Proof for Theorem A :

Let $a \sim \pi_G$ be an action sampled from the generator, p_G be the probability of the generated action having reward 1. Let $R_{\text{gt}}(a) \in \{0,1\}$ be the ground-truth reward (i.e., whether the action is good or bad), and $V(a) \in \{0,1\}$ be the verifier output.

We define:

$$\begin{aligned} p_G &= P\big(R_{\text{gt}}(a) = 1 \;\big|\; a \sim \pi_G\big) \\ p_V &= P\left(V(a) = R_{\text{gt}}(a)\right) \end{aligned}$$

that is, p_G is the probability the generator generates a good action, and p_V is the probability the verifier predicts the correct reward given an action, **independent of the ground truth action reward**.

Theorem A. Given $p_G \in (0,1)$, $N > 1$, then $\mathbb{E}\left[R_{\text{gt}}(a_{\text{w/ver}})\right] > \mathbb{E}\left[R_{\text{gt}}(a_{\text{naive}})\right]$ if and only if $p_V > 0.5$

Proof.

The expected reward of naive sampling from generator: $\mathbb{E}\left[R_{\text{gt}}(a_{\text{naive}})\right] = p_G \times 1 + (1-p_G) \times 0 = p_G$.

The expected reward of the selected action using the verifier is:

$$\begin{aligned} \mathbb{E}[R_{\text{gt}}(a_{\text{w/ver}})] &= P(\exists i,\ V(a_i)=1) \cdot \mathbb{E}[R_{\text{gt}}(a) \mid V(a)=1] \\ &\quad + P(\forall i,\ V(a_i)=0) \cdot \mathbb{E}[R_{\text{gt}}(a) \mid V(a)=0] \\ &= \big(1-(1-Q)^N\big) \cdot \frac{P(R=1,V=1)}{P(V=1)} \\ &\quad + (1-Q)^N \cdot \frac{P(R=1,V=0)}{P(V=0)} \end{aligned}$$

where $Q = P\left(V(a)=1\right) = (1-p_G)(1-p_V) + p_G p_V$. and

$$P(R=1,V=1) = p_G \cdot p_V, \quad P(R=1,V=0) = p_G \cdot (1-p_V).$$

Substituting into the expression, we get:

$$\mathbb{E}[R_{\text{gt}}(a_{\text{w/ver}})] = \big(1-(1-Q)^N\big) \cdot \frac{p_G p_V}{Q} + (1-Q)^N \cdot \frac{p_G(1-p_V)}{1-Q}.$$

We will first prove $\mathbb{E}[R_{\text{gt}}(a_{\text{w/ver}})] > \mathbb{E}[R_{\text{gt}}(a_{\text{naive}})] \Rightarrow p_V > 0.5$, and show that each step is reversible to prove the other direction.

$$\begin{aligned} &\mathbb{E}[R_{\text{gt}}(a_{\text{w/ver}})] > \mathbb{E}[R_{\text{gt}}(a_{\text{naive}})] = p_G. \\ \underset{p_G}{\overset{divide\ by}{\iff}}\ &\big(1-(1-Q)^N\big) \cdot \frac{p_V}{Q} + (1-Q)^N \cdot \frac{1-p_V}{1-Q} > 1. \end{aligned}$$

Rewriting and simplifying:

$$\begin{aligned} &\big(1-(1-Q)^N\big) \cdot \frac{p_V}{Q} + (1-Q)^N \cdot \frac{1-p_V}{1-Q} > 1 \\ \iff &\big(1-(1-Q)^N\big) \cdot \frac{p_V}{Q} + (1-p_V)(1-Q)^{N-1} > 1 \\ \iff &\left(\frac{p_V}{Q} - \frac{p_V}{Q}(1-Q)^N\right) + (1-p_V)(1-Q)^{N-1} > 1 \\ \iff &\frac{p_V}{Q} - \frac{p_V}{Q}(1-Q)^N + (1-p_V)(1-Q)^{N-1} > 1 \\ \iff &\left(\frac{p_V}{Q} - 1\right) + \left[-\frac{p_V}{Q}(1-Q)^N + (1-p_V)(1-Q)^{N-1}\right] > 0 \\ \iff &\left(\frac{p_V}{Q} - 1\right) + (1-Q)^{N-1}\left[-\frac{p_V}{Q}(1-Q) + (1-p_V)\right] > 0 \\ \iff &\left(\frac{p_V}{Q} - 1\right) + (1-Q)^{N-1}\left[1 - p_V - \frac{p_V}{Q}(1-Q)\right] > 0 \\ \iff &\left(\frac{p_V}{Q} - 1\right) + (1-Q)^{N-1}\left(1 - \frac{p_V}{Q}\right) > 0 \\ \iff &\left(\frac{p_V}{Q} - 1\right)\left[1-(1-Q)^{N-1}\right] > 0. \end{aligned}$$

Since $p_G \in (0,1)$, we have $Q \in (0,1)$, then $1-(1-Q)^{N-1} > 0$, so the inequality holds if and only if:

$$\frac{p_V}{Q} > 1 \quad \iff \quad p_V > Q.$$

Substituting the expression for Q:

$$p_V > p_G p_V + (1-p_G)(1-p_V),$$

Rearranging:

$$\begin{aligned} &p_V - p_G p_V > (1-p_G)(1-p_V), \\ \iff &p_V(1-p_G) > (1-p_G)(1-p_V). \end{aligned}$$

Since $p_G \in (0,1)$, we can divide both sides by $1-p_G$, yielding:

$$p_V > 1 - p_V \quad \iff \quad p_V > 0.5.$$

Since all of the above steps are reversible, we prove that the verifier improves the expected reward over naive sampling if and only if $p_V > 0.5$.