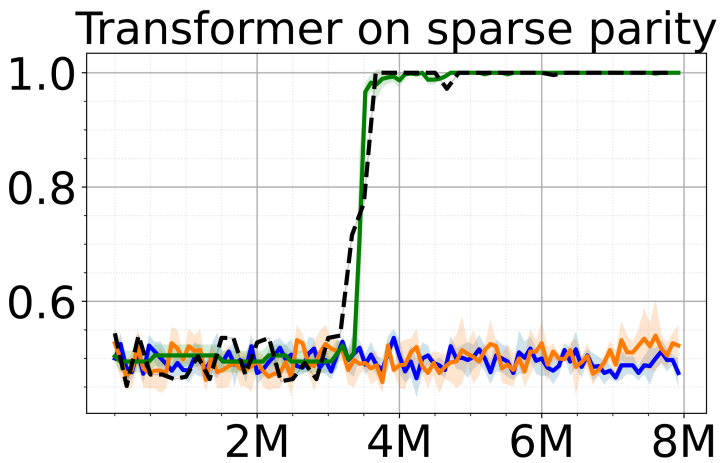
Progressive distillation

Distilling from checkpoints at certain intervals.





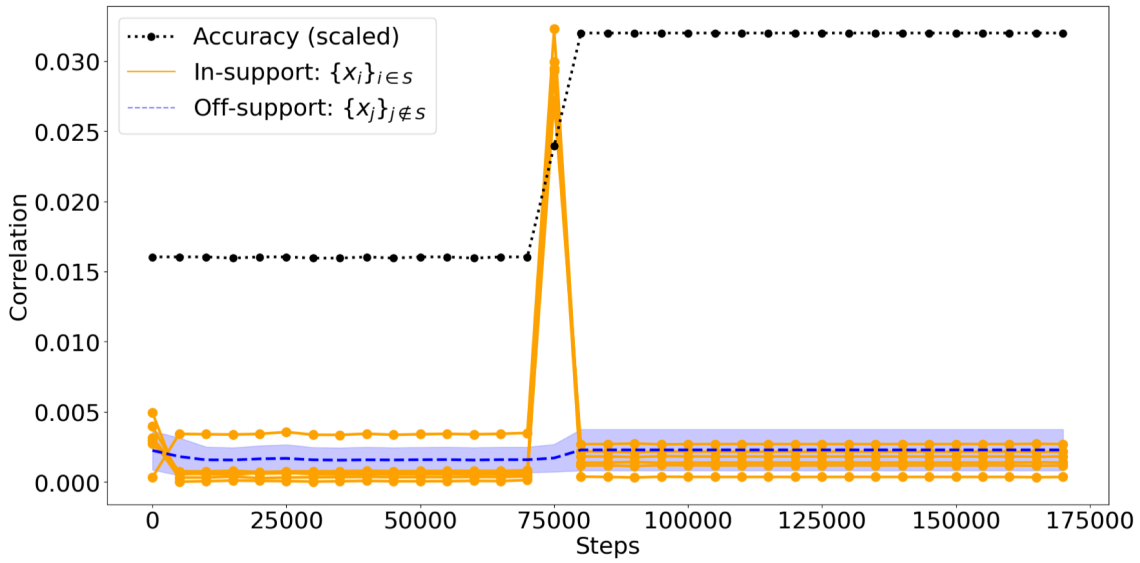
at 100k intervals

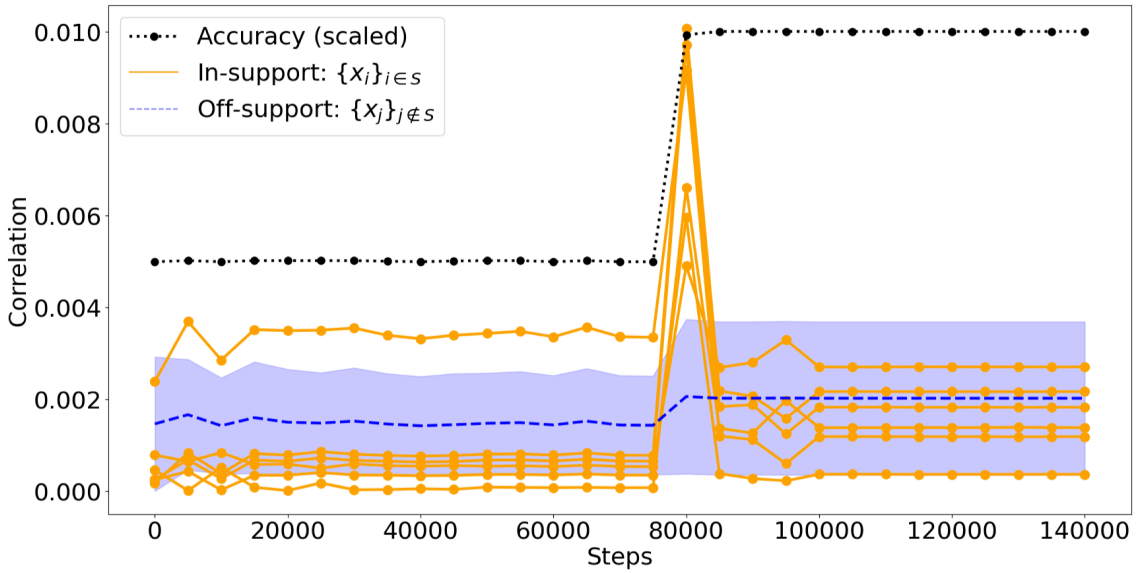


at 10k intervals

Transformer on sparse parity

1. Implicit curriculum emerges: Higher $\hat{f}_{\{i\}}$ for $i \in S$.

