Supplementary Material for 'Action-Priority Driven Policy Optimization for Flexible Job Shop Scheduling Problem'

Wenjun Zheng[1], Weilin Cai[2], Wei Qu[3], Jianfeng Mao*
[1,2,3]:{wenjunzheng,weilincai,weiqu}@link.cuhk.edu.cn
*: jfmao@cuhk.edu.cn; corresponding author

I. Model Explanation

A. Detailed Version of System Dynamics

In this subsection, we introduce the system dynamics in details. Since the FJSP problem is deterministic, the next state $s_{t+1} = (\mathbf{J}_{t+1}, \mathbf{M}_{t+1}, \mathbf{W}_{t+1}, \mathbb{T}_t)$ is uniquely determined given the current state $s_t = (\mathbf{J}_t, \mathbf{M}_t, \mathbf{W}_t)$ and action $a_t = (m, j, i)$. The state update follows three steps: Step (1) is an **Intermediate Information Storage**: In this step, intermediate job, machine, and working information are created to store the state after applying action a. Here Step (2) is for **Checking Available Actions**: Since, in our model, a stage represents a decision point when an available machine can be assigned to an operation, we need to check whether any available actions remain after executing a. If no actions are available, we proceed to Step (3), i.e. **Processing Progression**: The system advances the processing stage, during which some operations are completed, and some machines are released. After this update, we again check for available actions, leading back to Step (2) if any exist. The details for each step are shown below:

(1) First, obtain intermediate job information $\mathbf{J}$, machine information $\mathbf{M}$, working information $\mathbf{W}$ and current system time $\mathbb{T}$. The newly assigned machine M_m for operation $O_{j,i}$ and its remaining processing time are added to the work information:

$$\mathbf{W} = \mathbf{W}_t \cup \{(m, j, i, (\mathcal{P})_{m,j,i})\}$$

The first column of $\mathbf{J}_t$ remains unchanged:

$$(\mathbf{J})_{*,1} = (\mathbf{J}_t)_{*,1}$$

The status of job J_j is updated from idle to in progress:

$$(\mathbf{J})_{j',2} = \begin{cases} 1, & \text{if } j' = j, \\ (\mathbf{J}_t)_{j',2}, & \text{otherwise.} \end{cases}$$

The status of machine M_m is changed from idle to busy:

$$(\mathbf{M})_{m'} = \begin{cases} 1, & \text{if } m' = m, \\ (\mathbf{M}_t)_{m'}, & \text{otherwise.} \end{cases}$$

The system time remains the same: $\mathbb{T} = \mathbb{T}_t$.

(2) If based on $\mathbf{J}$, $\mathbf{M}$ and $\mathcal{P}$, there is no available action, then goes to Step (3), otherwise let $\mathbf{J}_{t+1} = \mathbf{J}, \mathbf{M}_{t+1} = \mathbf{M}, \mathbf{W}_{t+1} = \mathbf{W}, \mathbb{T}_{t+1} = \mathbb{T}$.

(3) Since there is no action available, the system should precede the processing progress. Let $\hat{p}$ represent the minimum remaining processing time, defined as:

$$\hat{p} = \min \{p_{m,j,i} \mid (m, j, i, p_{m,j,i}) \in \mathbf{W}\}.$$

Then the system time should updated as $\hat{\mathbb{T}} = \mathbb{T} + \hat{p}$. The set $\hat{\mathbf{W}}$ is then updated as follows:

$$\hat{\mathbf{W}} = \{(m, j, i, p_{m,j,i} - \hat{p}_t) \mid p_{m,j,i} > \hat{p}_t\}.$$

This update process removes machine-operation pairs where the remaining processing time is exactly $\hat{p}$, while reducing the remaining time for all other pairs accordingly. We define the set of machines that are released as:

$$\mathcal{M}^{\text{released}} = \{m \mid (m, j, i, p_{m,j,i}) \in \mathbf{W}, p_{m,j,i} = \hat{p}\}.$$

Similarly, the set of jobs that need to be updated is given by:

$$\mathcal{J}^{\text{released}} = \{j \mid (m, j, i, p_{m,j,i}) \in \mathbf{W}, p_{m,j,i} = \hat{p}\}.$$

Additionally, let

$$\mathcal{J}^{\text{done}} = \{j \mid (\mathbf{J}_t)_{j,1} = \mathbb{k}_j\}$$

denote the set of jobs that have already completed their final operation. The job status is updated as follows for the first column:

$$(\hat{\mathbf{J}})_{j,1} = \begin{cases} 0, & \text{if } j \in \mathcal{J}^{\text{released}} \cap \mathcal{J}^{\text{done}}, \\ (\mathbf{J})_{j,1} + 1, & \text{if } j \in \mathcal{J}^{\text{released}} \setminus \mathcal{J}^{\text{done}}, \\ (\mathbf{J})_{j,1}, & \text{if } j \notin \mathcal{J}^{\text{released}}. \end{cases}$$

- If $j \in \mathcal{J}^{\text{released}} \cap \mathcal{J}^{\text{done}}$, the job has completed its final operation.

- If $j \in \mathcal{J}^{\text{released}} \setminus \mathcal{J}^{\text{done}}$, the job has finished its current operation but still has remaining steps.

For the second column, the update rule is:

$$(\hat{\mathbf{J}})_{j,2} = \begin{cases} 0, & \text{if } j \in \mathcal{J}^{\text{released}}, \\ (\mathbf{J})_{j,2}, & \text{if } j \notin \mathcal{J}^{\text{released}}. \end{cases}$$

The machine status is updated as follows:

$$(\mathbf{M}_{t+1})_m = \begin{cases} 0, & \text{if } m \in \mathcal{M}^{\text{released}}, \\ (\mathbf{M})_m, & \text{if } m \notin \mathcal{M}^{\text{released}}. \end{cases}$$

This ensures that machines in $\mathcal{M}^{\text{released}}$ are marked as available, while others retain their previous status. Let $\mathbf{J} = \hat{\mathbf{J}}, \mathbf{M} = \hat{\mathbf{M}}, \mathbf{W} = \hat{\mathbf{W}}, \mathbb{T} = \hat{\mathbb{T}}$, then goes to Step (2).

B. Model Illustration

II. Proofs in Details

A. Proof of Lemma 1 (with the original lemma provided)

Lemma 1. *Given a policy π and its corresponding action-priority vector Φ, suppose $G(\tau') > V^\pi(s_0')$, then for the new policy π' and its corresponding Φ' constructed according to (8), we have following relations:*

- *$\pi(a|s) = \pi'(a|s), \forall a \neq a_0', a \in \mathcal{A}(s)$, if a_0' is invalid for state s.*
- *$\pi'(a|s) < \pi(a|s), \forall a \neq a_0', a \in \mathcal{A}(s)$, if a_0' is valid for state s. More specifically, if s is the state that satisfies $\mathcal{A}(s) \subseteq \mathcal{A}(s_0')$, then $\pi'(a|s) \leq \delta\pi(a|s)$ with equality holds when $\mathcal{A}(s) = \mathcal{A}(s_0')$.*

For the case $G(\tau') < V^\pi(s_0')$, the first result still holds, while the second result changes the inequality $\leq$ to $\geq$.

Proof. Here, we provide the proof for the case $G(\tau') > V^\pi(s_0')$. The proof for the converse case is similar.

