Appendix A: Proofs

Belief Estimation

In this section, we will give the formal definition of model identifiably, and proofs of Proposition 1 & Theorem 1.

Definition 3 (Identifiability) *Consider a probability space $(\mathscr{X}, \mathscr{A}, \mathscr{P})$, where $\mathscr{X}$ is the sample space, $\mathscr{A}$ is the σ-algebra defined on it and $\mathscr{P} = \{P_\theta : \theta \in \Theta\}$ is a family of probability measure, where Θ is the parametric space. Two points of Θ, θ_1 and θ_2 are said to be observationally equivalent (written as $\theta_1 \sim \theta_2$) if*

$$P_{\theta_1}(A) = P_{\theta_2}(A), \forall A \in \mathscr{A}$$

The equivalence relation $\sim$ partitions Θ in to equivalent classes, e.g. $[\theta^0] = \{\theta \in \Theta : \theta \sim \theta^0\}$.

1) *The point θ^0 is said (globally) identifiable if $[\theta^0] = \{\theta^0\}$.*
2) *The model P is said identifiable if the quotient set $\Theta/\sim$ is the finest possible partition.*
3) *A function $\varphi(\theta)$ is identifiable if $\forall \theta_1, \theta_2 \in \Theta, \theta_1 \in [\theta_2] \Rightarrow \varphi(\theta_1) = \varphi(\theta_2)$ (Paulino and de Bragança 1994).*

Proposition 1 *In the pedagogical game, the student's belief can be unidentifiable, whereas the student performance is always identifiable.*

Proof. The probability model of the observation can be written in the matrix form $\boldsymbol{P}_\theta = \boldsymbol{B} \cdot \boldsymbol{\rho}_\theta$, where $\boldsymbol{B}$ is a $(q \times p)$-matrix such that $q = 2|\mathcal{Z}|$, $p = |\mathcal{H}|$, called *emission matrix*. $B_{(\tilde{x},\tilde{y}),h} = P_{\mathcal{D}}(x) \cdot \mathbb{1}\{h(\tilde{x}) = \tilde{y}\}$ is irrelevant to $\theta \in \mathbb{R}^{|H|}$.

The model is unidentifiable since there are many θ's derive the same $\boldsymbol{\rho}_\theta$. If the linear system $\boldsymbol{B} \cdot \boldsymbol{\rho}_\theta = \boldsymbol{P}_\theta$ has multiple solution for $\boldsymbol{\rho}_\theta$, then $\boldsymbol{\rho}_\theta$ as a function of θ is also unidentifiable. In practice, this emission matrix $\boldsymbol{B}$ is very likely to be column rank-deficient. If the feature set $\mathcal{Z}$ can be shattered by the hypothesis class $\mathcal{H}$, then p is at least $2^{|\mathcal{Z}|}$, which is far larger than the number of rows, then B is obviously column rank-deficient. In fact, as long as $p > \frac{1}{2}q + 1$, or equivalently $|\mathcal{H}| > |\mathcal{Z}|+1$, the emission matrix is column rank-deficient.

To prove this, we show that $\texttt{rank}(\boldsymbol{B}) \leq \frac{1}{2}q + 1$. Consider $|\mathcal{Z}|$ hypotheses $h_0, \ldots, h_{|\mathcal{Z}|}$ such that $h_0(x) = 0$ for all x in $\mathcal{Z}$, as well as $h_i(x_i) = 1$ and $h_i(x_{-i}) = 0$ for $i \in [|\mathcal{Z}|]$, where x_{-i} denotes all the elements in $\mathcal{Z}$ other than x_i, for any column $\boldsymbol{B}_{\cdot,h'}$ in the emission matrix, it can be represented as a linear combination of $\boldsymbol{B}_{\cdot,h_0}, \ldots, \boldsymbol{B}_{\cdot,h_{|\mathcal{Z}|}}$. Suppose the hypothesis h' assigns $x_{i_1}, \ldots, x_{i_w}$ positive,

$$\boldsymbol{B}_{\cdot,h'} = \sum_{j=1}^{w} \boldsymbol{B}_{\cdot,h_{i_j}} - (w-1) \cdot \boldsymbol{B}_{\cdot,h_0},$$

therefore the rank of $\boldsymbol{B}$ is no greater than $|\mathcal{Z}| + 1$. When $|\mathcal{H}| > |\mathcal{Z}| + 1$, the model must be unidentifiable, and the student's belief can be unidentifiable in this case.

