
SUPPLEMENTARY MATERIALS: MULTI-AGENT SEQUENTIAL LEARNING VIA GRADIENT ASCENT CREDIT ASSIGNMENT

Oussama Sabri
University Medical Center
Hamburg-Eppendorf (UKE)
Hamburg, Germany
o.sabri@uke.de

Luc Lehericy
Laboratoire JAD, CNRS
Université Côte d'Azur (UCA)
Nice, France
luc.lehericy@univ-cotedazur.fr

Alexandre Muzy
Laboratoire I3S, CNRS
Université Côte d'Azur (UCA)
Sophia Antipolis, France
alexandre.muzy@univ-cotedazur.fr

1 Simulation environment and parameters

1.1 Simulation environment

We describe the simulation environment realized in this paper. All the Monte Carlo simulations were performed over $N = 30$ realizations, using the pseudo-random number generator Mersenne Twister from of library numpy 1.11.1, with seed = 748596. Each realization, indexed by $r \in [N]$, at episode $e \geq 1$, produces $Y^{(e),[r]}$ for any variable Y . Therefore, the mean and the standard deviation of Y over the realizations is given as

$$\forall e \geq 1 \mapsto \bar{Y}^{(e)} = \frac{1}{N} \sum_{r=1}^N Y^{(e),[r]} = \text{mean}\left((Y^{(e),[r]})_{r \in [N]}\right),$$
$$\forall e \geq 1 \mapsto \sigma_Y^{(e)} = \text{SD}\left((Y^{(e),[r]})_{r \in [N]}\right).$$

The confidence interval at episode e is given as follow

$$\left[\bar{Y}^{(e)} - c \frac{\sigma_Y^{(e)}}{\sqrt{N}}, \bar{Y}^{(e)} + c \frac{\sigma_Y^{(e)}}{\sqrt{N}} \right],$$

where $c \approx 2.045$ is the 97.5th percentile of a student's t -distribution of degree of freedom $N - 1 = 29$.

1.2 Parameters

We recall the parameters used in the simulation results.

1.2.1 Permutation sampling

In Sec. 5.1, we have considered $n = 12$ machines and agents, a constant learning rate $\alpha = 0.01$ and a total number of $15 \cdot 10^4$ episodes. For the first case, the target order τ^* was drawn randomly, whereas in the second case, $\nu_{t,i}$ was drawn from a normal distribution with zero mean and unit variance *i.e.*, $\mathcal{N}(0, 1)$.

1.2.2 Permutation and action sampling

In Sec. 5.2, we have considered $n = 9$ machines and agents, each machine has $k_i \in \llbracket 1; 7 \rrbracket$ action(s). Two learning rates were used: write α_β the learning rate used to update the permutation credit β and α_θ the one used to update the action credit θ . We considered the following: $\alpha_\beta = \frac{\alpha_\theta}{n}$. In all simulations, we considered a constant learning rate, with $\alpha_\theta = 0.01$. For all $(t, i) \in [n]^2$ and $a \in [k_i]$, $\nu_{t,i,a}$ is sampled from a normal distribution $\mathcal{N}(\mu_{t,i}, 1)$ with mean $\mu_{t,i}$ and unit variance, where $\mu_{t,i}$ is sampled from a uniform distribution $\mathcal{U}[-1, 1]$ for all $(t, i) \in [n]^2$.

2 Proofs

2.1 Proof of Theorem 1

Let $i \in [n]$. Recall that

$$\pi \mapsto \mathbb{E}_\pi[R^{(1)}] = \sum_{\tau \in \mathfrak{S}(n)} \pi(\sigma^{(1)} = \tau, A^{(1)} = \mathbf{a}) \lambda_{\tau, \mathbf{a}},$$

where $\lambda_{\tau, \mathbf{a}} = \mathbb{E}[R^{(1)} | \sigma^{(1)} = \tau, A^{(1)} = \mathbf{a}]$.