- If $\tilde{a}_0$ is invalid for state s, then $\mathbb{I}(s, \tilde{a}_0) = 0$.

$$\begin{aligned} \pi'(a|s) &= \frac{\phi_a'}{\sum_j \mathbb{I}(s,j)\phi_j'} = \frac{\phi_a'}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\phi_j'} \\ &= \frac{\delta\phi_a}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\delta\phi_j} = \frac{\delta\phi_a}{\delta\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\phi_j} \\ &= \frac{\phi_a}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\phi_j} = \frac{\phi_a}{\sum_j \mathbb{I}(s,j)\phi_j} \\ &= \pi(a|s) \end{aligned}$$

- If $\tilde{a}_0$ is valid for state s, then $\mathbb{I}(s, \tilde{a}_0) = 1$.

$$\begin{aligned} \pi'(a|s) &= \frac{\phi_a'}{\sum_j \mathbb{I}(s,j)\phi_j'} = \frac{\phi_a'}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\phi_j' + \phi_{\tilde{a}_0}'} \\ &= \frac{\delta\phi_a}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\delta\phi_j + \phi_{\tilde{a}_0} + (1-\delta)\sum_{j\neq\tilde{a}_0} \mathbb{I}(\tilde{s}_0,j)\phi_j} < \frac{\delta\phi_a}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\delta\phi_j + \delta\phi_{\tilde{a}_0}} \\ &= \frac{\phi_a}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\phi_j + \phi_{\tilde{a}_0}} = \frac{\phi_a}{\sum_j \mathbb{I}(s,j)\phi_j} \\ &= \pi(a|s) \end{aligned}$$

Now, if $\mathcal{A}(s) \subseteq \mathcal{A}(\tilde{s}_0)$, we have $\sum_{j\neq\tilde{a}_0} \mathbb{I}(\tilde{s}_0,j)\phi_j \geq \sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\phi_j$ since $\phi_i > 0$. Therefore,

$$\begin{aligned} \frac{\delta\phi_a}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\delta\phi_j + \phi_{\tilde{a}_0} + (1-\delta)\sum_{j\neq\tilde{a}_0} \mathbb{I}(\tilde{s}_0,j)\phi_j} &\leq \frac{\delta\phi_a}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\delta\phi_j + \phi_{\tilde{a}_0} + (1-\delta)\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\phi_j} \\ &= \frac{\delta\phi_a}{\sum_{j\neq\tilde{a}_0} \mathbb{I}(s,j)\phi_j + \phi_{\tilde{a}_0}} = \delta\pi(a|s) \end{aligned}$$

That is $\pi'(a|s) \leq \pi(a|s)$. ■

B. Proof of Lemma 2 (with the original lemma provided)

Lemma 2. *Suppose $\mathcal{A}_m(s_0')$ satisfying **CD3** and $\tau \in \mathcal{T}$, then $t_\tau^+ \geq 2$.*

Proof. Since $\tau \in \mathcal{T}$, by the definition of $\mathcal{T}$, we have $\tau = (s_0', a_0, \cdots)$, where either $a_0 \in \mathcal{A}_m(s_0') \cup \mathcal{A}_{(j,i)}(s_0')$ or $a_0 \notin \mathcal{A}_m(s_0') \cup \mathcal{A}_{(j,i)}(s_0')$. If $a_0 \in \mathcal{A}_m(s_0') \cup \mathcal{A}_{(j,i)}(s_0')$, then by **CD3**, we directly obtain $t_\tau^+ \geq 2$. Otherwise, if $a_0 \notin \mathcal{A}_m(s_0')$, suppose $a_0 = (m, j, i)$ and $a_0' = (m', j', i')$. Since $m \neq m'$, and $j \neq j'$, $i \neq i'$, it follows that after executing action a_0, action a_0' is still available. Hence, we conclude that $t_\tau^+ \geq 2$. ■

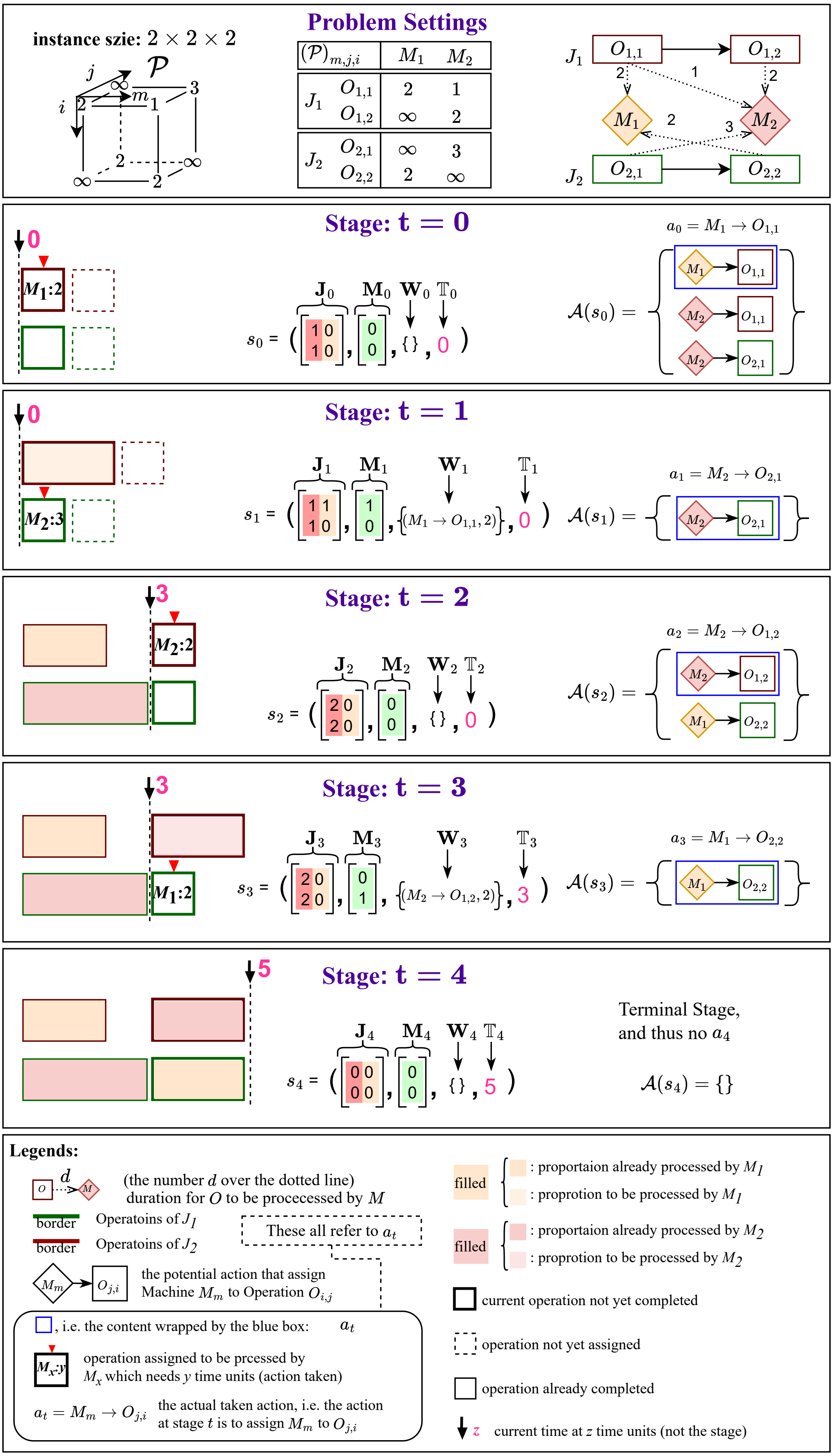

Fig. 1. Model Illustration

C. Proof of Lemma 3 (with the original lemma provided)

Lemma 3. *Suppose $\tau \in \mathcal{T}_1$, then $a'_0 \notin \mathcal{A}_{t^+_\tau}(\tau)$. If π' is updated according to updating rule (8), then*

$$\mathrm{P}^{\pi'}_\tau < \frac{\delta^2(1-\delta\eta)\sum_j \mathbb{I}_j(s)\phi_j}{(1-\eta)(\delta\sum_j \mathbb{I}_j(s)\phi_j + (1-\delta)\sum_j \mathbb{I}(s'_0, j)\phi_j)}\mathrm{P}^\pi_\tau, s \in \tau$$

where $\mathrm{P}^{\pi'}_\tau$ represents the probability of τ given policy π'. π' is the new policy and π is the old one.

Proof. Suppose $\tau = (s'_0, a_0, \dots, a_{t^+_\tau - 1}, s_{t^+_\tau}, \dots)$. By the definition of $\mathcal{T}_1$, we have $a_t \neq a'_0, \forall 0 \leq t \leq t^+_\tau - 1$. Then, by Lemma 1, it follows that $\pi'(a_0|s'_0) = \delta\pi(a_0|s'_0)$ and $\pi'(a_t|s_t) < \delta\pi(a_t|s_t)$ for $1 \leq t \leq t^+_\tau - 1$.

