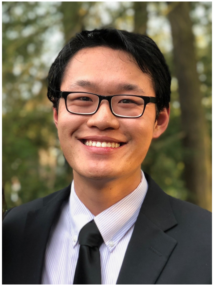
Learning to Answer from Correct Demonstrations

Cong Ma
Department of Statistics, UChicago

BDAI Conference, University of Chicago, Oct. 2025

Nirmit Joshi



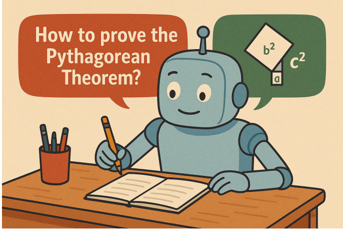
Gene Li

Siddharth Bhandari

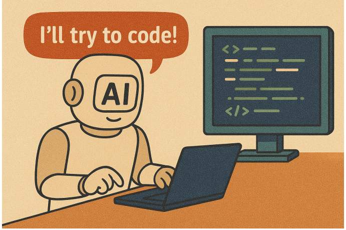
Shiva Kasiviswanathan

Nati Srebro

Answering questions is a big part of our life

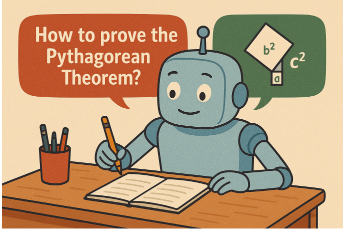


(a) Math Problem Solving

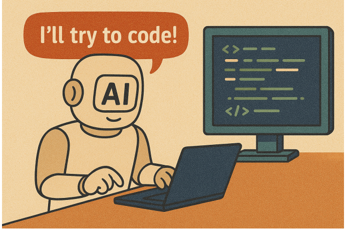


(b) Coding

Answering questions is a big part of our life



(a) Math Problem Solving

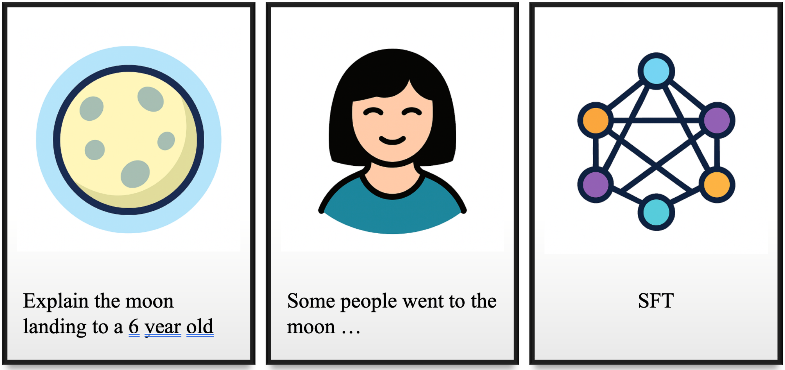


(b) Coding

- **Feature:** many equally good answers
- **Challenge:** *not* to reproduce all correct responses, but to generate *a single good answer*

Learning from correct demonstrations

A timely example: supervised fine-tuning in large language models



A prompt is sampled from the prompt dataset

A labeler demonstrates the desired output

Fine-tune GPT-3 with supervised learning

Formulation via contextual bandits

- Question = context $x \in \mathcal{X}$
- Candidate response = action $y \in \mathcal{Y}$
- Rewards $r_*(x, y) \in \{0, 1\}$ indicating correct or not