Therefore, for any $u \in \{1, 2\}$ and $v \in \{A, B\}$,

$$\begin{aligned} \frac{\partial \mathbb{E}_{\beta, \theta}^{u, v}[R^{(1)}]}{\partial \beta_i} &= \frac{\partial}{\partial \beta_i} \left(\sum_{\tau \in \mathfrak{S}(n)} \pi_\beta^u(\sigma^{(1)} = \tau) \pi_\theta^v(A^{(1)} = \mathbf{a} | \sigma^{(1)} = \tau) \lambda_{\tau, \mathbf{a}} \right), \\ &= \sum_{\tau \in \mathfrak{S}(n)} \frac{\partial \pi_\beta^u(\sigma^{(1)} = \tau)}{\partial \beta_i} \mathbb{E}_\theta^v[R^{(1)} | \sigma^{(1)} = \tau]. \end{aligned}$$

Model 1. Recall that

$$\begin{aligned} \pi_\beta^1(\sigma^{(1)} = \tau) &= \prod_{t \in [n]} \pi_\beta^1(\sigma_t^{(1)} = \tau_t | \sigma_{1:(t-1)}^{(1)} = \tau_{1:(t-1)}), \\ &= \prod_{t \in [n]} \Phi_{\tau_{t:n}}(\beta, \tau_t). \end{aligned}$$

So that

$$\begin{aligned} \frac{\partial \pi_\beta^1(\sigma^{(1)} = \tau)}{\partial \beta_i} &= \frac{\partial}{\partial \beta_i} \left(\prod_{t \in [n]} \Phi_{\tau_{t:n}}(\beta, \tau_t) \right), \\ &= \pi_\beta^1(\sigma^{(1)} = \tau) \left[\sum_{t \in [n]} \frac{1}{\Phi_{\tau_{t:n}}(\beta, \tau_t)} \underbrace{\frac{\partial \Phi_{\tau_{t:n}}(\beta, \tau_t)}{\partial \beta_i}}_{(*)} \right]. \end{aligned} \tag{S.1}$$

Since

$$\begin{aligned} (*) &= \frac{\partial \Phi_{\tau_{t:n}}(\beta, \tau_t)}{\partial \beta_i}, \\ &= \mathbf{1}_{t \leq \tau^{-1}(i)} \frac{\partial}{\partial \beta_i} \left(\frac{e^{\beta \tau_t}}{\sum_{j \in \tau_{t:n}} e^{\beta_j}} \right), \\ &= \mathbf{1}_{t \leq \tau^{-1}(i)} \frac{\mathbf{1}_{\tau_t = i} e^{\beta \tau_t}}{\sum_{j \in \tau_{t:n}} e^{\beta_j}} - \mathbf{1}_{t \leq \tau^{-1}(i)} \frac{e^{\beta \tau_t}}{\sum_{j \in \tau_{t:n}} e^{\beta_j}} \frac{e^{\beta_i}}{\sum_{j \in \tau_{t:n}} e^{\beta_j}}, \\ &= \mathbf{1}_{t \leq \tau^{-1}(i)} \Phi_{\tau_{t:n}}(\beta, \tau_t) (\mathbf{1}_{\tau_t = i} - \Phi_{\tau_{t:n}}(\beta, i)). \end{aligned}$$

Eq. (S.1) becomes

$$\begin{aligned} \frac{\partial \pi_{\beta}^1(\sigma^{(1)} = \tau)}{\partial \beta_i} &= \pi_{\beta}^1(\sigma^{(1)} = \tau) \left(\sum_{t \in [n]} \mathbb{1}_{t \leq \tau^{-1}(i)} (\mathbb{1}_{\tau_t = i} - \Phi_{\tau_{t:n}}(\beta, i)) \right), \\ &= \pi_{\beta}^1(\sigma^{(1)} = \tau) \left(1 - \sum_{t=1}^{\tau^{-1}(i)} \Phi_{\tau_{t:n}}(\beta, i) \right). \end{aligned}$$

Therefore, for any $v \in \{A, B\}$, the gradient of the objective function is

$$\begin{aligned} \frac{\partial \mathbb{E}_{\beta, \theta}^{1,v}[R^{(1)}]}{\partial \beta_i} &= \sum_{\tau \in \mathfrak{S}(n)} \pi_{\beta}^1(\sigma^{(1)} = \tau) \left(1 - \sum_{t=1}^{\tau^{-1}(i)} \Phi_{\tau_{t:n}}(\beta, \tau_t) \right) \mathbb{E}_{\theta}^v[R^{(1)} | \sigma^{(1)} = \tau], \\ &= \mathbb{E}_{\beta, \theta}^{1,v} \left[\left(1 - \sum_{t=1}^{\sigma^{(1)^{-1}(i)}(\sigma^{(1)})} \Phi_{\sigma_{t:n}^{(1)}}(\beta, i) \right) (R^{(1)} - B^{(1)}) \right], \end{aligned}$$

for any random variable $B^{(1)}$ that is independent of $\sigma^{(1)}$ under π_{β}^1 .