In the worst case, if action a'_0 is selected at some state s in τ, then

$$\pi(a'_0|s) = \frac{\phi_{a'_0}}{\sum_{j\neq a'_0}\mathbb{I}_j(s)\phi_j + \phi_{a'_0}}$$

and

$$\pi'(a'_0|s) = \frac{\phi_{a'_0} + (1-\delta)\sum_{j\neq a'_0}\mathbb{I}_j(s'_0)\phi_j}{\delta\sum_{j\neq a'_0}\mathbb{I}_j(s)\phi_j + \phi_{a'_0} + (1-\delta)\sum_{j\neq a'_0}\mathbb{I}_j(s'_0)\phi_j}.$$

Again, by Lemma 1, once action a'_0 is selected, we have $\pi'(a|s) = \pi(a|s)$. Thus, for some $s \in \tau$, it follows that

$$\begin{aligned} p^{\pi'}_\tau &< \frac{\delta^{t^+_\tau}[\phi_{a'_0} + (1-\delta)\sum_{j\neq a'_0}\mathbb{I}_j(s'_0)\phi_j]}{\delta\sum_{j\neq a'_0}\mathbb{I}_j(s)\phi_j + \phi_{a'_0} + (1-\delta)\sum_{j\neq a'_0}\mathbb{I}_j(s'_0)\phi_j} \cdot \frac{\sum_{j\neq a'_0}\mathbb{I}_j(s)\phi_j + \phi_{a'_0}}{\phi_{a'_0}} p^\pi_\tau, \quad \text{for some } s \in \tau \\ &= \frac{\delta^{t^+_\tau}(1-\delta\eta)\sum_j \mathbb{I}_j(s)\phi_j}{(1-\eta)(\delta\sum_j \mathbb{I}_j(s)\phi_j + (1-\delta)\sum_j \mathbb{I}_j(s'_0)\phi_j)} p^\pi_\tau. \end{aligned}$$

Since, by Lemma 2, we have $t^+_\tau \geq 2$, it follows that

$$p^{\pi'}_\tau < \frac{\delta^2(1-\delta\eta)\sum_j \mathbb{I}_j(s)\phi_j}{(1-\eta)(\delta\sum_j \mathbb{I}_j(s)\phi_j + (1-\delta)\sum_j \mathbb{I}_j(s'_0)\phi_j)} p^\pi_\tau.$$

■

D. Proof of Lemma 4 (with the original lemma provided)

Lemma 4. *Suppose $\tau \in \mathcal{T}$, $a'_0 \in \mathcal{A}_{t^+_\tau}(\tau)$ and a'_0 is executed at stage t, we know $0 \leq t < t^+_\tau$. Then,*

$$\mathrm{P}^{\pi'}_\tau < \frac{\delta^t(1-\delta\eta)}{1-\eta}\mathrm{P}^\pi_\tau$$

Proof. By definition, the set $\mathcal{A}_{t^+_\tau}(\tau)$ consists of the actions taken in trajectory τ before stage t^+_τ. Since $a'_0 \in \mathcal{A}_{t^+_\tau}(\tau)$, we can express τ as

$$\tau = (s'_0, a_0, \dots, s_t, a'_0, \dots, a_{t^+_\tau - 1}, s_{t^+_\tau}, \dots).$$

Then, by Lemma 1, it follows that $\pi'(a_0|s'_0) = \delta\pi(a_0|s'_0)$ and $\pi'(a_{t'}|s_{t'}) < \delta\pi(a_{t'}|s_{t'})$ for $1 \leq t' \leq t-1$. At state s_t, we have

$$\pi(a'_0|s_t) = \frac{\phi_{a'_0}}{\sum_j \mathbb{I}_j(s_t)\phi_j}$$

and

$$\pi'(a'_0|s_t) = \frac{\phi_{a'_0} + (1-\delta)\sum_{j\neq a'_0}\mathbb{I}_j(s'_0)\phi_j}{\delta\sum_{j\neq a'_0}\mathbb{I}_j(s_t)\phi_j + \phi_{a'_0} + (1-\delta)\sum_{j\neq a'_0}\mathbb{I}_j(s'_0)\phi_j}.$$

From Lemma 1, once action a'_0 is taken, we have $\pi'(a|s_{t'}) = \pi(a|s_{t'})$ for $t' > t$. Thus, we obtain

$$\begin{aligned} p^{\pi'}_\tau &< \frac{\delta^t(\phi_{a'_0} + (1-\delta)\sum_{j\neq a'_0}\mathbb{I}_j(s'_0)\phi_j)}{\delta\sum_{j\neq a'_0}\mathbb{I}_j(s_t)\phi_j + \phi_{a'_0} + (1-\delta)\sum_{j\neq a'_0}\mathbb{I}_j(s'_0)\phi_j} \cdot \frac{\sum_j \mathbb{I}_j(s_t)\phi_j}{\phi_{a'_0}} p^\pi_\tau \\ &< \frac{\delta^t(\delta\phi_{a'_0} + (1-\delta)\sum_j \mathbb{I}_j(s'_0)\phi_j)}{\phi_{a'_0}} p^\pi_\tau \\ &= \frac{\delta^t(1-\delta\eta)}{1-\eta} p^\pi_\tau. \end{aligned}$$

Such relaxation exactly match the required inequality in Lemma 4. ■

E. Proof of Lemma 5 (with the original lemma provided)

Lemma 5. *Suppose that $\tilde{\tau} := (s'_0, a'_0, \tilde{s}_1, \tilde{a}_1, \ldots, \tilde{a}_{t^+_{\tilde{\tau}}-1}, \tilde{s}_{t^+_{\tilde{\tau}}}, ...) \in \mathcal{T}_3$, and the set $\{a'_0, \tilde{a}_1, \ldots, a_{t^+_{\tilde{\tau}}-1}\}$ contains all the actions taken before the first change in system time within trajectory $\tilde{\tau}$. Next, define $\tilde{\tau}^{(t)}$ as the trajectory whose first $t^+_{\tilde{\tau}}$ actions match the set $\{a'_0, \tilde{a}_1, \ldots, a_{t^+_{\tilde{\tau}}-1}\}$, and satisfy $\tilde{\tau}^{(t)}|\tilde{s}_{t^+_{\tilde{\tau}}} = \tilde{\tau}|\tilde{s}_{t^+_{\tilde{\tau}}}$. By the existence of the **Permutation Irrelevance** Property (Proposition 1), the equality $\tilde{\tau}^{(t)}|\tilde{s}_{t^+_{\tilde{\tau}}} = \tilde{\tau}|\tilde{s}_{t^+_{\tilde{\tau}}}$ is well-defined.*

The order of the first $t^+_{\tilde{\tau}}$ actions in $\tilde{\tau}^{(t)}$ is given by

$$a^{(t)}_{t'} = \begin{cases} a_1, & \text{if } t' = 0, \\ a'_0, & \text{if } t' = t, \\ a_{t'+1}, & \text{if } 0 < t' < t, \\ a_{t'}, & \text{if } t' > t, \end{cases} \tag{1}$$

where $0 \leq t' \leq t^+_{\tilde{\tau}} - 1$.

Then, an equivalent expression for the set $\mathcal{T}_2$ is given by

$$\mathcal{T}_2 = \bigcup_{\tilde{\tau} \in \mathcal{T}_3} \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}.$$

Similarly, the set $\mathcal{T}_4$ can be expressed in the same form. In other words, $\mathcal{T}_4$ is a special case of this relation, since $\mathcal{T}_5$ contains only a single element, τ'. Specifically, we have

$$\mathcal{T}_4 = \{\tau'^{(0)}, \ldots, \tau'^{(t^+_{\tau'}-1)}\}.$$

where $\tau'^{(t)}$ for $0 \leq t \leq t^+_{\tau'} - 1$ is defined similarly, satisfying $\tau'^{(t)}|s'_{t^+_{\tau'}} = \tau'|s'_{t^+_{\tau'}}$.