	ACTIONS (ANSWERS)			
CONTEXTS (QUESTIONS)	1	0	1	1
	0	1	0	1
	1	0	0	0
	0	1	1	0

Learning goal

Suppose we observe $\{(x_i, y_i)\}_{1 \le i \le m}$ with

$$x_i \sim \mathcal{D}, \quad \text{and} \quad y_i \sim \pi_*(\cdot \mid x_i),$$

where $\pi_*(\cdot \mid x)$ is supported on the set of optimal actions for the context x, given by

$$\sigma_*(x) := \{y \in \mathcal{Y} : r_*(x, y) = 1\}$$

Goal: learn policy $\widehat{\pi}$ with small loss

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}) = \mathbb{E}_{x\sim\mathcal{D}, \widehat{y}\sim\widehat{\pi}(\cdot|x)}\left[\mathbb{1}\{\widehat{y} \notin \sigma_*(x)\}\right]$$

Existing approach based on policy class assumption

Policy class assumption

A common approach to solve this problem is to assume that

$$\pi_* \in \Pi \quad \text{for some small } \Pi \subseteq (\Delta(\mathcal{Y}))^{\mathcal{X}}$$

This motivates maximum likelihood estimator (MLE):

$$\widehat{\pi}_{\mathrm{MLE}} \in \arg\max_{\pi \in \Pi} \prod_{i=1}^{m} \pi(y_i \mid x_i)$$

This is exactly how people solve supervised fine-tuning

Theory and practice of MLE

Proposition 1 (JGBKMS '25, adapted from Foster et al. '24)

Assume $\pi_* \in \Pi$. With high probability, any $\widehat{\pi}_{\mathrm{MLE}}$ obeys

$$L_{\mathcal{D},\sigma_{\pi_*}}(\widehat{\pi}_{\mathrm{MLE}}) \lesssim \frac{\log(|\Pi|)}{m}$$

- **Pro:** minimax optimal for finite Π
- **Con:** small $\log|\Pi|$ is often unrealistic

An alternative: Reward class assumption

Reward class assumption

We assume the underlying reward model class is small, i.e.,

$$\sigma_* \in \mathcal{S} \qquad \text{for some small } \mathcal{S} \subseteq (2^{\mathcal{Y}})^{\mathcal{X}}$$

$$\begin{array}{cc} & \begin{array}{c}\text{ACTIONS}\\ \text{(ANSWERS)}\end{array} \\ \begin{array}{c}\text{CONTEXTS}\\ \text{(QUESTIONS)}\end{array} & \begin{array}{|c|c|c|c|}\hline 1 & 0 & 1 & 1 \\ \hline 0 & 1 & 0 & 1 \\ \hline 1 & 0 & 0 & 0 \\ \hline 0 & 1 & 1 & 0 \\ \hline \end{array} \end{array}$$

Comparisons between two assumptions

Given policy class Π, it is natural to define its associated reward class

$$\mathcal{S}_\Pi := \bigcup_{\pi \in \Pi} \{\sigma_\pi \mid \sigma_\pi(x) = \operatorname{supp} \pi(\cdot \mid x), \forall x \in \mathcal{X}\}$$

Similarly, given reward class $\mathcal{S}$, define its associated policy class

$$\Pi_\mathcal{S} := \bigcup_{\sigma \in \mathcal{S}} \Pi_\sigma, \text{ where } \Pi_\sigma := \{\pi \mid \operatorname{supp} \pi(\cdot \mid x) \subseteq \sigma(x)\,, \forall x \in \mathcal{X}\}\,.$$

Comparisons between two assumptions

Given policy class Π, it is natural to define its associated reward class

$$\mathcal{S}_\Pi := \bigcup_{\pi \in \Pi} \{\sigma_\pi \mid \sigma_\pi(x) = \operatorname{supp} \pi(\cdot \mid x), \forall x \in \mathcal{X}\}$$

Similarly, given reward class $\mathcal{S}$, define its associated policy class

$$\Pi_\mathcal{S} := \bigcup_{\sigma \in \mathcal{S}} \Pi_\sigma, \text{ where } \Pi_\sigma := \{\pi \mid \operatorname{supp} \pi(\cdot \mid x) \subseteq \sigma(x)\,, \forall x \in \mathcal{X}\}\,.$$

Our assumption is weaker: $|\mathcal{S}_\Pi| \leq |\Pi|$ while $|\Pi_\mathcal{S}| \gg |\mathcal{S}|$

Can we learn when $\mathcal{S}$ is small?