Model 2. In this case,

$$\begin{aligned} \pi_{\beta}^2(\sigma^{(1)} = \tau) &= \prod_{t \in [n]} \pi_{\beta}^2(\sigma_t^{(1)} = \tau_t | \sigma_{1:(t-1)}^{(1)} = \tau_{1:(t-1)}), \\ &= \prod_{t \in [n]} \Phi_{\tau_{t:n}}(\beta_t, \tau_t). \end{aligned}$$

Hence,

$$\begin{aligned} \frac{\partial \pi_{\beta}^1(\sigma^{(1)} = \tau)}{\partial \beta_{t,i}} &= \frac{\partial}{\partial \beta_{t,i}} \left(\prod_{l \in [n]} \Phi_{\tau_{l:n}}(\beta_l, \tau_l) \right), \\ &= \prod_{\substack{l \in [n] \\ l \neq t}} \Phi_{\tau_{l:n}}(\beta_l, \tau_l) \underbrace{\frac{\partial}{\partial \beta_{t,i}} \Phi_{\tau_{t:n}}(\beta_t, \tau_t)}_{(*)}. \end{aligned} \tag{S.2}$$

Since

$$\begin{aligned} (*) &= \frac{\partial}{\partial \beta_{t,i}} \Phi_{\tau_{t:n}}(\beta_t, \tau_t), \\ &= \frac{\partial}{\partial \beta_{t,i}} \left(\mathbb{1}_{i \in \tau_{t:n}} \frac{e^{\beta_t, \tau_t}}{\sum_{j \in \tau_{t:n}} e^{\beta_t, j}} \right), \\ &= \mathbb{1}_{i \in \tau_{t:n}} \left[\frac{\mathbb{1}_{i=\tau_t} e^{\beta_t, \tau_t}}{\sum_{j \in \tau_{t:n}} e^{\beta_t, j}} - \frac{e^{\beta_t, \tau_t} e^{\beta_t, i}}{\left(\sum_{j \in \tau_{t:n}} e^{\beta_t, j} \right)^2} \right], \\ &= \Phi_{\tau_{t:n}}(\beta_t, \tau_t) (\mathbb{1}_{i=\tau_t} - \Phi_{\tau_{t:n}}(\beta_t, i)). \end{aligned}$$

Eq. (S.2) becomes

$$\frac{\partial \pi_{\beta}^2(\sigma^{(1)} = \tau)}{\partial \beta_{t,i}} = \pi_{\beta}^2(\sigma^{(1)} = \tau) (\mathbb{1}_{i=\tau_t} - \Phi_{\tau_{t:n}}(\beta_t, i)).$$

Therefore, for any $v \in \{A, B\}$, the gradient of the objective function is

$$\begin{aligned} \frac{\partial \mathbb{E}_{\beta, \theta}^{2,v}[R^{(1)}]}{\partial \beta_{t,i}} &= \sum_{\tau \in \mathfrak{S}(n)} \pi_{\beta}^2(\sigma^{(1)} = \tau) (\mathbb{1}_{i=\tau_t} - \Phi_{\tau_{t:n}}(\beta_t, i)) \mathbb{E}_{\theta}^v[R^{(1)} | \sigma^{(1)} = \tau], \\ &= \mathbb{E}_{\beta, \theta}^{2,v} \left[(\mathbb{1}_{i=\sigma_t^{(1)}} - \Phi_{\sigma_{t:n}^{(1)}}(\beta_t, i)) (R^{(1)} - B^{(1)}) \right], \end{aligned}$$

for any random variable $B^{(1)}$ that is independent of $\sigma^{(1)}$ under π_{β}^2 .

2.2 Proof of Theorem. 2

For any $u \in \{1, 2\}$ and $v \in \{A, B\}$,

$$\begin{aligned} \frac{\partial \mathbb{E}_{\beta, \theta}^{u, v}[R^{(1)}]}{\partial \theta_x} &= \frac{\partial}{\partial \theta_x} \left(\sum_{\substack{\tau \in \mathfrak{S}(n) \\ \mathbf{a} \in \mathcal{A}}} \pi_{\beta}^u(\sigma^{(1)} = \tau) \pi_{\theta}^v(A^{(1)} = \mathbf{a} | \sigma^{(1)} = \tau) \lambda_{\tau, \mathbf{a}} \right), \\ &= \sum_{\substack{\tau \in \mathfrak{S}(n) \\ \mathbf{a} \in \mathcal{A}}} \pi_{\beta}^u(\sigma^{(1)} = \tau) \frac{\partial \pi_{\theta}^v(A^{(1)} = \mathbf{a} | \sigma^{(1)} = \tau)}{\partial \theta_x} \lambda_{\tau, \mathbf{a}}. \end{aligned}$$