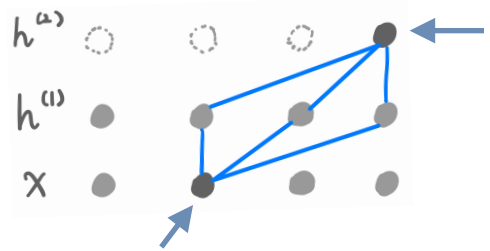


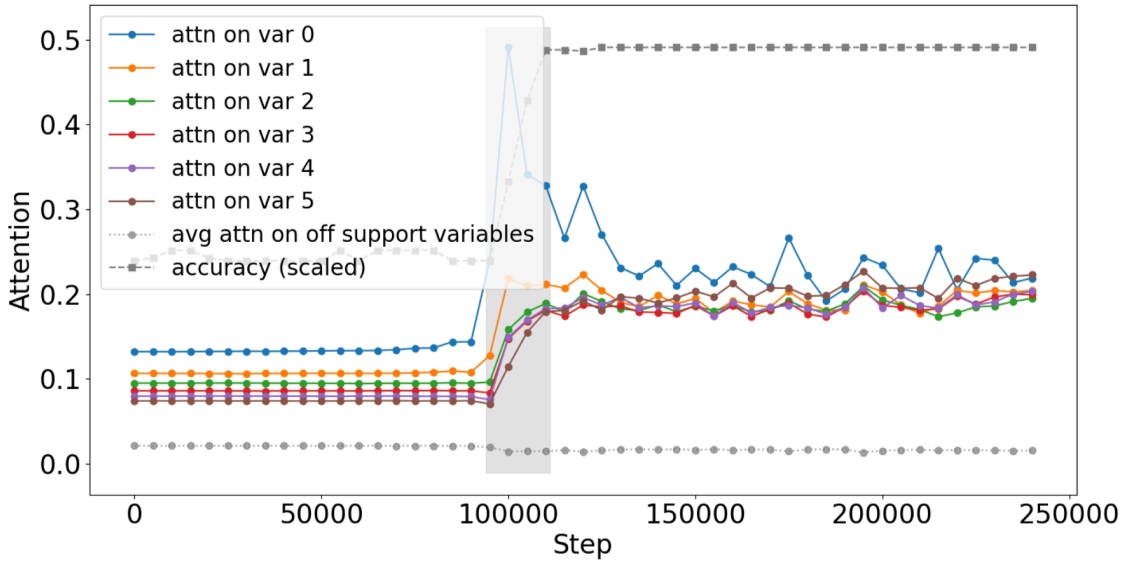


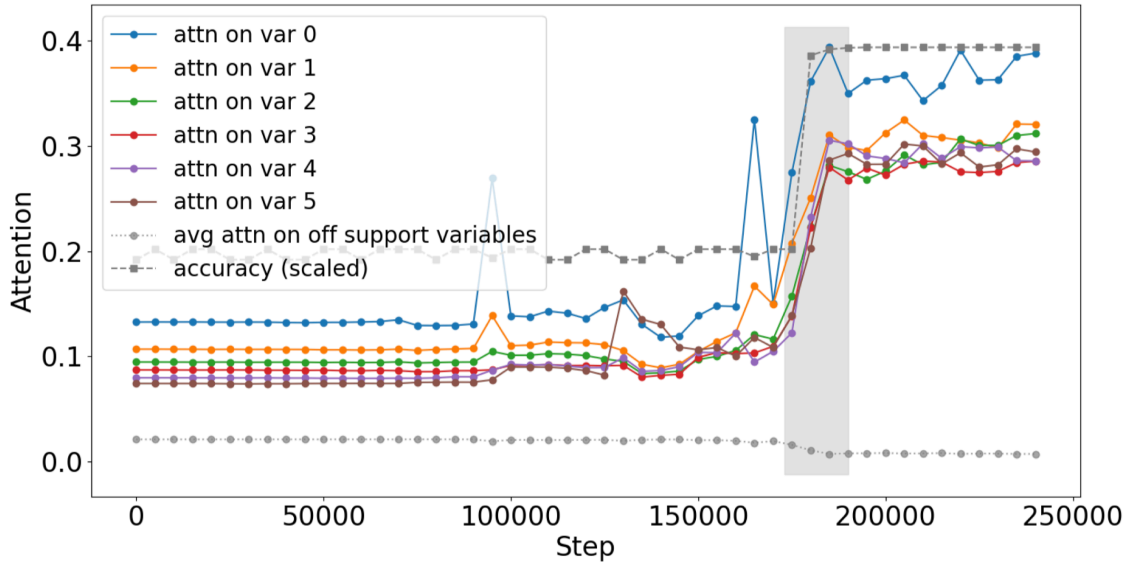
Transformer on sparse parity

2. True support is learned: more attention weights on in-support coordinates.

sum of length-2 paths

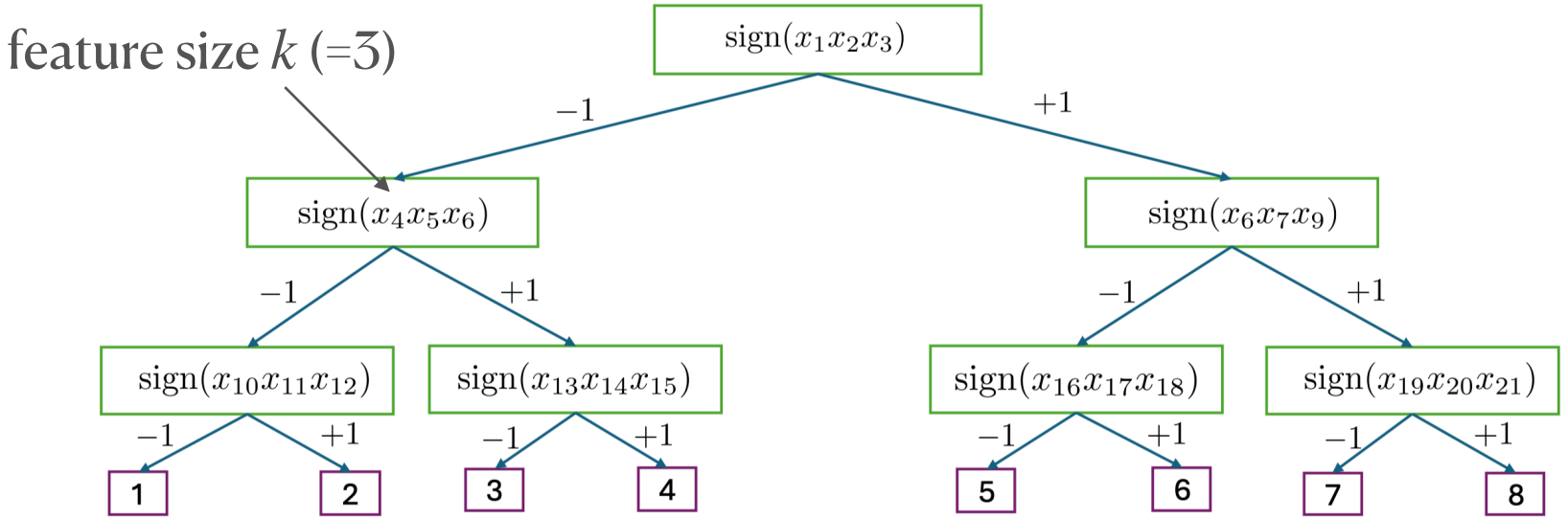






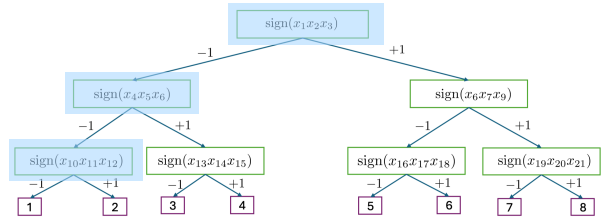