On the other hand, we observe the student performance as a function of θ, which can rewrite as

$$\begin{aligned}
\eta(\rho_\theta) &= \frac{1}{|\mathcal{G}|} \sum_{(x,y)\in\mathcal{G}} \mathbb{E}_{h\sim\rho_t}\left[\mathbb{1}\{h(x) = y\}\right] \\
&= \frac{1}{|\mathcal{G}|} \sum_{(x,y)\in\mathcal{G}} \sum\nolimits_{h\in\mathcal{H}} \rho_\theta(h) \cdot \mathbb{1}\{h(x) = y\} \\
&= \frac{1}{|\mathcal{G}|} \sum_{(x,y)\in\mathcal{G}} \frac{1}{P_{\mathcal{D}}(x)} \cdot P_\theta(x,y).
\end{aligned}$$

If $\theta_1 \sim \theta_2$, then $P_{\theta_1}(x,y) = P_{\theta_2}(x,y)$ for all $(x,y) \in \mathcal{G} \subseteq \mathcal{Z} \times \{0,1\} \Rightarrow \eta(\rho_{\theta_1}) = \eta(\rho_{\theta_2})$ always holds. Thus the student performance is aways identifiable. □

Theorem 1 *If $\hat{\rho}_\theta$ is a maximum likelihood estimate (MLE) of the student's belief, the induced estimator of student performance, $\eta(\hat{\rho}_\theta)$, is unbiased.*

Proof. The observation is k i.i.d random variables $(\tilde{X}_1, \tilde{Y}_1)$, $\ldots, (\tilde{X}_k, \tilde{Y}_k) \sim P_\theta$. Maximum likelihood estimates of θ is derived by maximizing the likelihood $\mathcal{L}(\theta; \tilde{X}, \tilde{Y})$. Using random variable indicators $\Lambda_i^{x,y} = \mathbb{1}\{\tilde{X}_i = x, \tilde{Y}_i = y\}$, we can write $\mathcal{L}(\theta; \tilde{X}, \tilde{Y}) = \prod_{i=1}^k \prod_{(x,y)\in\tilde{\mathcal{G}}} P_\theta(x,y)^{\Lambda_i^{x,y}}$. Note that $\sum_{(x,y)\in\tilde{\mathcal{G}}} P_\theta(x,y) = 1$, the concave function $\mathcal{L}(\theta; \tilde{X}, \tilde{Y}) = \prod_{(x,y)\in\tilde{G}} P_\theta(x,y)^{\sum_{i=1}^k \Lambda_i^{x,y}}$ of $P_\theta(x,y)$ has a global maximum $\mathcal{L}^*$. Using the Lagrange function

$$\mathcal{F}(\theta) = \log \mathcal{L}(\theta; \tilde{X}, \tilde{Y}) - \lambda \left(\sum_{(x,y)\in\tilde{G}} P_\theta(x,y) - 1 \right)$$

If $\hat{\theta}$ is a maximum estimate, then $\mathcal{F}(\hat{\theta})$ achieves its maximum $\log L^*$. Assume $P_\theta(x,y) > 0$ for all $(x,y) \in \tilde{\mathcal{G}}$ (otherwise $(x,y) \notin \tilde{\mathcal{G}}$ almost surely), we have

$$\frac{\partial \mathcal{F}(\theta)}{\partial P_\theta(x,y)}\Big|_{\theta=\hat{\theta}} = \left(\frac{\sum_{i=1}^k \Lambda_i^{x,y}}{P_\theta(x,y)} - \lambda \right) \Big|_{\theta=\hat{\theta}} = 0,$$

for all $(x,y) \in \tilde{\mathcal{G}}$. Therefore, when $P_{\hat{\theta}}(x,y) = \frac{\sum_{i=1}^k \Lambda_i^{x,y}}{\lambda}$, where $\lambda = \sum_{(x,y)\in\tilde{\mathcal{G}}} \sum_{i=1}^k \Lambda_i^{x,y} = k$, the likelihood function achieves its maximum. Thus, the $\hat{\theta}$ is a maximum likelihood estimator if and only if $\hat{\theta}$ is a solution of the normal equation $\boldsymbol{B}^T \boldsymbol{B} \boldsymbol{\rho}_\theta = \boldsymbol{B}^T \boldsymbol{P}_{\hat{\theta}}$. When the model is unidentifiable, there would be multiple MLEs.