Model A. In this case, for any $i \in [n]$ and $a \in [k_i]$,

$$\begin{aligned} \frac{\partial \pi_{\theta}^A(A^{(1)} = \mathbf{a} | \sigma^{(1)} = \tau)}{\partial \theta_{i, a}} &= \frac{\partial}{\partial \theta_{i, a}} \left(\prod_{j \in [n]} \pi_{\theta}^A(A_j^{(1)} = a_j | \sigma^{(1)} = \tau) \right), \\ &= \left(\prod_{\substack{j \in [n] \\ j \neq i}} \pi_{\theta}^A(A_j^{(1)} = a_j | \sigma^{(1)} = \tau) \right) \underbrace{\frac{\partial}{\partial \theta_{i, a}} \left(\frac{e^{\theta_{i, a_i}}}{\sum_{b \in [k_i]} e^{\theta_{i, b}}} \right)}_{(*)}. \end{aligned}$$

Since

$$\begin{aligned} (*) &= \frac{\mathbb{1}_{a_i=a} e^{\theta_{i, a_i}}}{\sum_{b \in [k_i]} e^{\theta_{i, b}}} - \frac{e^{\theta_{i, a_i}} e^{\theta_{i, a}}}{\left(\sum_{b \in [k_i]} e^{\theta_{i, b}} \right)^2}, \\ &= \mathbb{1}_{a_i=a} \Phi(\theta_i, a_i) - \Phi(\theta_i, a_i) \Phi(\theta_i, a), \\ &= \pi_{\theta}^A(A_i^{(1)} = a_i | \sigma^{(1)} = \tau) (\mathbb{1}_{a_i=a} - \Phi(\theta_i, a)). \end{aligned}$$

Thus,

$$\frac{\partial \pi_{\theta}^A(A^{(1)} = \mathbf{a} | \sigma^{(1)} = \tau)}{\partial \theta_{i, a}} = \pi_{\theta}^A(A^{(1)} = \mathbf{a} | \sigma^{(1)} = \tau) (\mathbb{1}_{a_i=a} - \Phi(\theta_i, a)).$$

Therefore, the gradient of the objective function is

$$\begin{aligned} \frac{\partial \mathbb{E}_{\beta, \theta}^{u, A}[R^{(1)}]}{\partial \theta_{i, a}} &= \sum_{\substack{\tau \in \mathfrak{S}(n) \\ \mathbf{a} \in \mathcal{A}}} (\mathbb{1}_{a_i=a} - \Phi(\theta_i, a)) \pi_{\beta, \theta}^{u, A}(\sigma^{(1)} = \tau, A^{(1)} = \mathbf{a}) \lambda_{\tau, \mathbf{a}}, \\ &= \mathbb{E}_{\beta, \theta}^{u, A} \left[(\mathbb{1}_{A_i^{(1)}=a} - \Phi(\theta_i, a)) (R^{(1)} - B^{(1)}) \right], \end{aligned}$$

for any random variable $B^{(1)}$ that is independent of $(\sigma^{(1)}, A^{(1)})$ under $\pi_{\beta, \theta}^{u, A}$.

Model B. In this case, for any $i, t \in [n]$ and $a \in [k_i]$,

$$\begin{aligned} \frac{\partial \pi_{\theta}^B(A^{(1)} = \mathbf{a} | \sigma^{(1)} = \tau)}{\partial \theta_{i, t, a}} &= \frac{\partial}{\partial \theta_{i, t, a}} \left(\prod_{j \in [n]} \pi_{\theta}^B(A_j^{(1)} = a_j | \sigma^{(1)} = \tau) \right), \\ &= \frac{\partial}{\partial \theta_{i, t, a}} \left(\prod_{s \in [n]} \Phi(\theta_{\tau_s, s}, a_{\tau_s}) \right), \\ &= \left(\prod_{\substack{s \in [n] \\ s \neq t}} \pi_{\theta}^B(A_{\tau_s}^{(1)} = a_{\tau_s} | \sigma^{(1)} = \tau) \right) \mathbb{1}_{\tau_t=i} \underbrace{\frac{\partial}{\partial \theta_{i, t, a}} \left(\frac{e^{\theta_{i, t, a_i}}}{\sum_{b \in [k_i]} e^{\theta_{i, t, b}}} \right)}_{(*)}. \end{aligned}$$