Beyond sparse parity — a hierarchical task

Hierarchical parity ... depth-D $\rightarrow$ 2^D-way classification.

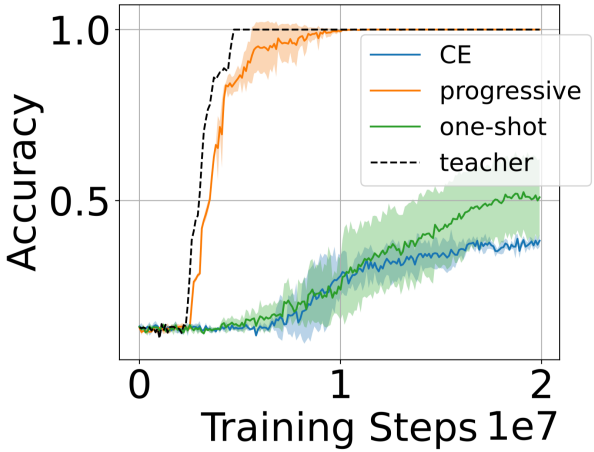


Beyond sparse parity — a hierarchical task

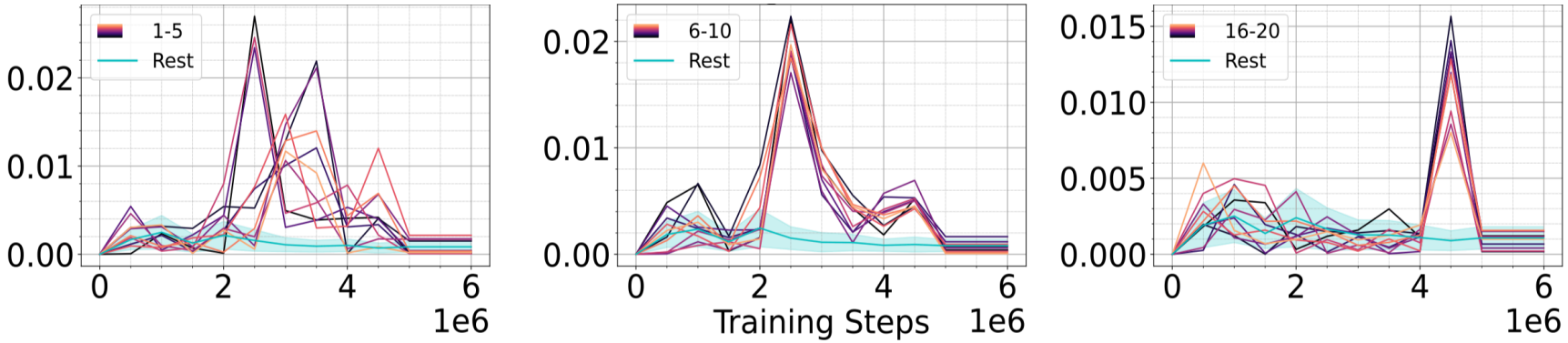
Hierarchical parity ... depth-D $\rightarrow$ 2^D-way classification.