Proof. We first derive an equivalent expression for the set $\mathcal{T}_2$. Recall its definition:

$$\mathcal{T}_2 = \left\{ \tau \mid \exists \tilde{\tau} \in \mathcal{T}_3, \text{ s.t. } \mathcal{A}_{t^+_{\tilde{\tau}}}(\tau) = \mathcal{A}_{t^+_{\tilde{\tau}}}(\tilde{\tau}), \tau \neq \tilde{\tau}, \tau \notin \mathcal{T}_3, \tau \in \mathcal{T} \right\}.$$

For any $\tau \in \mathcal{T}_2$, there exists some $\tilde{\tau} \in \mathcal{T}_3$ such that $\mathcal{A}_{t^+_{\tilde{\tau}}}(\tau) = \mathcal{A}_{t^+_{\tilde{\tau}}}(\tilde{\tau})$. This implies that the action set before the first system change is identical for both trajectories. Since $\tau \neq \tilde{\tau}$, the sequence of actions before stage $t^+_{\tilde{\tau}}$ must be a permutation of the set $\mathcal{A}_{t^+_{\tilde{\tau}}}$. Consequently, $\tau \in \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}$. Thus, we conclude that

$$\mathcal{T}_2 \subseteq \bigcup_{\tilde{\tau} \in \mathcal{T}_3} \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}.$$

Conversely, suppose $\tau \in \bigcup_{\tilde{\tau} \in \mathcal{T}_3} \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}$. Then, there must exist some $\tilde{\tau} \in \mathcal{T}_3$ such that $\tau \in \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}$. By the definition of the set $\{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}$ (see equation (1)), we have

$$\mathcal{A}_{t^+_{\tilde{\tau}}}(\tau) = \mathcal{A}_{t^+_{\tilde{\tau}}}(\tilde{\tau}) \quad \text{and} \quad \tau \neq \tilde{\tau}.$$

This implies $\tau \in \mathcal{T}_2$. Thus, $\bigcup_{\tilde{\tau} \in \mathcal{T}_3} \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\} \subseteq \mathcal{T}_2$. Combining both directions, we can then conclude that $\mathcal{T}_2 = \bigcup_{\tilde{\tau} \in \mathcal{T}_3} \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}$. The proof of the equivalence relation for the set $\mathcal{T}_4$ is similar and follows the same structure as the proof above. ■

F. Proof of Proposition 1 (with the original proposition provided)

Proposition 1 (Permutation Irrelevant). *Suppose $(s_0, a_0, \cdots, s_t, a_t, s_{t+1})$ is a state-action sequence with $\mathbb{T}_0 = \mathbb{T}_t < \mathbb{T}_{t+1}$, meaning that the system times of states s_0 and s_t are the same and both are less than that of s_{t+1}. Then, consider any permutation of the action set $\{a_0, \ldots, a_t\}$, which forms a new ordered action sequence. If we execute these actions in the new given order, the resulting state will be exactly s_{t+1}.*

Proof. It's equivalent to prove the existence of some function (or more generally, some mapping) h that maps the input $(s_0, \{a_1, a_2, \cdots, a_t\})$ to s_{t+1}, i.e. we shall have $s_{t+1} = h(s_0, \{a_1, a_2, \cdots, a_t\})$.

As we write the action as $a = (m, j, i)$ and the calculation of $\mathbf{W}$ actually adds the term $(m, j, i, (\mathcal{P})_{m,j,i})$, we may rewrite such term as $(a, (\mathcal{P})_a)$. Denote the set $\underline{\mathbf{W}} \doteq \mathbf{W}_0 \cup \{(a_0, (\mathcal{P})_{a_0}), (a_1, (\mathcal{P})_{a_1}), \cdots, (a_t, (\mathcal{P})_{a_t})\}$, the processing time $\tilde{p} \doteq \min\{(\mathcal{P})_a | (a, (\mathcal{P})_a) \in \underline{\mathbf{W}}\}$, the action set $\tilde{\mathcal{A}} \doteq \{a \mid (\mathcal{P})_a = \tilde{p}\}$, and the sets $\overline{\mathbf{W}} \doteq \underline{\mathbf{W}} \setminus \left\{(a, (\mathcal{P})_a) \middle| a \in \tilde{\mathcal{A}}\right\}$, $\tilde{\mathbf{W}} \doteq \overline{\mathbf{W}} \setminus \mathbf{W}_0$. We also define the following auxiliary sets:

- $\tilde{\mathcal{J}} \doteq \left\{ j \middle| (m, j, i) \in \tilde{\mathcal{A}} \right\}$
- $\hat{\mathcal{J}} \doteq \{ j | (m, j, i, (\mathcal{P})_{m,j,i}) \in \overline{\mathbf{W}} \}$
- $\bar{\mathcal{J}} \doteq \{ j \mid (\mathbf{J}_0)_{j,1} = \Bbbk_j \}$
- $\tilde{\mathcal{M}} \doteq \left\{ m \middle| (m, j, i, (\mathcal{P})_{m,j,i}) \in \tilde{\mathbf{W}} \right\}$
- $\hat{\mathcal{M}} \doteq \left\{ m \middle| (m, j, i) \in \tilde{\mathcal{A}} \right\} \cap \{ m | (m, j, i, (\mathcal{P})_{m,j,i}) \in \mathbf{W}_0 \}$

Now, we can calculate s_{t+1} in as follows according to the system dynamics:

- $(\mathbf{J}_{t+1})_{j,1} = \begin{cases} 0, & \text{if } j \in \tilde{\mathcal{J}} \cap \bar{\mathcal{J}}; \\ (\mathbf{J}_0)_{j,1} + 1, & \text{if } j \in \tilde{\mathcal{J}} \setminus \bar{\mathcal{J}}; \\ (\mathbf{J}_0)_{j,1}, & \text{otherwise.} \end{cases}$
- $(\mathbf{J}_{t+1})_{j,1} = \begin{cases} 0, & \text{if } j \in \tilde{\mathcal{J}}; \\ 1, & \text{if } j \in \hat{\mathcal{J}}. \end{cases}$
- $(\mathbf{M}_{t+1})_m = \begin{cases} 0, & \text{if } m \in \hat{\mathcal{M}}; \\ 1, & \text{if } j \in \tilde{\mathcal{M}}. \end{cases}$
- $\mathbf{W}_{t+1} = \left\{ (a, (\mathcal{P})_a - \tilde{p}) \middle| (a, (\mathcal{P}_a \in \overline{\mathbf{W}}) \right\};$
- $\mathbb{T}_{t+1} = \mathbb{T}_0 + \tilde{p}.$

Here the second part of the inputs of h is some set $\{a_0, a_1, \dots, a_t\}$ rather than some sequence. Since a set is identical no matter how the elements are permuted, with $s_{t+1} = h(s_0, \{a_0, a_1, \dots, a_t\})$, we can tell that the permutation irrelevant property exists.

A potential issue is that when the sequence $(a_0, a_1, \cdots, a_t)$ is reordered, the proof above requires that these actions are still feasible at each time step. This is true: for $t_x, t_y \in \{0, 1, \cdots, t\}$, suppose $a_{t_x} = (m_{t_x}, j_{t_x}, i_{t_x})$ and $a_{t_y} = (m_{t_y}, j_{t_y}, i_{t_y})$, as long as $t_x \neq t_y$, we shall have $m_{t_x} \neq m_{t_y}$ and $(j_{t_x} \neq j_{t_y})$. Since the assignments at each time step are mutually exclusive, the feasibility can be guaranteed even if the action sequence is permuted.