But these MLEs $\hat{\theta}$ all have the same induced estimator of student performance $\hat{\eta}$ since it is identifiable as we argued in Proposition 1. Recall the definition 2 , the $\hat{\eta}$ is

$$\hat{\eta} = \eta(\rho_{\hat{\theta}}) = \frac{1}{|\mathcal{G}|} \sum_{(x,y)\in\mathcal{G}} \frac{1}{P_{\mathcal{D}}(x)} \cdot \frac{\sum_{i=1}^k \Lambda_i^{x,y}}{k}$$

And its expectation over all possible observations is

$$\begin{aligned}
\mathbb{E}_{P_\theta}[\hat{\eta}] &= \frac{1}{|\mathcal{G}|} \sum_{(x,y)\in\mathcal{G}} \frac{1}{P_{\mathcal{D}}(x)} \cdot \frac{\sum_{i=1}^k \mathbb{E}_{P_\theta}[\Lambda_i^{x,y}]}{k} \\
&= \frac{1}{|\mathcal{G}|} \sum_{(x,y)\in\mathcal{G}} \frac{1}{P_{\mathcal{D}}(x)} \cdot P_\theta(x,y) \\
&= \eta(\rho_\theta)
\end{aligned}$$

Therefore, the induced $\hat{\eta}$ is an unbiased estimator. □

Optimal Teaching

In this section, we will first show two lemmas. In Lemma 1 we derive upper and lower bounds of real performance by using a surrogate student performance. Lemma 2 shows the surrogate loss is an monotonic submodular function. Using these two lemmas, we show the effectiveness and efficiency of greedy improving the surrogate performance in Theorems 2 & 3, even when the student belief is unidentifiable.

Lemma 1 *Let $\mathbf{\Psi}' = Z(\hat{\boldsymbol{\rho}}_\theta)\cdot\mathbf{\Psi}$, and $\tilde{\eta}' = 1-(\mathbf{1}-\boldsymbol{u})^T\mathbf{\Psi}'\hat{\boldsymbol{\rho}}_\theta$, we have*

$$\tilde{\eta}' \geq \tilde{\eta} \geq \frac{1}{\beta}\cdot\tilde{\eta}' - \frac{1-\beta}{\beta},$$

where $\beta \in (0, Z(\hat{\boldsymbol{\rho}}_\theta)]$ controls the scaling ratio.

Proof. The prior belief $\hat{\boldsymbol{\rho}}_\theta$ and the poster belief $\mathbf{\Psi}\hat{\boldsymbol{\rho}}_\theta$ are two sum to one vectors, i.e. $\mathbf{1}^T\hat{\boldsymbol{\rho}}_\theta = \mathbf{1}^T\mathbf{\Psi}\hat{\boldsymbol{\rho}}_\theta = 1$. We have

$$\begin{aligned}\tilde{\eta} &= \boldsymbol{u}^T\mathbf{\Psi}\hat{\boldsymbol{\rho}}_\theta = 1-(1-\boldsymbol{u}^T\mathbf{\Psi}\hat{\boldsymbol{\rho}}_\theta)\\ &= 1-(\mathbf{1}^T\mathbf{\Psi}\hat{\boldsymbol{\rho}}_\theta - \boldsymbol{u}^T\mathbf{\Psi}\hat{\boldsymbol{\rho}}_\theta)\\ &= 1-(\mathbf{1}-\boldsymbol{u})^T\Psi\hat{\boldsymbol{\rho}}_\theta\end{aligned}$$

Note that $Z(\hat{\boldsymbol{\rho}}_\theta) \leq 1$ and $\boldsymbol{u} \leq \mathbf{1}$, we have

$$\tilde{\eta}' - \tilde{\eta} = (1-Z(\hat{\boldsymbol{\rho}}_\theta))\cdot(\mathbf{1}-\boldsymbol{u})^T\mathbf{\Psi}\hat{\boldsymbol{\rho}}_\theta \geq 0$$