Since

$$\begin{aligned}
 (*) &= \frac{\mathbb{1}_{a_i=a} e^{\theta_{i,t}, a_i}}{\sum_{b \in [k_i]} e^{\theta_{i,t}, b}} - \frac{e^{\theta_{i,t}, a_i} e^{\theta_{i,t}, a}}{\left(\sum_{b \in [k_i]} e^{\theta_{i,t}, b} \right)^2}, \\
 &= \mathbb{1}_{a_i=a} \Phi(\theta_{i,t}, a_i) - \Phi(\theta_{i,t}, a_i) \Phi(\theta_{i,t}, a), \\
 &= \pi_{\theta}^B(A_i^{(1)} = a_i | \sigma_t^{(1)} = i) (\mathbb{1}_{a_i=a} - \Phi(\theta_{i,t}, a)).
 \end{aligned}$$

Thus,

$$\frac{\partial \pi_{\theta}^B(A^{(1)} = \mathbf{a} | \sigma^{(1)} = \tau)}{\partial \theta_{i,t}, a} = \pi_{\theta}^B(A^{(1)} = \mathbf{a} | \sigma^{(1)} = \tau) \mathbb{1}_{\tau_t=i} (\mathbb{1}_{a_i=a} - \Phi(\theta_{i,t}, a)).$$

Therefore, the gradient of the objective function is

$$\begin{aligned}
 \frac{\partial \mathbb{E}_{\beta, \theta}^{u, B}[R^{(1)}]}{\partial \theta_{i,t}, a} &= \sum_{\substack{\tau \in \mathfrak{S}(n) \\ \mathbf{a} \in \mathcal{A}}} \mathbb{1}_{\tau_t=i} (\mathbb{1}_{a_i=a} - \Phi(\theta_{i,t}, a)) \pi_{\beta, \theta}^{u, B}(\sigma^{(1)} = \tau, A^{(1)} = \mathbf{a}) \lambda_{\tau, \mathbf{a}}, \\
 &= \mathbb{E}_{\beta, \theta}^{u, B} \left[\mathbb{1}_{\sigma_t^{(1)}=i} (\mathbb{1}_{A_i^{(1)}=a} - \Phi(\theta_{i,t}, a)) (R^{(1)} - B^{(1)}) \right],
 \end{aligned}$$

for any random variable $B^{(1)}$ that is independent of $(\sigma^{(1)}, A^{(1)})$ under $\pi_{\beta, \theta}^{u, B}$.