Further explanation is shown in Figure 2. ■

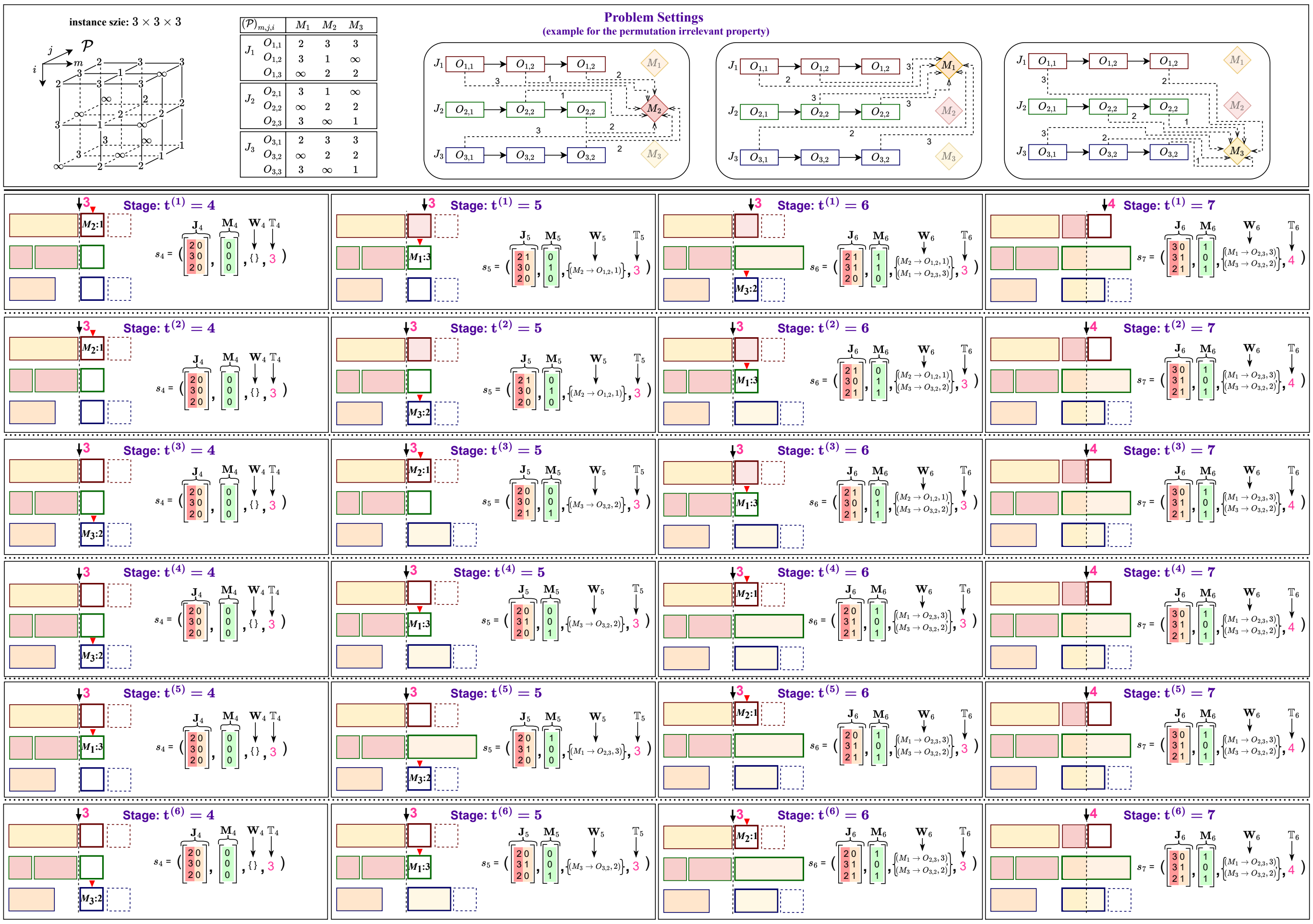

Fig. 2. Permutation Irrelevant Property Explanation

G. Proof of Proposition 2 (with the original proposition provided)

Proposition 2. *For any $\delta \in (0, 1)$, condition **CD1** is satisfied, i.e.,*

$$\frac{\delta^2(1-\delta\eta)}{(1-\eta)\left(\delta + (1-\delta)\frac{\sum_j \mathbb{I}(s_0', j)\phi_j}{\sum_j \mathbb{I}_j(s)\phi_j}\right)} < \frac{1-(1-\delta\eta)\mathrm{P}^{\pi}_{\tau'|(s_0', a_0')}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s_0', a_0')}}, \quad \forall s \neq s_0'.$$

Proof. Denote $\sigma = \frac{\sum_j \mathbb{I}(s_0',j)\phi_j}{\sum_j \mathbb{I}_j(s)\phi_j} > 0$, then

$$\begin{aligned}
&\frac{\delta^2(1-\delta\eta)}{(1-\eta)(\delta+(1-\delta)\frac{\sum_j \mathbb{I}(s_0',j)\phi_j}{\sum_j \mathbb{I}_j(s)\phi_j})} < \frac{1-(1-\delta\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}} \\
\Leftrightarrow &\frac{\delta^2(1-\delta\eta)}{(1-\eta)((1-\sigma)\delta+\sigma)} < \frac{1-(1-\delta\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}} \\
\Leftrightarrow &\delta^2(1-\delta\eta)\frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta} < ((1-\sigma)\delta+\sigma)(1-(1-\delta\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}) \\
\Leftrightarrow &\delta^2(1-\delta\eta)\frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta} < (1-\sigma)\eta\mathrm{P}^\pi_{\tau'|(s_0',a_0')}\delta^2 + [(1-\sigma)(1-\mathrm{P}^\pi_{\tau'|(s_0',a_0')}) + \sigma\eta\mathrm{P}^\pi_{\tau'|(s_0',a_0')}]\delta + \sigma(1-\mathrm{P}^\pi_{\tau'|(s_0',a_0')}) \\
\Leftrightarrow &\left[\frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta} - (1-\sigma)\eta\mathrm{P}^\pi_{\tau'|(s_0',a_0')}\right]\delta^2 - [(1-\sigma)(1-\mathrm{P}^\pi_{\tau'|(s_0',a_0')}) + \sigma\eta\mathrm{P}^\pi_{\tau'|(s_0',a_0')}]\delta - \sigma(1-\mathrm{P}^\pi_{\tau'|(s_0',a_0')}) < \\
&\frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta}\eta\delta^3
\end{aligned}$$

Let $f_1(\delta) = \left[\frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta} - (1-\sigma)\eta\mathrm{P}^\pi_{\tau'|(s_0',a_0')}\right]\delta^2 - [(1-\sigma)(1-\mathrm{P}^\pi_{\tau'|(s_0',a_0')}) + \sigma\eta\mathrm{P}^\pi_{\tau'|(s_0',a_0')}]\delta - \sigma(1-\mathrm{P}^\pi_{\tau'|(s_0',a_0')})$ and $f_2(\delta) = \frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta}\eta\delta^3$. We find that $f_1(1) = f_2(1)$ and $f_1(0) = -\sigma(1-\mathrm{P}^\pi_{\tau'|(s_0',a_0')}) < 0$. Since

$$\begin{aligned}
\frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta} - (1-\sigma)\eta\mathrm{P}^\pi_{\tau'|(s_0',a_0')} &< \frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta} - \eta\mathrm{P}^\pi_{\tau'|(s_0',a_0')} \\
&= \frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')} - \eta(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta} \\
&= \frac{1-(1-\eta^2)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta} \\
&> \frac{1-(1-\eta^2)}{1-\eta} = \frac{\eta^2}{1-\eta} > 0
\end{aligned}$$

Hence, $f_1(\delta)$ is a convex quadratic function. Since

$$\frac{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-\eta}\eta > 0,$$

it follows that, as shown in Figure 3, we have $f_1(\delta) < f_2(\delta), \quad \forall\delta \in (0,1)$. ∎

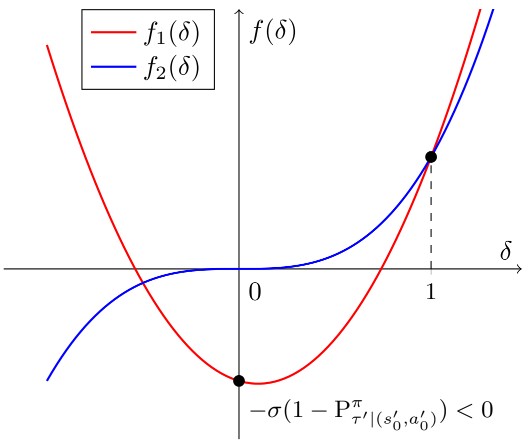

Fig. 3. $f_1(\delta), f_2(\delta)$

H. Proof of Proposition 3 (with the original proposition provided)

Proposition 3. *If* $1 - \frac{\eta}{1-\eta} > \alpha + \beta$*, where*

$$\alpha = \frac{\eta\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}, \quad \beta = \frac{1}{1-\eta}\frac{\eta}{1-(1-\eta)\mathrm{P}^\pi_{\tau'|(s_0',a_0')}}\frac{\sum_j \mathbb{I}(s_0',j)\phi_j}{\sum_j \mathbb{I}(s_1',j)\phi_j},$$

then there exist a $\epsilon > 0$*, such that when* $\delta \in (\epsilon, 1)$*,* ***CD2*** *is satisfied.*

Remark 1. *By the definition of*

$$\eta = \frac{\sum_{j \neq a_0'} \mathbb{I}(s_0', j)\phi_j}{\sum_j \mathbb{I}(s_0', j)\phi_j},$$

as well as the trajectory τ', states s_0' and s_1', the values of α and β can be easily computed. Thus, the condition

$$1 - \frac{\eta}{1-\eta} > \alpha + \beta$$

can be efficiently verified.