Since $\beta \leq 1$ and $\beta/Z(\hat{\boldsymbol{\rho}}_\theta) \leq 1$, let $\mathbf{\Psi}'' = \frac{Z(\hat{\boldsymbol{\rho}}_\theta)}{\beta}\mathbf{\Psi}$ and $\tilde{\eta}'' = 1-(\mathbf{1}-\boldsymbol{u})^T\Psi''\hat{\boldsymbol{\rho}}_\theta$. We have

$$\begin{aligned}\tilde{\eta}'' - \tilde{\eta} &= \left(\frac{1}{\beta}\cdot\tilde{\eta}' - \frac{1-\beta}{\beta}\right) - \tilde{\eta}\\ &= (1-Z(\hat{\boldsymbol{\rho}}_\theta)/\beta)\cdot(\mathbf{1}-\boldsymbol{u})^T\mathbf{\Psi}\hat{\boldsymbol{\rho}}_\theta\\ &\leq 0\end{aligned}$$

Therefore, $\tilde{\eta}' \geq \tilde{\eta} \geq \tilde{\eta}'/\beta - (1-\beta)/\beta$ holds. Especially, we have $\tilde{\eta} = \tilde{\eta}'/\beta - (1-\beta)/\beta$ when $\beta = Z(\tilde{\boldsymbol{\rho}}_\theta)$. □

Lemma 2 *The surrogate performance $\tilde{\eta}'$ is a monotonic submodular function of the teaching examples o_t.*

Proof. We can write the surrogate performance $\tilde{\eta}'$ as

$$\begin{aligned}\tilde{\eta}' &= (\tilde{\eta}' - \hat{\eta}) + \hat{\eta}\\ &= (\mathbf{1}-\boldsymbol{u})^T(\boldsymbol{I}-\mathbf{\Psi}')\hat{\boldsymbol{\rho}}_\theta + \hat{\eta}\\ &= \sum_{h\in\mathcal{H}}(1-u(h))\cdot\hat{\rho}_\theta(h)\cdot P(o_t|h) + \hat{\eta},\end{aligned}$$

where $P(o_t|h) = 1 - \prod_{(x,y)\in o_t}\sigma_\alpha(h(x),y)$. Since the noise-tolerant likelihood $\sigma_\alpha(\cdot,\cdot) \in (0,1)$, for any $o_a \subseteq o_b \subseteq \mathcal{G}$, thus we have

1. (Monotonicity):

$$\begin{aligned}&P(o_b|h) = 1 - \prod_{(x,y)\in o_b}\sigma_\alpha(h(x),y)\\ =\ & 1 - \prod_{(x,y)\in o_a}\sigma_\alpha(h(x),y)\cdot\prod_{(x,y)\in o_b\setminus o_a}\sigma_\alpha(h(x),y)\\ \geq\ & 1 - \prod_{(x,y)\in o_a}\sigma_\alpha(h(x),y) = P(o_a|h)\end{aligned}$$

2. (Submodularity):

$$\begin{aligned}&P(o_b\cup(x,y)|h) - P(o_b|h)\\ =\ & (1-\sigma_\alpha(h(x),y))\prod_{(x,y)\in o_b}\sigma_\alpha(h(x),y)\\ \leq\ & (1-\sigma_\alpha(h(x),y))\prod_{(x,y)\in o_a}\sigma_\alpha(h(x),y)\\ =\ & P(o_a\cup(x,y)|h) - P(o_a|h)\end{aligned}$$

It indicates $P(o_t|h)$ is monotonic and submodular for all h. Besides, the coefficient $(1-u(h))\hat{\rho}_\theta(h) \geq 0$ for all h. We conclude the surrogate student performance $\tilde{\eta}'$ is a monotonic submodular function of o_t. □

Theorem 2 *Suppose the model is identifiable. Greedily giving the student examples to improve the surrogate student performance $\tilde{\eta}'$ until it is no less than $(1-\beta(1-\gamma))(1-\hat{\eta}) + \hat{\eta}$ guarantees $\tilde{\eta} \geq \gamma(1-\hat{\eta}) + \hat{\eta}$, the real performance achieves the teaching target.*

