Human-assisted Computation: Autograding

Let a student's answer be A_1 with a probability of p_1 for their answer being correct. i.e.

$$p_1 = P(A_1 = 1 \mid A_0 = 1) = P(A_1 = -1 \mid A_0 = -1).$$

From Baye's Theorem, we have

$$P(A_0 = 1 \mid A_1 = 1) = \frac{P(A_1 = 1 \mid A_0 = 1)P(A_0 = 1)}{P(A_1 = 1)}.$$

We can further rewrite it using p_0 and p_1:

$$P(A_0 = 1 \mid A_1 = 1) = \frac{p_1 p_0}{p_1 p_0 + (1 - p_1)(1 - p_0)}.$$

$$P(A_0 = 1 \mid A_1 = a_1) = \text{logit}^{-1}(\text{logit}(p_0) + a_1 \text{logit}(p_1))$$

where the logit function is defined as $\text{logit}(\cdot) = \log(\frac{\cdot}{1-\cdot})$, and $\text{logit}^{-1}(\cdot) = \frac{\exp \cdot}{1+\exp \cdot}$.