- Results on $d = 100, D = 3, k = 5$:



Corr. to degree-2 monomials

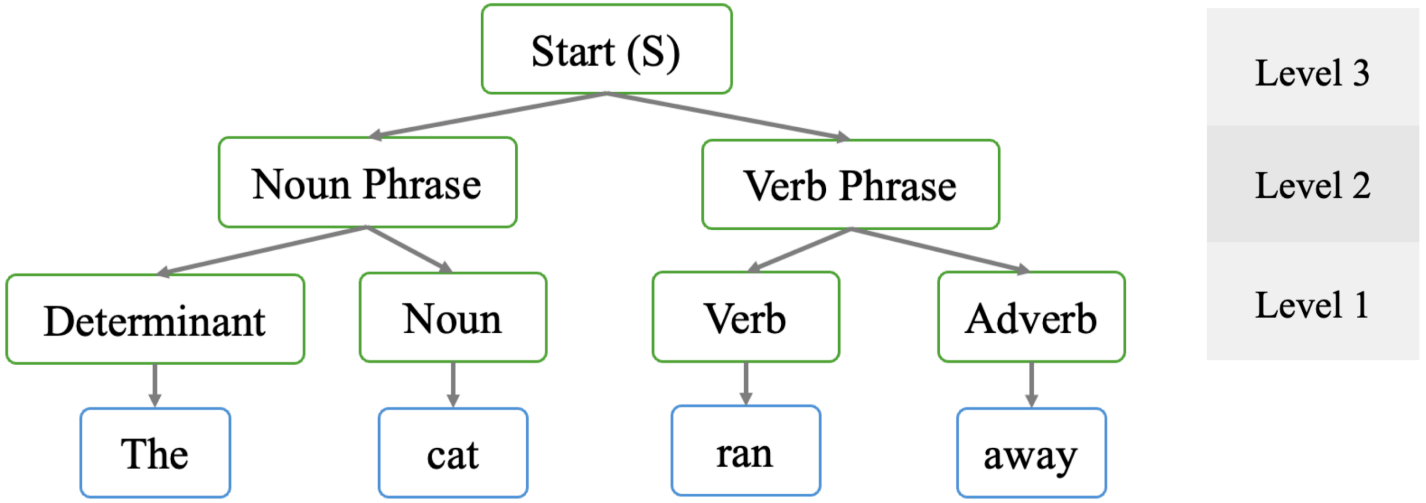


Learning at diff speed → need multiple teachers.

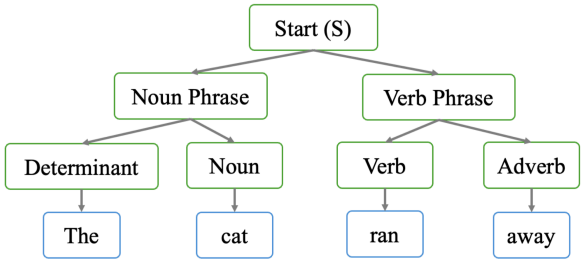
Beyond sparse parity — PCFG

Data: PCFG (probabilistic context-free grammar) [Allen-Zhu & Li 23]

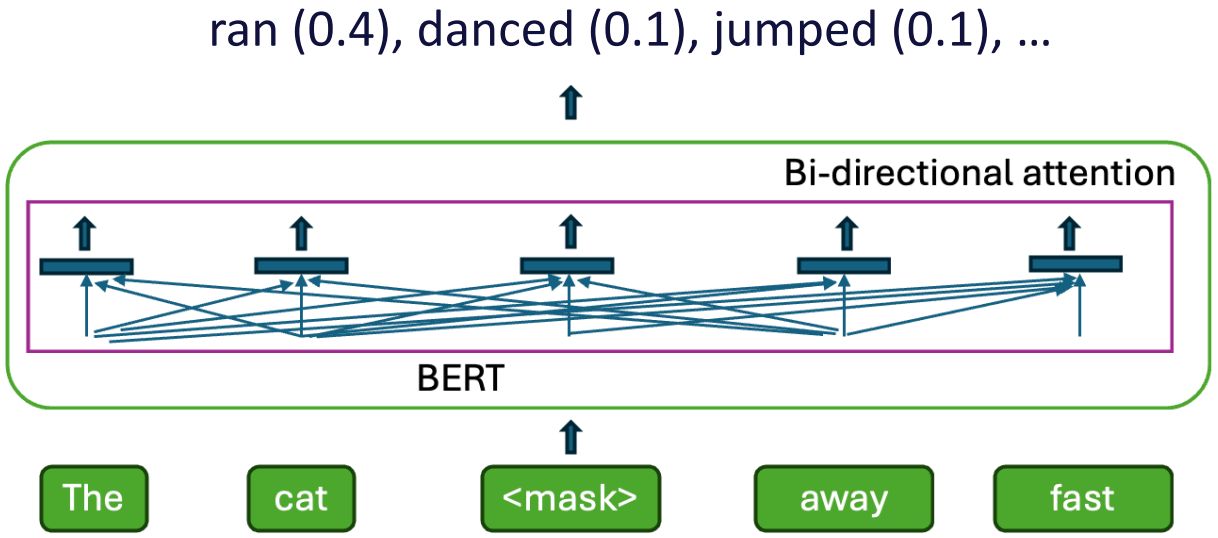
- Defined by: vocab; non-terminals; rules & probabilities.



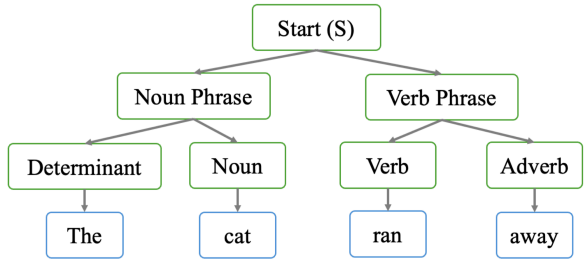
Beyond sparse parity — PCFG



Task: masked prediction … loss averaged over the masked set.