Proof. Theorem 2 is a direct consequence of Lamma 1. □

Corollary 1 *When the model is unidentifiable, i.e., there is a equivalent class $[\hat{\boldsymbol{\rho}}_\theta] \neq \{\hat{\boldsymbol{\rho}}_\theta\}$, we can greedily select examples $a_t^{\mathbf{T}} \subseteq \mathcal{G}\backslash o_t$ to increase the worst improvement*

$$E_{a_t^{\mathbf{T}}} = \min_{\boldsymbol{\rho}_\theta\in[\hat{\boldsymbol{\rho}}_\theta]}(\mathbf{1}-\boldsymbol{u})^T(\boldsymbol{I}-\mathbf{\Psi}'_{a_t^{\mathbf{T}}})\boldsymbol{\rho}_\theta,$$

until $E_{a_t^{\mathbf{T}}}$ is no less than $(1-\beta(1-\gamma))(1-\hat{\eta})$, where $\beta = \min_{\boldsymbol{\rho}_\theta\in[\hat{\boldsymbol{\rho}}_\theta]}\mathbf{\Psi}'^T_{a_t^{\mathbf{T}}}\boldsymbol{\rho}_\theta$, to ensure $\tilde{\eta}$ achieves the teaching target.

Theorem 3 *Given the current student performance $\hat{\eta}$ and the target improvement ratio γ, by greedily providing $\texttt{OPT}(\tilde{\eta}'_\xi)\cdot\log\frac{1}{\xi\beta(1-\gamma)}$ examples it is guaranteed to improve the student performance to the teaching target, where*

$$\tilde{\eta}'_\xi = \hat{\eta} + [1-\beta(1-\xi)(1-\gamma)]\cdot(1-\hat{\eta}),$$

and $\texttt{OPT}(\cdot)$ is the minimal number of examples to increase surrogate performance to a certain value.

Proof. Following lamma 2, we know that $\tilde{\eta}'$ is a nonnegative monotone submodular function. Using the result of greedy maximization of such submodular function (Krause and Golovin 2014) that, for any ℓ and k,

$$f(S_\ell) \geq \left(1-e^{-\ell/k}\right)\max_{S:|S|=k}f(S),$$

where S_ℓ is the set picked after ℓ steps.

Let $k^* = \texttt{OPT}(\tilde{\eta}'_\xi)$, when $\ell \geq k^*\log\frac{1}{\xi\beta(1-\gamma)}$, we have

$$\begin{aligned}\tilde{\eta}'(S_\ell) - \hat{\eta} &\geq (1-\xi\beta(1-\gamma))(\tilde{\eta}'_\xi - \hat{\eta})\\ &\geq (1-\beta(1-\gamma))(1-\hat{\eta}).\end{aligned}$$

This indicates greedily providing $\texttt{OPT}(\tilde{\eta}'_\xi)\cdot\log\frac{1}{\xi\beta(1-\gamma)}$ more examples will absolutely improve the student performance to the teaching target.

□

Value-Alignment

To show the value alignment property, we first investigate a special case where the student earns a constant much bonus credit every round, then complete the proof by showing that the student earns most in that special case.

Lemma 3 *If a student* $\mathbf{S}$ *earns* $\omega \in (0,1)$ *in each round, then his his overall improvement* $\Delta = \eta_{N-1} - \eta_0$ *should be no less than* $\omega\bar{\Delta}$*, where* $\bar{\Delta} = \bar{\eta}_{N-1} - \bar{\eta}_0$ *is the overall improvement of a "model" student who has the same initial performance and exactly achieves each round target.*

Proof. Let $\Delta(\eta) = \gamma(1-\eta)$ be the required improvement for achieving the teaching target of η. Note that it has two properties: if $\eta^{(1)} \leq \eta^{(2)}$, then (a) $\Delta(\eta^{(1)}) \geq \Delta(\eta^{(2)})$, and (b) $\eta^{(1)} + \Delta(\eta^{(1)}) \leq \eta^{(2)} + \Delta(\eta^{(2)})$.