Proof. Given $\tau \in \mathcal{T}_3, \mathcal{T}_5$, let $g_1(\delta) = \sum_{t=1}^{t_\tau^+ - 1} \delta^t \frac{1-\delta\eta}{1-\eta} \mathrm{P}^\pi_{\tau^{(t)}}$ and

$$g_2(\delta) = \delta \left(\sum_{t=1}^{t_\tau^+ - 1} \frac{\eta \mathrm{P}^\pi_{\tau'|(s_0', a_0')} \mathrm{P}^\pi_{\tau^{(t)}}}{1 - (1-\eta)\mathrm{P}^\pi_{\tau'|(s_0', a_0')}} + \frac{\eta \mathrm{P}^\pi_{\tau|(s_0', a_0')}}{1 - (1-\eta)\mathrm{P}^\pi_{\tau'|(s_0', a_0')}} \right) \sum_{t=1}^{t_\tau^+ - 1} \frac{(1 - \mathrm{P}^\pi_{\tau'|(s_0', a_0')})\mathrm{P}^\pi_{\tau^{(t)}}}{1 - (1-\eta)\mathrm{P}^\pi_{\tau'|(s_0', a_0')}} - \frac{\eta \mathrm{P}^\pi_{\tau|(s_0', a_0')}}{1 - (1-\eta)\mathrm{P}^\pi_{\tau'|(s_0', a_0')}}$$

The following theorem is established in Spivak, M.'s book [1], and we will use it directly:

Theorem (for Proposition 3)**.** *Let $g_1(\delta)$ and $g_2(\delta)$ be continuously differentiable functions on $(0, 1]$. If*

$$g_1(1) = g_2(1) \quad \textit{and} \quad g_1'(1) > g_2'(1),$$

then there exists $\epsilon > 0$ such that

$$g_1(\delta) < g_2(\delta), \quad \forall \delta \in (\epsilon, 1).$$

Since $g_1(1) = g_2(1)$ and both functions are continuously differentiable, it suffices to show that $g_1'(1) > g_2'(1)$. If this holds, then for δ close to 1, we have $g_1(\delta) < g_2(\delta)$.

The derivatives at $\delta = 1$ are given by

$$g_1'(1) = \sum_{t=1}^{t_\tau^+ - 1} \left(t - \frac{\eta}{1-\eta} \right) \mathrm{P}^\pi_{\tau^{(t)}}$$

and

$$g_2'(1) = \sum_{t=1}^{t_\tau^+ - 1} \frac{\eta \mathrm{P}^\pi_{\tau'|(s_0', a_0')} \mathrm{P}^\pi_{\tau^{(t)}}}{1 - (1-\eta)\mathrm{P}^\pi_{\tau'|(s_0', a_0')}} + \frac{\eta \mathrm{P}^\pi_{\tau|(s_0', a_0')}}{1 - (1-\eta)\mathrm{P}^\pi_{\tau'|(s_0', a_0')}}.$$

Observing the dependency on t_τ^+, we see that when t_τ^+ increases by 1,

$$g_1'(1) \quad \text{increases by} \quad (t_\tau^+ - \frac{\eta}{1-\eta})\mathrm{P}^\pi_{\tau^{(t_\tau^+)}},$$

whereas

$$g_2'(1) \quad \text{increases by} \quad \frac{\eta \mathrm{P}^\pi_{\tau'|(s_0', a_0')} \mathrm{P}^\pi_{\tau^{(t_\tau^+)}}}{1 - (1-\eta)\mathrm{P}^\pi_{\tau'|(s_0', a_0')}}.$$

By Lemma 2, we know that $t_\tau^+ \geq 2$ and since the increase in $g_1'(1)$ is greater than that of $g_2'(1)$, therefore, we only need to consider the base case $t_\tau^+ = 2$, where

$$g_1'(1) = \left(1 - \frac{\eta}{1-\eta} \right) \mathrm{P}^\pi_{\tau^{(1)}}$$

and

$$g_2'(1) = \frac{\eta \mathrm{P}^\pi_{\tau'|(s_0', a_0')} \mathrm{P}^\pi_{\tau^{(1)}}}{1 - (1-\eta)\mathrm{P}^\pi_{\tau'|(s_0', a_0')}} + \frac{\eta \mathrm{P}^\pi_{\tau|(s_0', a_0')}}{1 - (1-\eta)\mathrm{P}^\pi_{\tau'|(s_0', a_0')}}.$$

Now, given that $t_\tau^+ = 2$, suppose $\tau = (s_0', a_0', s_1, a_1, s_2, \dots)$, then we have $\tau^{(1)} = (s_0', a_1, s_1^{(1)}, a_0', s_2, \dots)$. Thus,

$$\begin{aligned}
\mathrm{P}^\pi_{\tau|(s_0', a_0')} = \pi(a_1|s_1)\mathrm{P}^\pi_{\tau|(s_1, a_1)} &= \frac{\pi(a_1|s_1)\pi(a_1|s_0')\pi(a_0'|s_1^{(1)})\mathrm{P}^\pi_{\tau|(s_1, a_1)}}{\pi(a_1|s_0')\pi(a_0'|s_1^{(1)})} && (2) \\
&= \frac{\pi(a_1|s_1)\mathrm{P}^\pi_{\tau^{(1)}}}{\pi(a_1|s_0')\pi(a_0'|s_1^{(1)})} && (3) \\
&= \mathrm{P}^\pi_{\tau^{(1)}} \frac{\phi_{a_1}}{\sum_j \mathbb{I}(s_1, j)\phi_j} \frac{\sum_j \mathbb{I}(s_0', j)\phi_j}{\phi_{a_1}} \frac{\sum_j \mathbb{I}(s_1^{(1)}, j)\phi_j}{\phi_{a_0'}} && (4) \\
&= \mathrm{P}^\pi_{\tau^{(1)}} \frac{\sum_j \mathbb{I}(s_0', j)\phi_j}{\phi_{a_0'}} \frac{\sum_j \mathbb{I}(s_1^{(1)}, j)\phi_j}{\sum_j \mathbb{I}(s_1, j)\phi_j} && (5)
\end{aligned}$$

$$
\begin{aligned}
&= \mathrm{P}^{\pi}_{\tau^{(1)}} \frac{1}{1-\eta} \frac{\sum_j \mathbb{I}(s_1^{(1)}, j)\phi_j}{\sum_j \mathbb{I}(s_1, j)\phi_j} && (6)\\
&< \mathrm{P}^{\pi}_{\tau^{(1)}} \frac{1}{1-\eta} \frac{\sum_j \mathbb{I}(s_0', j)\phi_j}{\sum_j \mathbb{I}(s_1, j)\phi_j} && (7)
\end{aligned}
$$