Failure of MLE over $\Pi_{\mathcal{S}}$

Recall the associated policy class

$$\Pi_{\mathcal{S}} = \bigcup_{\sigma \in \mathcal{S}} \Pi_\sigma, \text{ where } \Pi_\sigma := \{\pi \mid \operatorname{supp} \pi(\cdot \mid x) \subseteq \sigma(x), \forall x \in \mathcal{X}\}.$$

It is natural to run MLE over $\Pi_{\mathcal{S}}$:

$$\widehat{\pi}_{\mathrm{MLE}} \in \arg\max_{\pi \in \Pi_{\mathcal{S}}} \prod_{i=1}^{m} \pi(y_i \mid x_i)$$

Failure of MLE over $\Pi_{\mathcal{S}}$

Recall the associated policy class

$$\Pi_{\mathcal{S}} = \bigcup_{\sigma \in \mathcal{S}} \Pi_\sigma, \text{ where } \Pi_\sigma := \{\pi \mid \operatorname{supp} \pi(\cdot \mid x) \subseteq \sigma(x)\,,\, \forall x \in \mathcal{X}\}\,.$$

It is natural to run MLE over $\Pi_{\mathcal{S}}$:

$$\widehat{\pi}_{\mathrm{MLE}} \in \arg\max_{\pi \in \Pi_{\mathcal{S}}} \prod_{i=1}^{m} \pi(y_i \mid x_i)$$

This fails: it overfits training data and does not generalize to unseen

Failure instance: $\sigma_*(x) = \sigma_0(x) = \{0\}$, $\sigma_{01}(x) = \{0,1\}$ with large missing mass

Failure of MLE over $\Pi_{\textsf{unif},\mathcal{S}}$

We may consider a restricted policy class $\Pi_{\textsf{unif},\mathcal{S}}$ with size $|\mathcal{S}|$:

$$\Pi_{\textsf{unif},\mathcal{S}} := \{\pi_{\textsf{unif},\sigma} : \sigma \in \mathcal{S}\} \quad \text{where} \quad \pi_{\textsf{unif},\sigma}(\cdot \mid x) = \mathrm{Unif}(\sigma(x))$$

and run MLE

$$\widehat{\pi}_{\mathrm{MLE}} \in \arg\max_{\pi \in \Pi_{\textsf{unif},\mathcal{S}}} \prod_{i=1}^{m} \pi(y_i \mid x_i)$$

Failure of MLE over $\Pi_{\mathsf{unif},\mathcal{S}}$

We may consider a restricted policy class $\Pi_{\mathsf{unif},\mathcal{S}}$ with size $|\mathcal{S}|$:

$$\Pi_{\mathsf{unif},\mathcal{S}} := \{\pi_{\mathsf{unif},\sigma} : \sigma \in \mathcal{S}\} \text{ where } \pi_{\mathsf{unif},\sigma}(\cdot \mid x) = \mathrm{Unif}(\sigma(x))$$

and run MLE

$$\widehat{\pi}_{\mathrm{MLE}} \in \arg\max_{\pi \in \Pi_{\mathsf{unif},\mathcal{S}}} \prod_{i=1}^{m} \pi(y_i \mid x_i)$$

This fails: $\Pi_{\mathsf{unif},\mathcal{S}}$ is misspecified in that π_* may not be in $\Pi_{\mathsf{unif},\mathcal{S}}$

Failure instance: $\sigma_1(x) = \{y^\star, a_1, \ldots, a_{s-1}\}, \sigma_*(x) = \sigma_2(x) = \{y^\star, b_1, \ldots, b_s\}$ and you only observe $y^\star$

Our learner

Online learning from correct demonstrations

Adversary chooses $\sigma_* \in \mathcal{S}$. In each round t:

- Adversary chooses $x_t \in \mathcal{X}$
- Learner predicts $\widehat{y}_t \in \mathcal{Y}$
- Adversary shows some $y_t \in \sigma_*(x_t)$

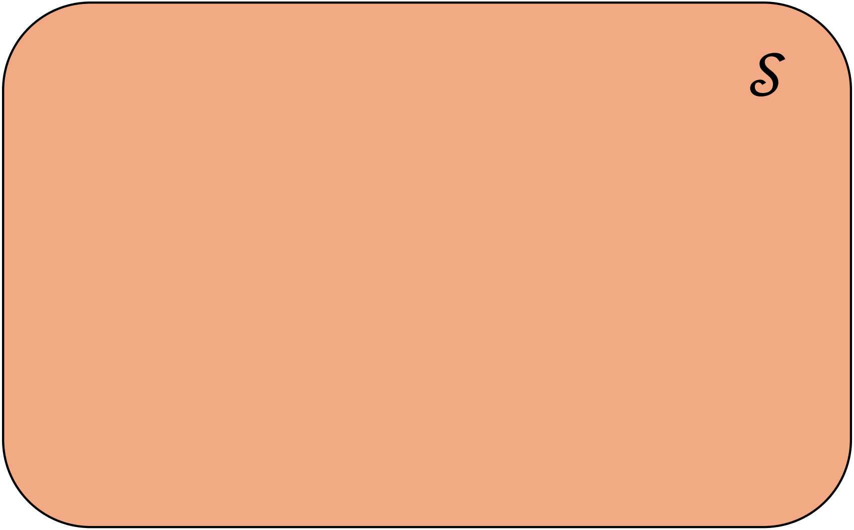
Online learning from correct demonstrations

Adversary chooses $\sigma_* \in \mathcal{S}$. In each round t:

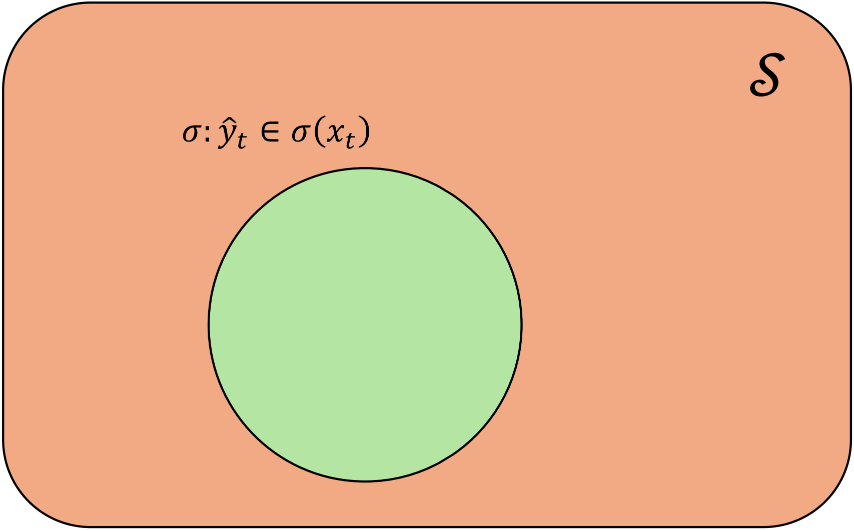
- Adversary chooses $x_t \in \mathcal{X}$
- Learner predicts $\widehat{y}_t \in \mathcal{Y}$
- Adversary shows some $y_t \in \sigma_*(x_t)$