Let $\bar{\eta}_0, \bar{\eta}_1, \ldots, \bar{\eta}_{n-1}$ be each round of performance of the model student, and $\eta_0, \eta_1, \ldots, \eta_{n-1}$ be each round of performance of student $\mathbf{S}$. $\eta_0 = \bar{\eta}_0$ Suppose the teacher estimates his performance accurately. Using the property (b), we know that $\eta_t \leq \bar{\eta}_t$, for all $t \in \{0, \ldots, N-1\}$, since $\eta_1 < \bar{\eta}_1$ and $\eta_{t-1} \leq \bar{\eta}_{t-1} \Rightarrow \eta_t \leq \bar{\eta}_t$. Furthermore, according to property (a) we know each round of $\mathbf{S}$'s improvement $\omega\Delta(\eta_{t-1})$ should be no less than $\omega\Delta(\bar{\eta}_{t-1})$. Therefore,

$$\Delta = \sum_{t=1}^{N-1} \omega\Delta(\eta_{t-1}) \geq \omega \sum_{t=1}^{N-1} \Delta(\bar{\eta}_{t-1}) = \omega\bar{\Delta}.$$

Hence, a student earning constant $\omega \in (0,1)$ each round has overall improvement at least $\omega\bar{\Delta}$, where $\bar{\Delta}$ is the overall improvement of the model student. □

Theorem 4 *If a student* $\mathbf{S}$ *earns* $\omega \in (0,1)$ *on average each round, then his overall improvement* $\Delta = \eta_{N-1} - \eta_0$ *hould be no less than* $\omega\bar{\Delta}$*, where* $\bar{\Delta} = \bar{\eta}_{N-1} - \bar{\eta}_0$ *is the overall improvement of a "model" student who has the same initial performance and exactly achieves each round target.*

Proof. The overall improvement of the model student can be written as

$$\begin{aligned}\bar{\Delta} &= \sum_{t=1}^{N-1} \gamma(1-\gamma)^{t-1}(1-\bar{\eta}_0) \\ &= \left(1-(1-\gamma)^{N-1}\right)(1-\eta_0)\end{aligned}$$

Let $\omega_1, \omega_2, \ldots, \omega_{N-1} \in (-\infty, 1)$ be each round of bonus credits received by the student $\mathbf{S}$, whose average is $\omega = \frac{1}{N-1}\sum_{t=1}^{N-1}\omega_t$. And let $\kappa_t = \frac{\eta_t - \eta_{t-1}}{\bar{\eta}_{t-1} - \eta_{t-1}} \in (-\infty, 1/\gamma), t \in [N-1]$ be the student real improvement ratio, whose average $\kappa \geq \omega$ since by definition $\omega_t = \begin{cases} 1, & \kappa_t > 1 \\ \kappa_t, & \kappa_t \leq 1 \end{cases}$. The student $\mathbf{S}$'s overall improvement is

$$\begin{aligned}\Delta &= \sum_{t=1}^{N-1} \kappa_t\gamma \left(\prod_{i=1}^{t-1}(1-\kappa_i\gamma)\right)(1-\eta_0) \\ &= \left(1 - \prod_{t=1}^{N-1}(1-\kappa_t\gamma)\right)\cdot(1-\eta_0) \\ &\geq \left(1-(1-\kappa\gamma)^{N-1}\right)\cdot(1-\eta_0) \\ &\geq \left(1-(1-\omega\gamma)^{N-1}\right)\cdot(1-\eta_0)\end{aligned}$$

The second last step is because of the AM-GM inequality,

$$\left(\prod_{t=1}^{N-1}(1-\kappa_t\gamma)\right)^{\frac{1}{(N-1)}} \leq \frac{\sum_{t=1}^{N-1}(1-\kappa_t\gamma)}{N-1} = 1-\kappa\gamma$$

where the equality holds when $\kappa_1 = \cdots = \kappa_{N-1} = \omega_1 = \cdots = \omega_{N-1} = \omega$. According to lemma 3, $\Delta \geq \omega\bar{\Delta}$ in this case. Therefore, we conclude that $\mathbf{S}$'s overall improvement is at least $\omega\bar{\Delta}$. □