The last inequality follows from the fact that $\mathcal{A}(s_1^{(1)}) \subset \mathcal{A}(s_0')$. Thus, we obtain

$$
\begin{aligned}
g_2'(1) &= \frac{\eta \mathrm{P}^{\pi}_{\tau'|(s_0',a_0')} \mathrm{P}^{\pi}_{\tau^{(1)}}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}} + \frac{\eta \mathrm{P}^{\pi}_{\tau|(s_0',a_0')}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}} \\
&< \left(\frac{\eta \mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}} + \frac{1}{1-\eta} \frac{\sum_j \mathbb{I}(s_0', j)\phi_j}{\sum_j \mathbb{I}(s_1, j)\phi_j} \right) \mathrm{P}^{\pi}_{\tau^{(1)}} \\
&= (\alpha + \beta)\, \mathrm{P}^{\pi}_{\tau^{(1)}}
\end{aligned}
$$

Thus, if $1 - \frac{\eta}{1-\eta} > \alpha + \beta$, then

$$
g_2'(1) < (\alpha+\beta)\mathrm{P}^{\pi}_{\tau^{(1)}} < \left(1 - \frac{\eta}{1-\eta}\right) \mathrm{P}^{\pi}_{\tau^{(1)}} = g_1'(1).
$$

This establishes the case for $t_\tau^+ = 2$. For $t_\tau^+ > 2$, the result follows from the previous discussion. Finally, by applying Theorem for Proposition 3, we complete the proof. ■

I. Proof of Theorem 1 (with the original theorem provided)

Theorem 1 (Policy Improvement-Version I)**.** *Given a policy π corresponding action-priority vector Φ and a trajectory $\tau' := (s_0', a_0', s_1', ...)$. Assume $a_0' = (m, j, i)$, let $\mathcal{A}_m(s_0') \subseteq \mathcal{A}(s_0')$ denote the set of feasible actions where the machine is fixed as M_m. Similarly, let $\mathcal{A}_{(j,i)}(s_0') \subseteq \mathcal{A}(s_0')$ denote the set of feasible actions where the job and operation are fixed as J_j and $O_{j,i}$, respectively. Suppose $G(\tau') > V^\pi(s_0')$, then we have new policy π' and its corresponding Φ' where*

$$
\phi_a' = \begin{cases} \phi_a + (1-\delta) \sum\limits_{a' \neq a} \mathbb{I}(s_0', a')\phi_{a'}, & \textit{if } a = a_0' \\ \delta \phi_a, & \textit{otherwise} \end{cases} \tag{8}
$$

*and $\delta \in (0,1)$. Here the policy π' and its corresponding Φ' updated according to (8), will perform better, that is $V^{\pi'}(s_0') > V^\pi(s_0')$ if δ satisfies **CD1**, **CD2** and $\mathcal{A}_m(s_0')$ satisfies **CD3**:*

CD1 $\frac{\delta^2(1-\delta\eta)}{(1-\eta)(\delta + (1-\delta)\frac{\sum_j \mathbb{I}(s_0',j)\phi_j}{\sum_j \mathbb{I}_j(s)\phi_j})} < \frac{1-(1-\delta\eta)\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}}, \forall s \neq s_0'$

CD2 $\sum_{t=1}^{t_\tau^+ - 1} \delta^t \frac{1-\delta\eta}{1-\eta} \mathrm{P}^{\pi}_{\tau(t)} < \delta(\sum_{t=1}^{t_\tau^+ - 1} \frac{\eta \mathrm{P}^{\pi}_{\tau'|(s_0',a_0')} \mathrm{P}^{\pi}_{\tau(t)}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}} + \frac{\eta \mathrm{P}^{\pi}_{\tau|(s_0',a_0')}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}}) + \sum_{t=1}^{t_\tau^+ - 1} \frac{(1-\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')})\mathrm{P}^{\pi}_{\tau(t)}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}} - \frac{\eta \mathrm{P}^{\pi}_{\tau|(s_0',a_0')}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}}, \forall \tau \in \mathcal{T}_3, \mathcal{T}_5$

CD3 *For any $a \in \mathcal{A}_m(s_0') \cup \mathcal{A}_{(j,i)}(s_0')$, after executing action a, the system time of the resulting state remains the same as $\mathbb{T}_0'$, which is the system time of s_0'.*

Proof. First, the value function $V^\pi(s_0')$ can be expressed as

$$
V^\pi(s_0') = \sum_{\tau \in \mathcal{T}_1} \mathrm{P}^{\pi}_{\tau} G(\tau) + \sum_{\tau \in \mathcal{T}_2} \mathrm{P}^{\pi}_{\tau} G(\tau) + \sum_{\tau \in \mathcal{T}_3} \pi(a_0'|s_0') \mathrm{P}^{\pi}_{\tau|(s_0',a_0')} G(\tau) + \sum_{\tau \in \mathcal{T}_4} \pi(a_0'|s_0') \mathrm{P}^{\pi}_{\tau|(s_0',a_0')} G(\tau) + \pi(a_0'|s_0') \mathrm{P}^{\pi}_{\tau'|(s_0',a_0')} G(\tau')
$$

Since $G(\tau') > V^\pi(s_0')$, we have

$$
G(\tau') > \frac{\sum_{\tau \in \mathcal{T}_1} \mathrm{P}^{\pi}_{\tau} G(\tau) + \sum_{\tau \in \mathcal{T}_2} \mathrm{P}^{\pi}_{\tau} G(\tau) + \sum_{\tau \in \mathcal{T}_3} \pi(a_0'|s_0') \mathrm{P}^{\pi}_{\tau|(s_0',a_0')} G(\tau) + \sum_{\tau \in \mathcal{T}_4} \pi(a_0'|s_0') \mathrm{P}^{\pi}_{\tau|(s_0',a_0')} G(\tau)}{1 - \pi(a_0'|s_0') \mathrm{P}^{\pi}_{\tau'|(s_0',a_0')}} \tag{9}
$$

Since in our modeling, once an action is executed, it will never be valid again, and together with Lemma 1, then

$$
\mathrm{P}^{\pi'}_{\tau|(s_0',a_0')} = \mathrm{P}^{\pi}_{\tau|(s_0',a_0')} \tag{10}
$$

Since

$$
\begin{aligned}
\pi'(a_0'|s_0') &= \frac{\phi_{a_0'}'}{\sum_j \mathbb{I}(s_0', j)\phi_j'} \\
&= \frac{\phi_{a_0'} + (1-\delta)\sum_{j \neq a_0'} \mathbb{I}(s_0', j)\phi_j}{\sum_{j \neq a_0'} \mathbb{I}(s_0', j)\delta\phi_j + \phi_{a_0'} + (1-\delta)\sum_{j \neq a_0'} \mathbb{I}(s_0', j)\phi_j} \\
&= \frac{\phi_{a_0'} + (1-\delta)\sum_{j \neq a_0'} \mathbb{I}(s_0', j)\phi_j}{\sum_{j \neq a_0'} \mathbb{I}(s_0', j)\phi_j + \phi_{a_0'}} = \frac{\phi_{a_0'} + (1-\delta)\sum_{j \neq a_0'} \mathbb{I}(s_0', j)\phi_j}{\sum_j \mathbb{I}(s_0', j)\phi_j} \\
&= \pi(a_0'|s_0') + \frac{(1-\delta)\sum_{j \neq a_0'} \mathbb{I}(s_0', j)\phi_j}{\sum_j \mathbb{I}(s_0', j)\phi_j}
\end{aligned}
$$

Since $\eta = \frac{\sum_{j \neq a'_0} \mathbb{I}(s'_0, j)\phi_j}{\sum_j \mathbb{I}(s'_0, j)\phi_j}$, then

$$\pi'(a'_0|s'_0) = \pi(a'_0|s'_0) + (1-\delta)\eta \tag{11}$$

and

$$\pi(a'_0|s'_0) = \frac{\phi_{a'_0}}{\sum_j \mathbb{I}(s'_0, j)\phi_j} = 1 - \eta \tag{12}$$