3 Additional figures

3.1 Permutation sampling

Case 1

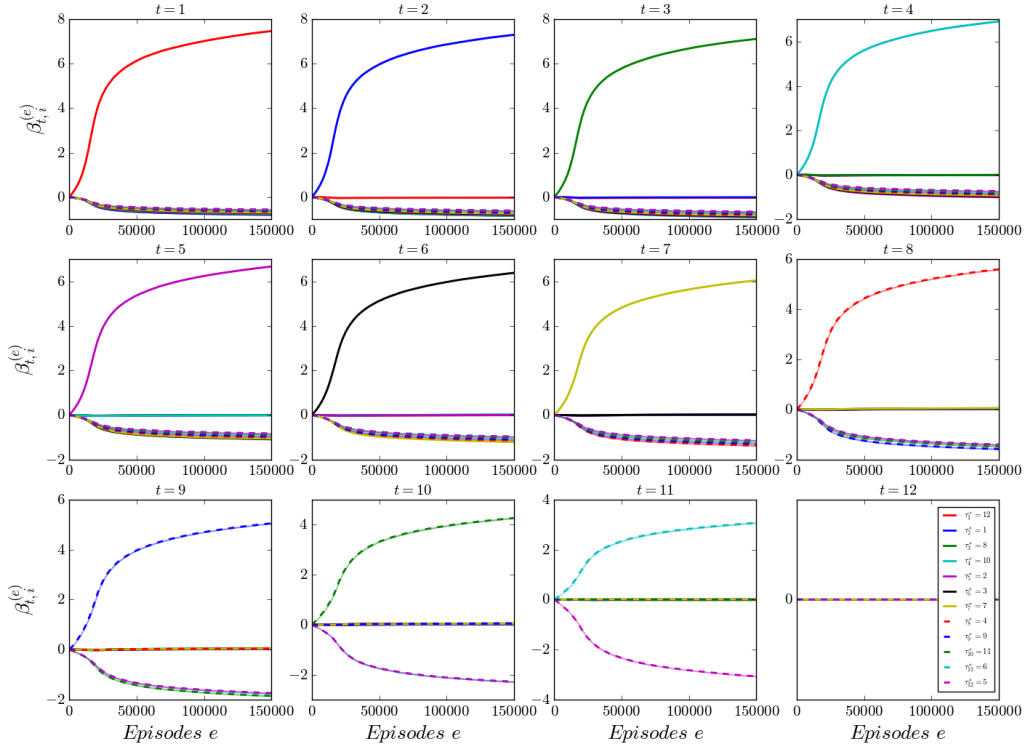
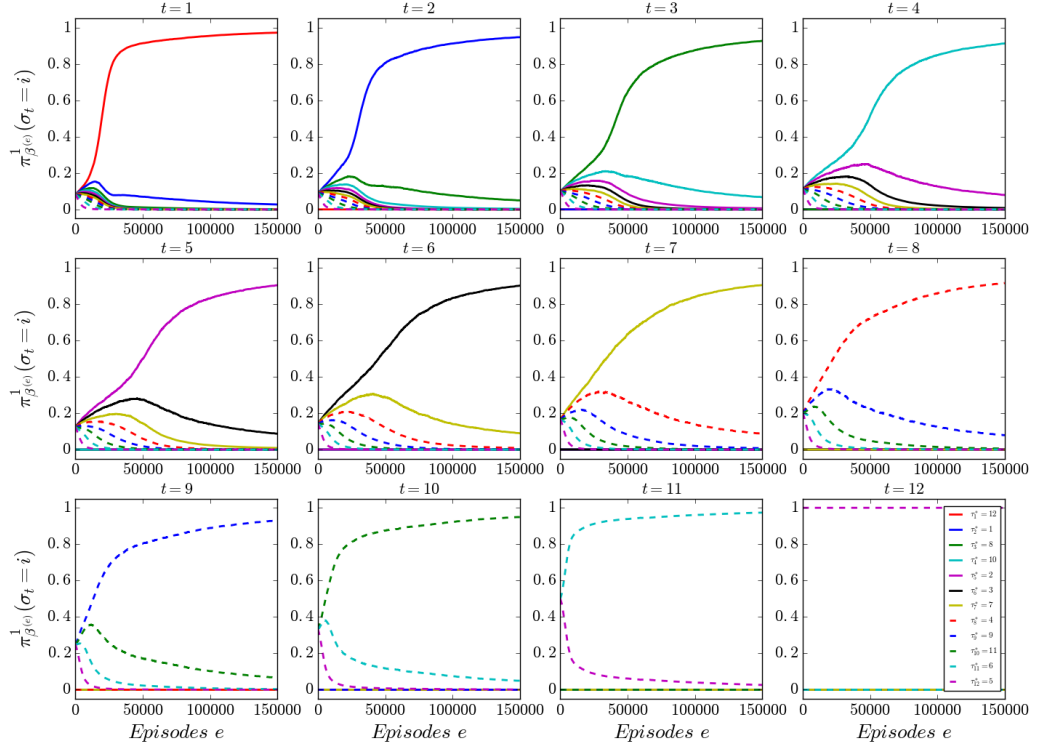
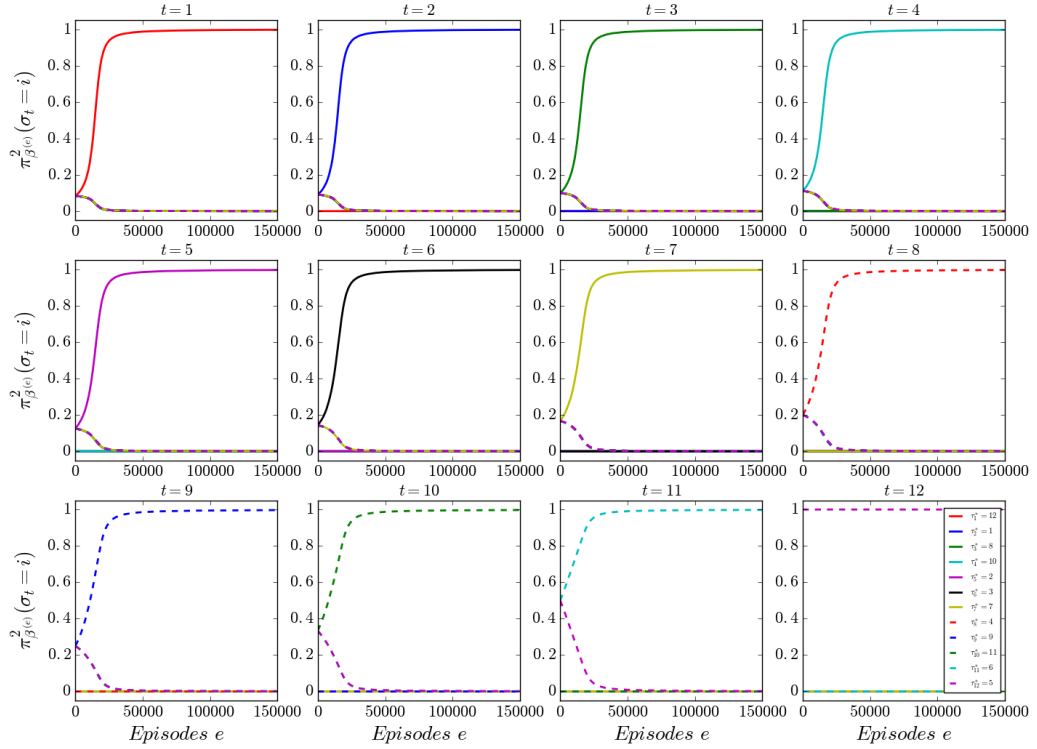