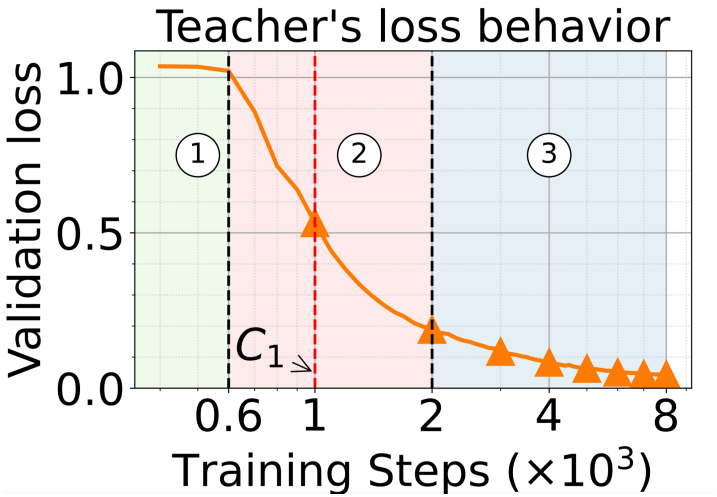
Beyond sparse parity — PCFG

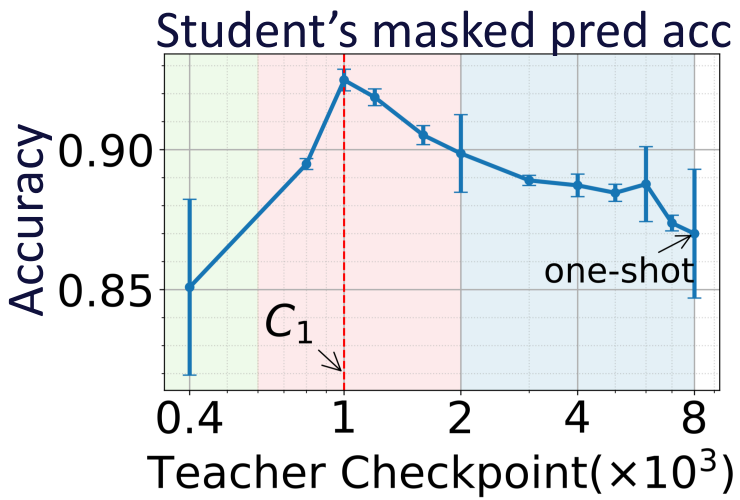


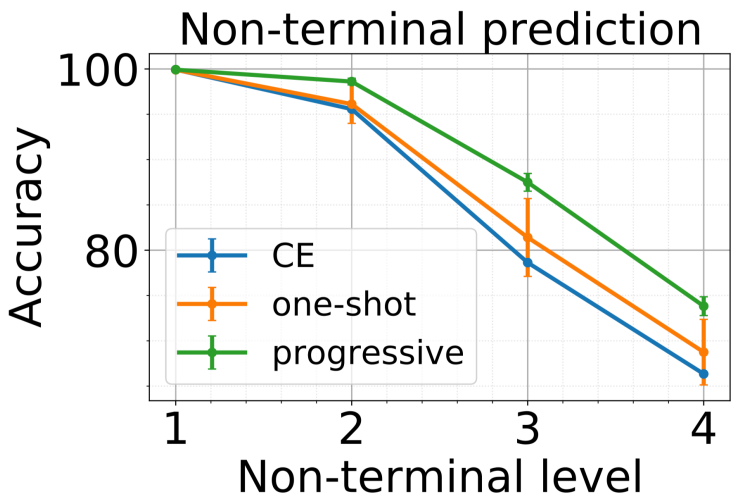
Task: masked prediction → optimal: following the tree hierarchy [Zhao et al. 23].

a quality measure

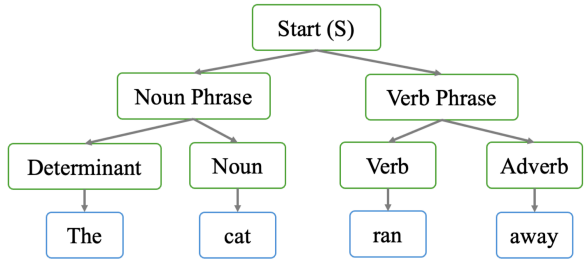
An implicit curriculum exists. *... what is it?*







Implicit curriculum for PCFG



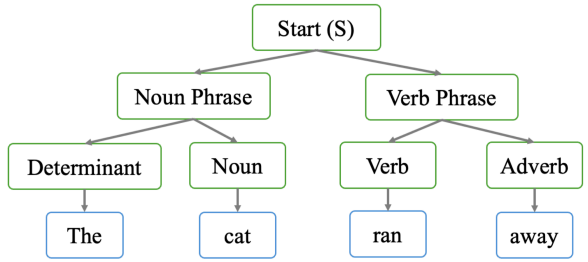
n-grams with an increasing n. (e.g. $n = 3$: cat ran away, cat danced away, cat jumped away, …)

- Smaller n (more local/lower sensitivity) is easier [Abbe et al. 23,24; Vasudeva et al. 24].

2 measures for the dependency on n-grams:

- $M_{\text{robust}} = \text{TV}\left(p(x_{\setminus\{i\}}),\ p(x_{\setminus n\text{-gram}(i)})\right)$ The cat ___ _?_ ____ after hearing…
  - "All but n-gram": smaller → the prediction depends less on n-gram.
- $M_{\text{close}} = \text{TV}\left(p(x_{\setminus\{i\}}),\ p(x_{n\text{-gram}(i)\setminus\{i\}})\right)$ ___ ___ ran _?_ away ____ _____…
  - "Only n-gram": smaller → the prediction is closer to a n-gram model.