Challenge: learner does not know $\widehat{y}_t$ was a mistake or not

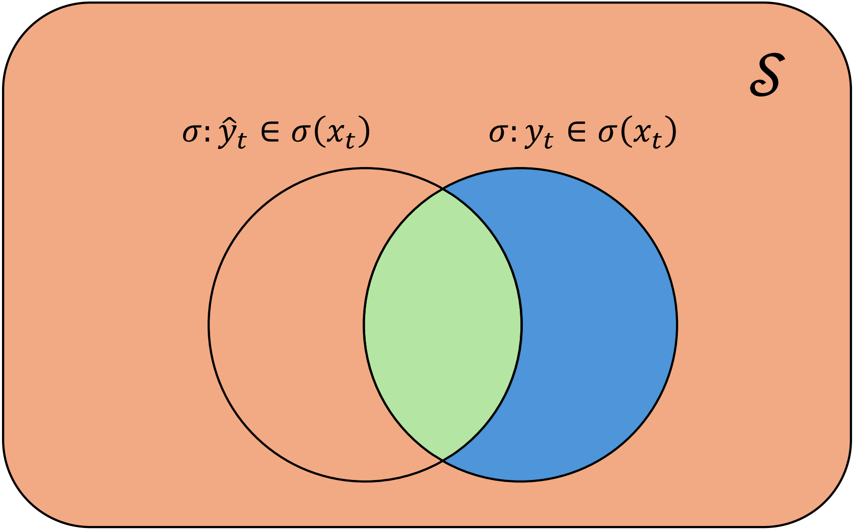
Online weight update



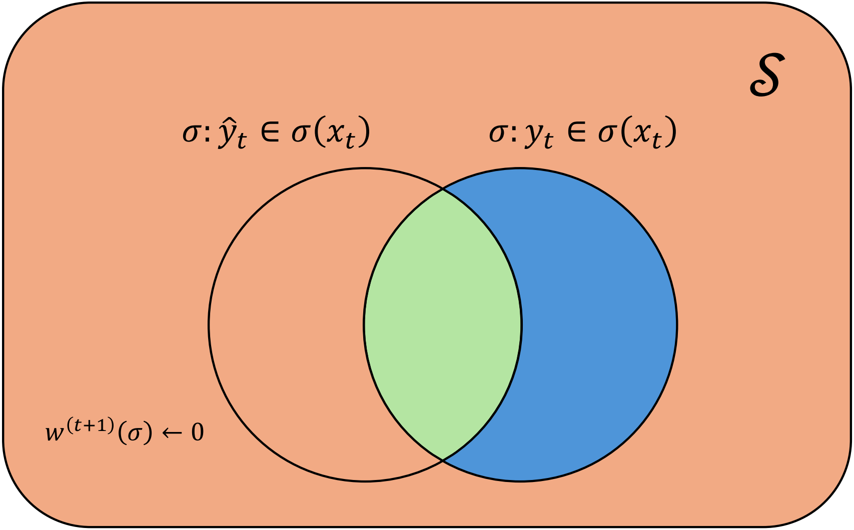
Online weight update



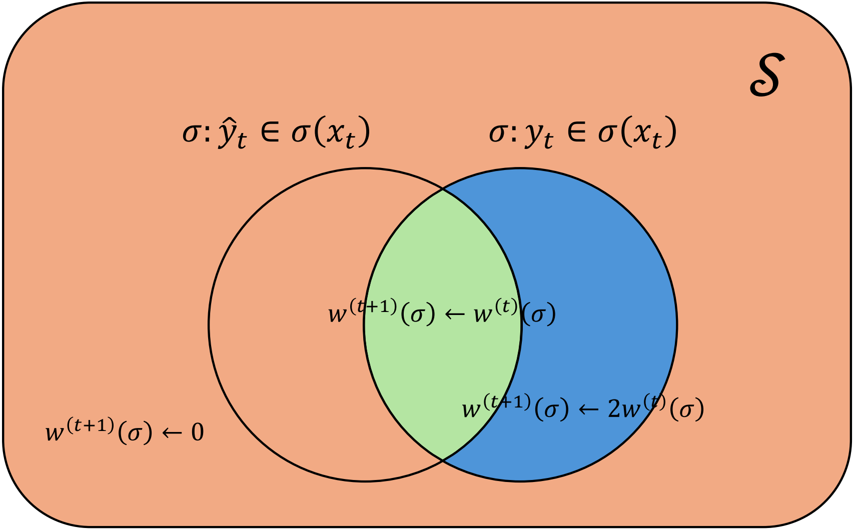
Online weight update



Online weight update



Online weight update



Online mistake bounds

Theorem 1 (JGBKMS '25)