By (9), (10), (11) and (12),

$$\begin{aligned}
V^{\pi'}(s'_0) - V^{\pi}(s'_0) &= \sum_{\tau \in \mathcal{T}_1} (\mathrm{P}^{\pi'}_\tau - \mathrm{P}^{\pi}_\tau) G(\tau) + \sum_{\tau \in \mathcal{T}_2} (\mathrm{P}^{\pi'}_\tau - \mathrm{P}^{\pi}_\tau) G(\tau) + \sum_{\tau \in \mathcal{T}_3} (1-\delta)\eta \mathrm{P}^{\pi}_{\tau|(s'_0,a'_0)} G(\tau) \\
&\quad + \sum_{\tau \in \mathcal{T}_4} (1-\delta)\eta \mathrm{P}^{\pi}_{\tau|(s'_0,a'_0)} G(\tau) + (1-\delta)\eta \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} G(\tau') \\
&> \sum_{\tau \in \mathcal{T}_1} \left[\mathrm{P}^{\pi'}_\tau + \frac{(1-\delta)\eta \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}}{1 - \pi(a'_0|s'_0)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_\tau - \mathrm{P}^{\pi}_\tau \right] G(\tau) + \sum_{\tau \in \mathcal{T}_2} \left[\mathrm{P}^{\pi'}_\tau + \frac{(1-\delta)\eta \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}}{1 - \pi(a'_0|s'_0)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_\tau - \mathrm{P}^{\pi}_\tau \right] G(\tau) \\
&\quad + \sum_{\tau \in \mathcal{T}_3} \left[(1-\delta)\eta + \frac{(1-\delta)\eta(1-\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}}{1 - \pi(a'_0|s'_0)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \right] \mathrm{P}^{\pi}_{\tau|(s'_0,a'_0)} G(\tau) + \sum_{\tau \in \mathcal{T}_4} \left[\mathrm{P}^{\pi'}_\tau + \frac{(1-\delta)\eta \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}}{1 - \pi(a'_0|s'_0)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_\tau - \mathrm{P}^{\pi}_\tau \right] G(\tau) \\
&= \sum_{\tau \in \mathcal{T}_1} \left[\mathrm{P}^{\pi'}_\tau + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_\tau \right] G(\tau) + \sum_{\tau \in \mathcal{T}_2} \left[\mathrm{P}^{\pi'}_\tau + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_\tau \right] G(\tau) \\
&\quad + \sum_{\tau \in \mathcal{T}_3} \left[\frac{(1-\delta)\eta}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \right] \mathrm{P}^{\pi}_{\tau|(s'_0,a'_0)} G(\tau) + \sum_{\tau \in \mathcal{T}_4} \left[\mathrm{P}^{\pi'}_\tau + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_\tau \right] G(\tau)
\end{aligned}$$

Since $G(\tau) < 0$, **CD1** and Lemma 3, we have

$$\begin{aligned}
\sum_{\tau \in \mathcal{T}_1} \left[\mathrm{P}^{\pi'}_\tau + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_\tau \right] G(\tau) &> \sum_{\tau \in \mathcal{T}_1} \left[\frac{\delta^2(1-\delta\eta)\sum_j \mathbb{I}_j(s)\phi_j}{(1-\eta)(\delta\sum_j \mathbb{I}_j(s)\phi_j + (1-\delta)\sum_j \mathbb{I}(s'_0,j)\phi_j)} + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \right] \mathrm{P}^{\pi}_\tau G(\tau) \\
&= \sum_{\tau \in \mathcal{T}_1} \left[\frac{\delta^2(1-\delta\eta)}{(1-\eta)(\delta + (1-\delta)\frac{\sum_j \mathbb{I}(s'_0,j)\phi_j}{\sum_j \mathbb{I}_j(s)\phi_j})} + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \right] \mathrm{P}^{\pi}_\tau G(\tau) \\
&\geq 0
\end{aligned}$$

By Lemma 5, we have

$$\mathcal{T}_2 = \bigcup_{\tilde{\tau} \in \mathcal{T}_3} \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}$$

and

$$G(\tilde{\tau}^{(0)}) = \cdots = G(\tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}) = G(\tilde{\tau}).$$

Furthermore, for any $\tilde{\tau}^{(t)} \in \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}$ with $1 \leq t \leq t^+_{\tilde{\tau}} - 1$, it follows from (1) that the action a'_0 is taken at stage t. Then, by Lemma 4, we obtain

$$\mathrm{P}^{\pi'}_\tau G(\tilde{\tau}^{(t)}) = \mathrm{P}^{\pi'}_\tau G(\tilde{\tau}) < \frac{\delta^t(1-\delta\eta)}{1-\eta} \mathrm{P}^{\pi}_\tau G(\tilde{\tau}).$$

Hence, we have

$$\begin{aligned}
&\sum_{\tau \in \mathcal{T}_2} \left[\mathrm{P}^{\pi'}_\tau + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_\tau \right] G(\tau) + \sum_{\tau \in \mathcal{T}_3} \left[\frac{(1-\delta)\eta}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \right] \mathrm{P}^{\pi}_{\tau|(s'_0,a'_0)} G(\tau) && (13) \\
&= \sum_{\tilde{\tau} \in \mathcal{T}_3} \left(\sum_{\tilde{\tau}^{(t)} \in \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}} \left[\mathrm{P}^{\pi'}_{\tilde{\tau}^{(t)}} + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_{\tilde{\tau}^{(t)}} \right] G(\tilde{\tau}^{(t)}) + \left[\frac{(1-\delta)\eta}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \right] \mathrm{P}^{\pi}_{\tilde{\tau}|(s'_0,a'_0)} G(\tilde{\tau}) \right) && (14) \\
&= \sum_{\tilde{\tau} \in \mathcal{T}_3} \left(\sum_{\tilde{\tau}^{(t)} \in \{\tilde{\tau}^{(0)}, \ldots, \tilde{\tau}^{(t^+_{\tilde{\tau}}-1)}\}} \left[\mathrm{P}^{\pi'}_{\tilde{\tau}^{(t)}} + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_{\tilde{\tau}^{(t)}} \right] + \left[\frac{(1-\delta)\eta}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \right] \mathrm{P}^{\pi}_{\tilde{\tau}|(s'_0,a'_0)} \right) G(\tilde{\tau}) && (15) \\
&> \sum_{\tilde{\tau} \in \mathcal{T}_3} \left(\sum_{t=1}^{t^+_{\tilde{\tau}}-1} \left[\frac{\delta^t(1-\delta\eta)}{1-\eta} \mathrm{P}^{\pi}_{\tilde{\tau}^{(t)}} - \frac{1 - (1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_{\tilde{\tau}^{(t)}} \right] + \left[\frac{(1-\delta)\eta}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \right] \mathrm{P}^{\pi}_{\tilde{\tau}|(s'_0,a'_0)} \right) G(\tilde{\tau}) && (16) \\
&> 0 && (17)
\end{aligned}$$

The last inequality follows from **CD2**. Similarly, for the set $\mathcal{T}_4$, applying Lemma 4 and Lemma 5, we obtain

$$\sum_{\tau \in \mathcal{T}_4} \left[\mathrm{P}^{\pi'}_\tau + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_\tau \right] G(\tau) = \sum_{\tau'^{(t)} \in \{\tau'^{(0)}, \ldots, \tau'^{(t^+_{\tau'}-1)}\}} \left[\mathrm{P}^{\pi'}_{\tau'^{(t)}} + \frac{(1-\delta\eta) \mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1 - (1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_{\tau'^{(t)}} \right] G(\tau'^{(t)})$$

$$
\begin{aligned}
&= \sum_{t=1}^{t^+_{\tau'}-1} \left[\mathrm{P}^{\pi'}_{\tau'(t)} + \frac{(1-\delta\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)} - 1}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_{\tau'(t)} \right] G(\tau') \\
&> \sum_{t=1}^{t^+_{\tau'}-1} \left[\frac{\delta^t(1-\delta\eta)}{1-\eta} \mathrm{P}^{\pi}_{\tau'(t)} - \frac{1-(1-\delta\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}}{1-(1-\eta)\mathrm{P}^{\pi}_{\tau'|(s'_0,a'_0)}} \mathrm{P}^{\pi}_{\tau'(t)} \right] G(\tau') \\
&> 0
\end{aligned}
$$

The last inequality follows from **CD2**. Consequently, we have proven that $V^{\pi'}(s'_0) > V^{\pi}(s'_0)$. ■

III. Numerical Results

References

[1] M. Spivak, *Calculus*. Cambridge University Press, 2006.