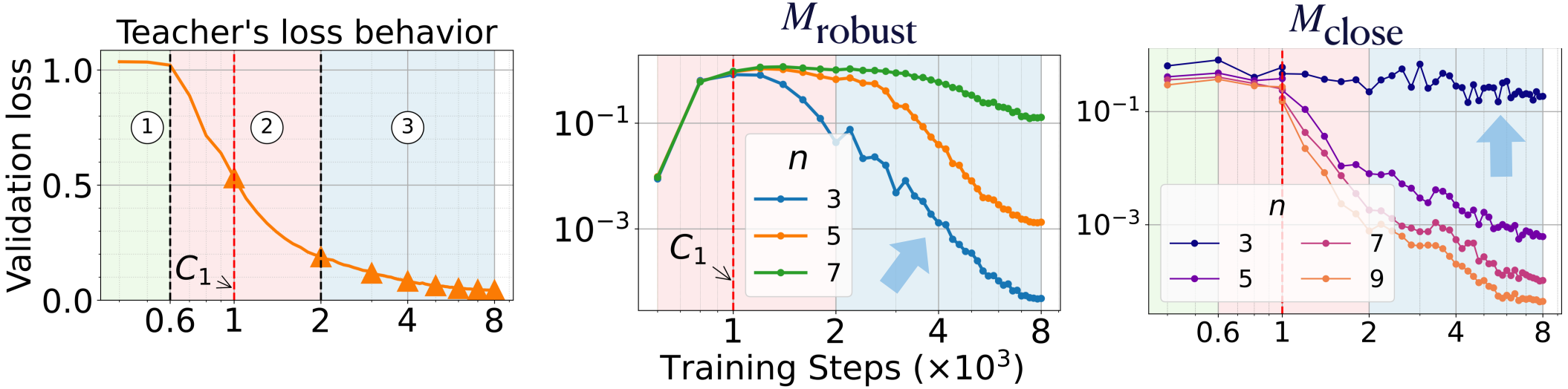
Implicit curriculum for PCFG



n-grams with an increasing n.

- Smaller n (more local/lower sensitivity) is easier [Abbe et al. 23,24; Vasudeva et al. 24].

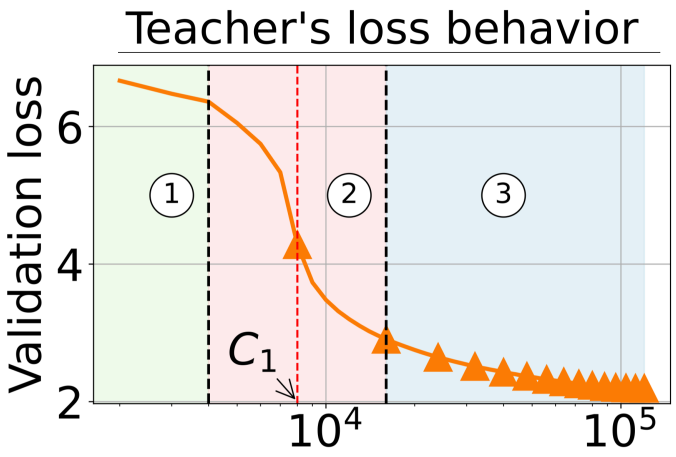
2 measures: later checkpoints depend more on higher n (i.e. harder).

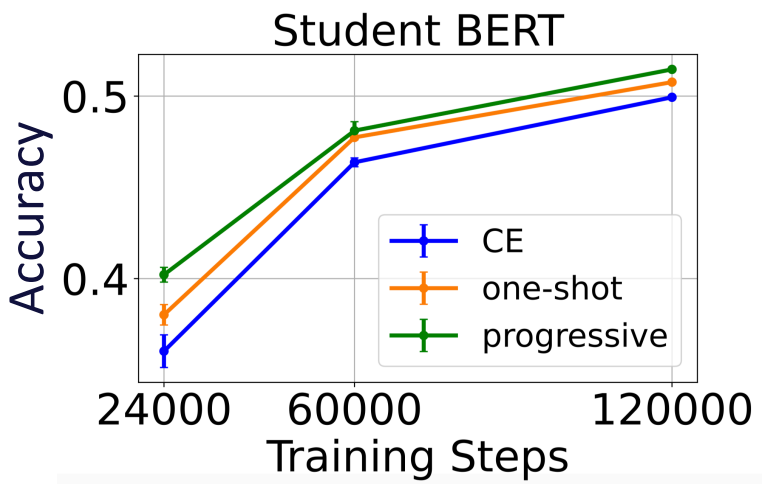


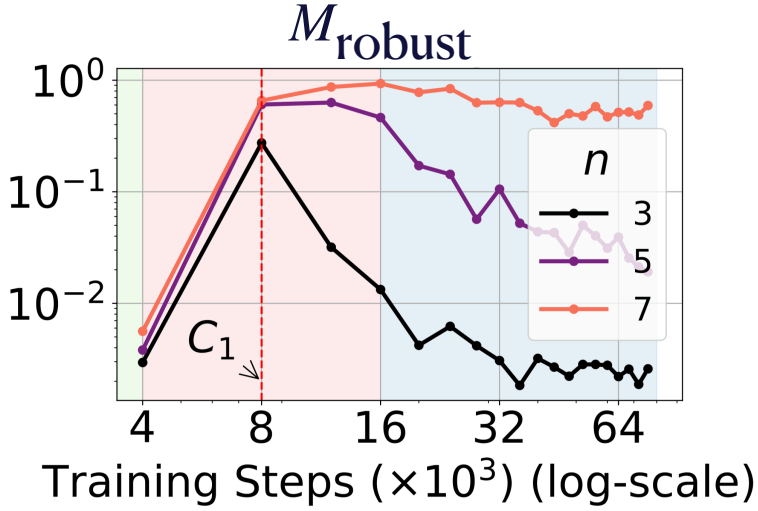
Natural languages

Masked prediction on Wikipedia and Books.