Our learner makes at most $\log_2 |\mathcal{S}|$ *mistakes.*

Online mistake bounds

Theorem 1 (JGBKMS '25)

Our learner makes at most $\log_2 |\mathcal{S}|$ *mistakes.*

Key proof idea:

- overall weight is decreasing
- mistake inflates $w(\sigma_*)$ by 2

Statistical guarantees

Theorem 2 (JGBKMS '25)

With high probability, online-batch-conversion estimator $\widehat{\pi}$ obeys

$$L_{\mathcal{D},\sigma_*}(\widehat{\pi}) \lesssim \frac{\log |\mathcal{S}|}{m}$$

Features:

- No dependence on $|\mathcal{X}|, |\mathcal{Y}|$, or $\sup_x |\sigma_*(x)|$
- Logarithmic dependence on $|\mathcal{S}|$, minimax optimal

Learning from suboptimal demonstrator

So far, we have assumed that π_* is optimal, i.e., $L_{\mathcal{D},\sigma_*}(\pi_*) = 0$

What if π_* is suboptimal?

Learning from suboptimal demonstrator

So far, we have assumed that π_* is optimal, i.e., $L_{\mathcal{D},\sigma_*}(\pi_*) = 0$

What if π_* is suboptimal?

Theorem 3 (JGBKMS '25)

A modification of our estimator $\widehat{\pi}$ obeys: for all $\sigma \in \mathcal{S}$

$$L_{\mathcal{D},\sigma}(\widehat{\pi}) \leq 5\, L_{\mathcal{D},\sigma}(\pi_*) + O\left(\frac{\log_2 |\mathcal{S}|}{m}\right)$$

- **Takeaway:** we can compete with arbitrary demonstrator

A notable extension

pass$@k$ error minimization

We check if the correct answer appears in the top-k guesses:

$$L_{\mathcal{D},\sigma_*}(\widehat{\mu}) = \mathbb{E}_{x\sim\mathcal{D}}, \mathbb{E}_{\boldsymbol{y}=(y^{(1)},\ldots,y^{(k)})\sim\widehat{\mu}(\cdot|x)}\left[\mathbb{1}\{y^{(i)} \notin \sigma_*(x); \forall i \in [k]\}\right].$$

pass@k error minimization

We check if the correct answer appears in the top-k guesses:

$$L_{\mathcal{D},\sigma_*}(\widehat{\mu}) = \mathbb{E}_{x\sim\mathcal{D}}, \mathbb{E}_{\boldsymbol{y}=(y^{(1)},\ldots,y^{(k)})\sim\widehat{\mu}(\cdot|x)} \left[\mathbb{1}\{y^{(i)} \notin \sigma_*(x); \forall i \in [k]\}\right].$$

Theorem 4 (JGBKMS '25)

Variant of our algorithm achieves $\frac{\log_{k+1}(|\mathcal{S}|)}{m}$ *error.*

- **Takeaway:** pass@k gives you $\log_{k+1}$ gain

Conclusions

Summary:

- Learning to answer from correct demonstrations
- An alternative assumption: low-complexity reward model class
- Optimal learner
- Extend to pass$@k$ and suboptimal demonstrators

Moving forward:

- Infinite $\mathcal{S}$?
- Computationally efficient methods?

N. Joshi, G. Li, S. Bhandari, S. Kasiviswanathan, C. Ma, N. Srebro, "Learning to Answer from Correct Demonstrations," forthcoming, 2025