Appendix B: Algorithms

Algorithm 1 Pedagogical Reasoning - Preparation

Initialize $|\mathcal{H}|$-vetors $\boldsymbol{u}$, $\{\boldsymbol{\psi}_{(x,y)}\}$, and $2|\mathcal{Z}| \times |\mathcal{H}|$-matrix $\boldsymbol{B}$ by filling them with zeros.
procedure PREPARATION($\mathcal{H}$, $\mathcal{G}$, $\mathcal{Z}$)
  for $h \in \mathcal{H}$ **do**
    for $x_i \in \mathcal{Z}$ **do**
      $B_{(x,0),h} \leftarrow \mathbb{1}\{h(x_i) = 0\}$
      $B_{(x,1),h} \leftarrow \mathbb{1}\{h(x_i) = 1\}$
    end for
    for $(x,y) \in \mathcal{G}$ **do**
      $u_h \leftarrow u_h + \frac{1}{|\mathcal{G}|} \cdot \mathbb{1}\{h(x) = y\}$
      $\psi_{(x,y),h} \leftarrow \sigma_\alpha(h(x), y)$
    end for
  end for
  $\boldsymbol{K}_B \leftarrow$ NULLSPACE($\boldsymbol{B}$)
  return $\boldsymbol{K}_B$, $\boldsymbol{u}$, $\{\boldsymbol{\psi}_{(x,y)}\}$
end procedure

Algorithm 2 Pedagogical Reasoning - Belief Estimation

procedure BELIEFESTIMATE(g_t, $\boldsymbol{u}$, $\boldsymbol{K}_B$)
  $\hat{\boldsymbol{\rho}}_{\theta_t} \leftarrow$ MLE(g_t)
  $\hat{\eta}_t \leftarrow \boldsymbol{u}^T\hat{\boldsymbol{\rho}}_{\theta_t}$
  $[\hat{\boldsymbol{\rho}}_{\theta_t}] \leftarrow (\hat{\boldsymbol{\rho}}_{\theta_t}, \boldsymbol{K}_B)$
  return $[\hat{\boldsymbol{\rho}}_{\theta_t}]$, $\hat{\eta}_t$
end procedure

Algorithm 3 Pedagogical Reasoning - Teaching

procedure TEACHING(o_t, $[\hat{\boldsymbol{\rho}}_{\theta_t}]$, $\mathcal{G}$, $\{\boldsymbol{\psi}_{(x,y)}\}$)
  $a_t^{\mathbf{T}} \leftarrow \{\}$
  $\boldsymbol{\psi} \leftarrow \bigotimes_{(x,y)\in o_t} \boldsymbol{\psi}_{(x,y)}$
  $\Delta_\beta \leftarrow (1-(1-\gamma))(1-\hat{\eta}_t)$
  while $\tilde{\eta}'_{t+1} - \hat{\eta}_t < \Delta_\beta$ **do**
    for $(x,y) \in \mathcal{G}\backslash(o_t \cup a_t^{\mathbf{T}})$ **do**
      $E_{(x,y)} \leftarrow \min_{\boldsymbol{\rho}\in[\hat{\boldsymbol{\rho}}_{\theta_t}]} 1 - \boldsymbol{\psi}_{(x,y)}^T(\boldsymbol{\psi}\otimes(1-\boldsymbol{u})\otimes\boldsymbol{\rho})$
    end for
    $a \leftarrow \arg\max_{(x,y)\in\mathcal{G}} E_{(x,y)}$
    $\tilde{\eta}'_{t+1} \leftarrow E_a$
    $\boldsymbol{\psi} \leftarrow \boldsymbol{\psi}_a \otimes \boldsymbol{\psi}$
    $\beta \leftarrow \min_{\boldsymbol{\rho}\in[\hat{\boldsymbol{\rho}}_{\theta_t}]} \boldsymbol{\psi}^T\boldsymbol{\rho}$
    $\Delta_\beta \leftarrow (1-\beta(1-\gamma))(1-\hat{\eta}_t)$
    $a_t^{\mathbf{T}} \leftarrow a_t^{\mathbf{T}} \cup \{a\}$
  end while
  return $a_t^{\mathbf{T}}$
end procedure