- Similar results for next-token prediction.







Summary: distillation for faster optimization

(Prior work: *generalization* benefits from soft logits)

Progressive distillation induces an implicit curriculum that accelerates optimization.

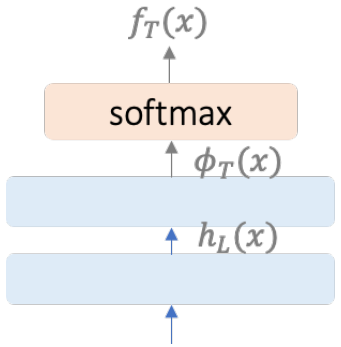
- Motivation: better teacher $\nrightarrow$ better student ("capacity gap").
- Sparse parity: a *low-degree curriculum* $\rightarrow$ improved sample complexity.
  - Analysis: larger Fourier coefficients on $\{i\}, i \in S$.
  - Generalization: hierarchical parity.
- PCFG & natural languages: *n-gram curriculum*.

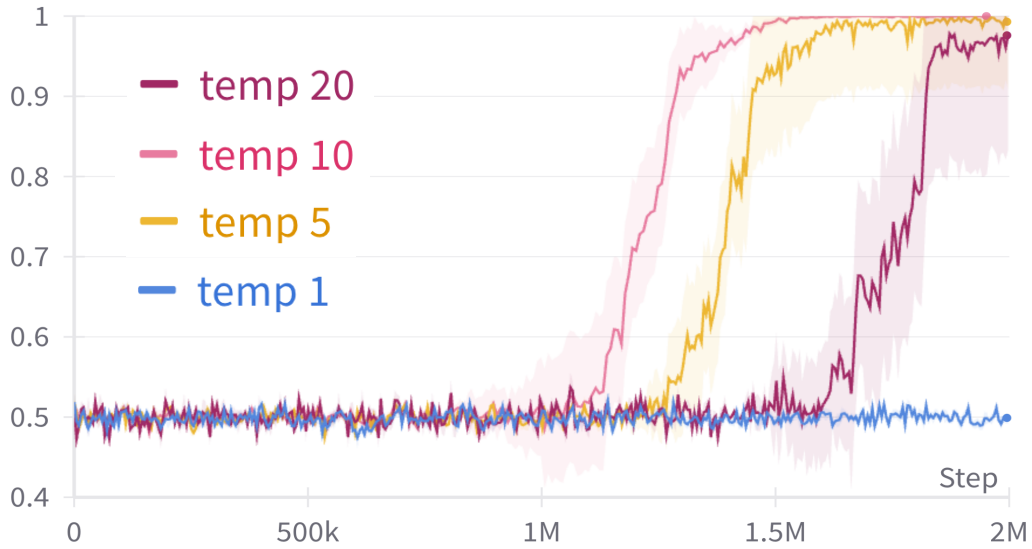
arxiv:2410.05464

Future: distillation for better efficiency

Using 1 checkpoint only? (e.g. 2-shot distillation for parity)

- A follow-up work: layerwise distillation [Gupta & Karmalkar 25].
- Final checkpoint, with a high temperature:





Future: distillation for better efficiency

Using 1 checkpoint only? (e.g. 2-shot distillation for parity)

- A follow-up work: layerwise distillation [Gupta & Karmalkar 25].
- Final checkpoint, with a high temperature.

**Temperature may be more important than methods.*

Setting	No KD	Offline KD		Online KD		Teacher
		$\tau = 1$	$\tau = 4$	$\tau = 1$	$\tau = 4$	
ResNet-56 → LeNet-5x8	47.3 ± 0.6	50.1 ± 0.4	59.9 ± 0.2	61.9 ± 0.2	66.1 ± 0.4	72.0
ResNet-56 → ResNet-20	67.7 ± 0.5	68.2 ± 0.3	71.6 ± 0.2	69.6 ± 0.3	71.4 ± 0.3	72.0
ResNet-110 → LeNet-5x8	47.2 ± 0.5	48.6 ± 0.8	59.0 ± 0.3	60.8 ± 0.2	65.8 ± 0.2	73.4
ResNet-110 → ResNet-20	67.8 ± 0.3	67.8 ± 0.2	71.2 ± 0.0	69.0 ± 0.3	71.4 ± 0.0	73.4

[Harutyunyan et al. 23]

Future: distillation for better efficiency

Transformer to state-space model (SSM)?

- [Bick et al. 24, Wang et al. 24]: distillation with weight initialization.

Model	Tokens / Dataset	WinoG.	Arc-E	Arc-C	PIQA	HellaS.	Lamb.	Avg. ↑
Phi-1.5-1.3B	150B / unknown	73.4	75.6	48.0	76.6	62.6	53.4	64.9
Phi-Mamba-1.5B	3.0B / C4	**71.7**	**74.0**	**44.1**	**75.5**	60.2	50.1	**62.6**
Mamba-1-1.4B	315B / The Pile	61.5	65.5	32.8	74.2	59.1	64.9	59.7
Mamba-2-1.3B	315B / The Pile	60.9	64.3	33.3	73.2	59.9	65.7	59.6

[Bick et al. 24]

Future: distillation for better efficiency

Transformer to state-space model (SSM)?

- [Bick et al. 24, Wang et al. 24]: distillation with weight initialization.

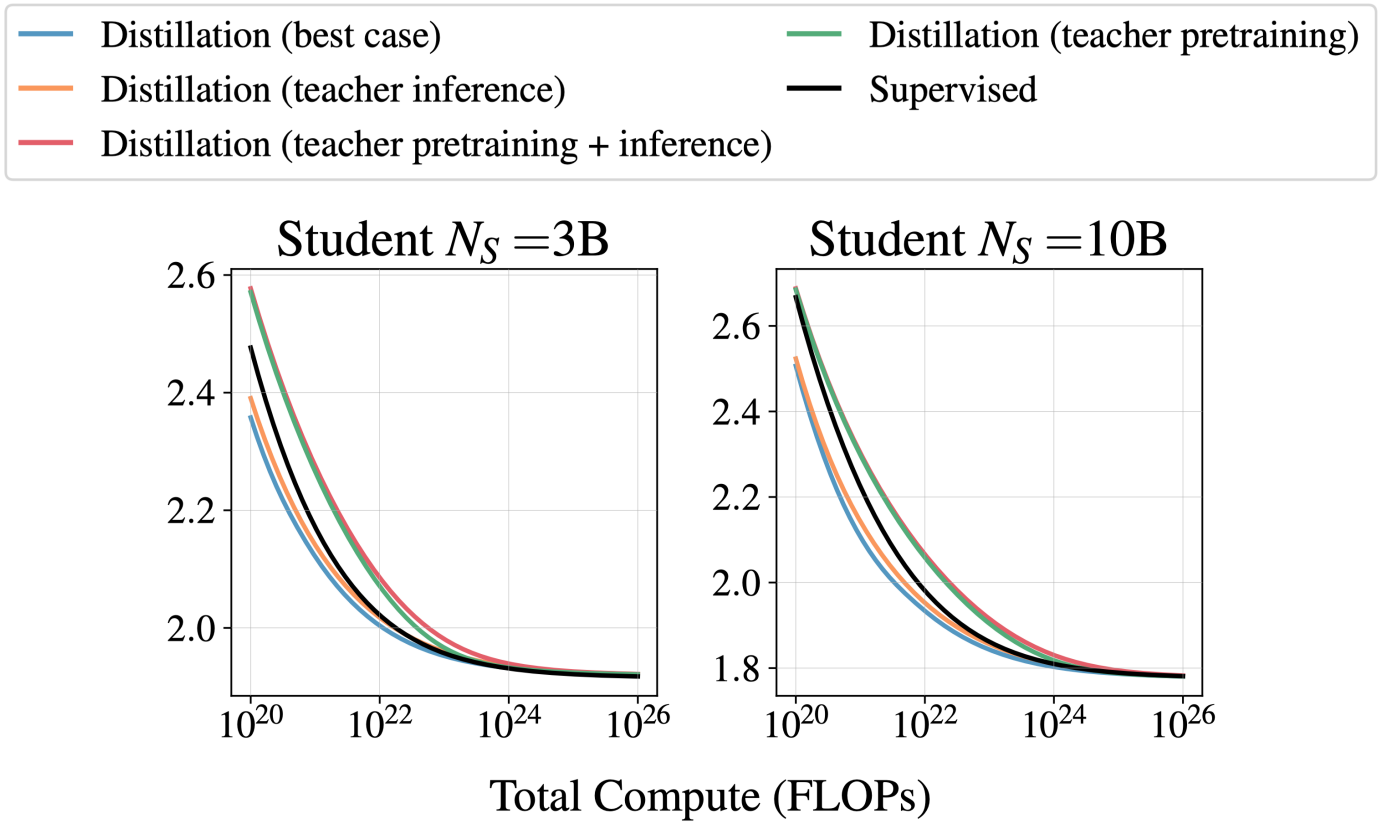
Model (% Att)	Size	Align	MT-Bench (score)	MT-Bench (Round 1)	MT-Bench (Round 2)	AlpacaEval (LC win %)	AlpacaEval (win %)
Llama-3-Instruct	8B	RLHF	**8.00**	-	-	$22.90_{1.26}$	$22.60_{1.26}$
Mamba-Llama3 (50%)	8B	DPO	7.35	7.82	6.88	$\mathbf{29.61}_{1.31}$	$\mathbf{26.69}_{1.31}$
Mamba-Llama3 (25%)	8B	DPO	6.86	7.56	6.15	$25.85_{1.26}$	$22.50_{1.26}$
Mamba-Llama3 (12.5%)	8B	DPO	6.46	6.91	6.01	$20.76_{1.16}$	$17.93_{1.16}$
Mamba2-Llama3 (50%)	8B	DPO	7.32	7.93	6.70	$26.78_{1.26}$	$22.69_{1.26}$
Mamba2-Llama3 (25%)	8B	DPO	6.74	7.24	6.24	$22.75_{1.18}$	$19.01_{1.18}$
Mamba2-Llama3 (12.5%)	8B	DPO	6.48	6.83	6.13	$20.25_{1.13}$	$16.88_{1.13}$
Mamba2-Llama3 (0%)	8B	DPO	5.64	6.16	5.11	$14.49_{0.93}$	$10.88_{0.93}$

[Wang et al. 24]

Future: distillation for better efficiency

A new scaling law?

[Busbridge et al. 25]



Summary: distillation for faster optimization

(Prior work: *generalization* benefits from soft logits)

Progressive distillation induces an implicit curriculum that accelerates optimization.

- Motivation: better teacher $\not\Rightarrow$ better student ("capacity gap").
- Sparse parity: a low-degree curriculum $\rightarrow$ improved sample complexity.
  - Analysis: larger Fourier coefficients on $\{i\}, i \in S$.
  - Generalization: hierarchical parity.
- PCFG & natural languages: n-gram curriculum.

arxiv:2410.05464