Figure S.1: The average parameter $(\beta)_i$ on model 2 for each agent t in Sec. 5.1.1. Each color/shape corresponds to a machine i .



(a)



(b)

Figure S.2: $\pi_{\beta(e)}^u(\sigma_t = \hat{\tau})$ for model 1 in (a) and model 2 in (b). $\hat{\tau}$ is obtained as in Eq.8 for each model.

Case 2

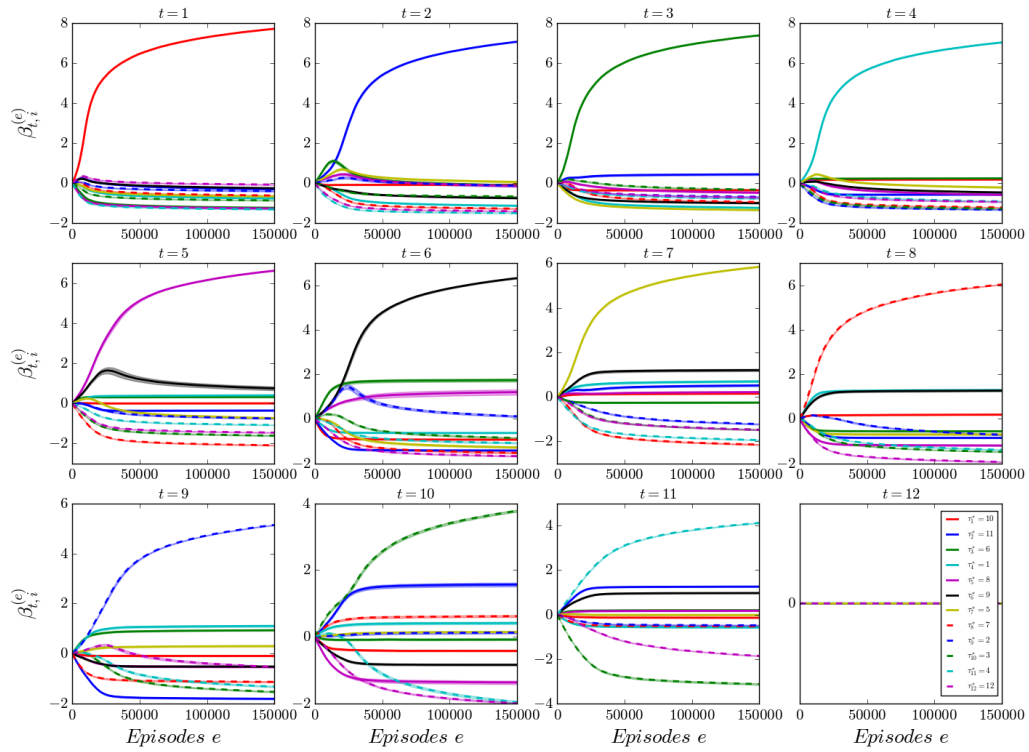
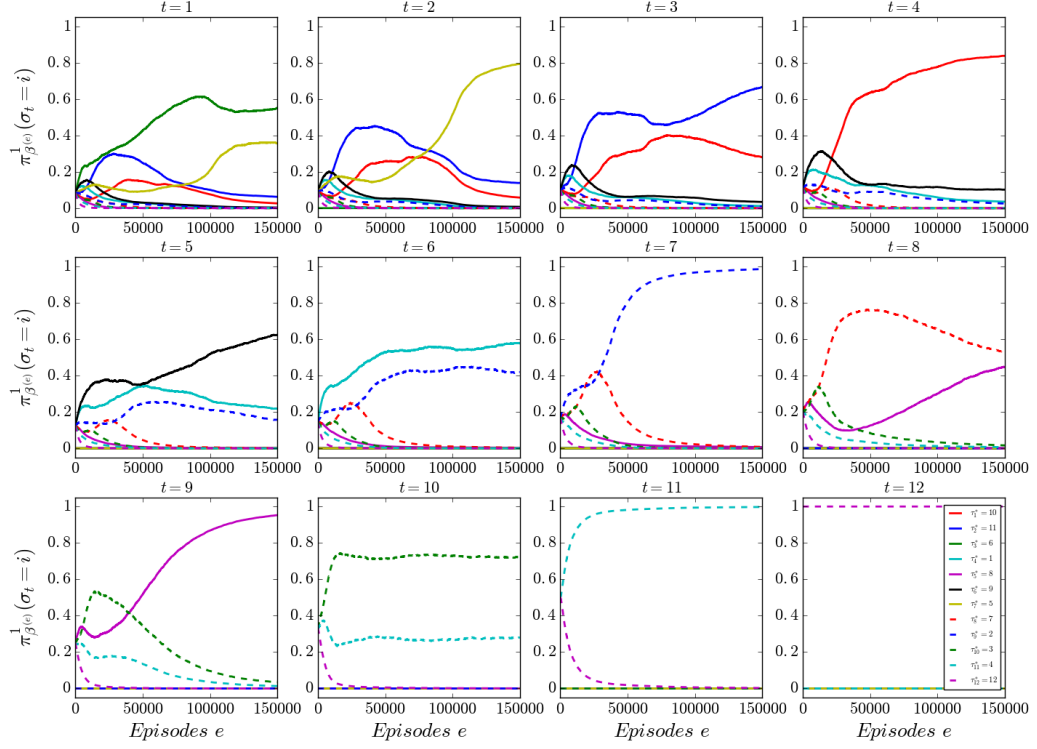
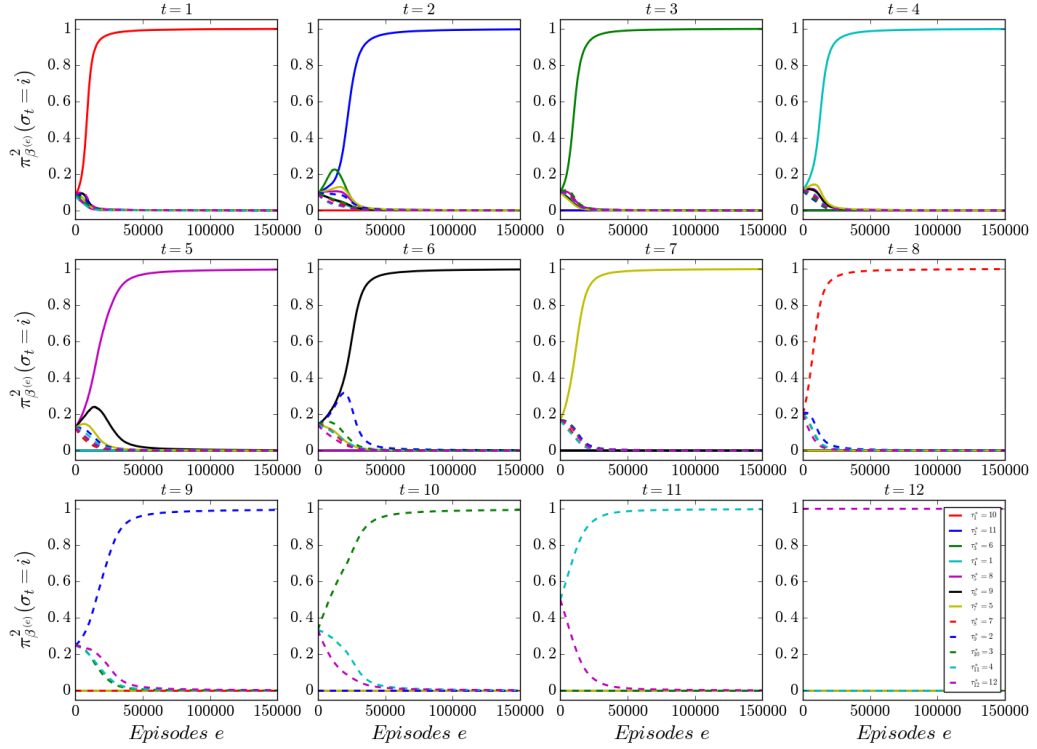


Figure S.3: The average parameter $(\beta)_i$ on model 2 for each agent t in Sec. 5.1.2. Each color/shape corresponds to a machine i .



(a)

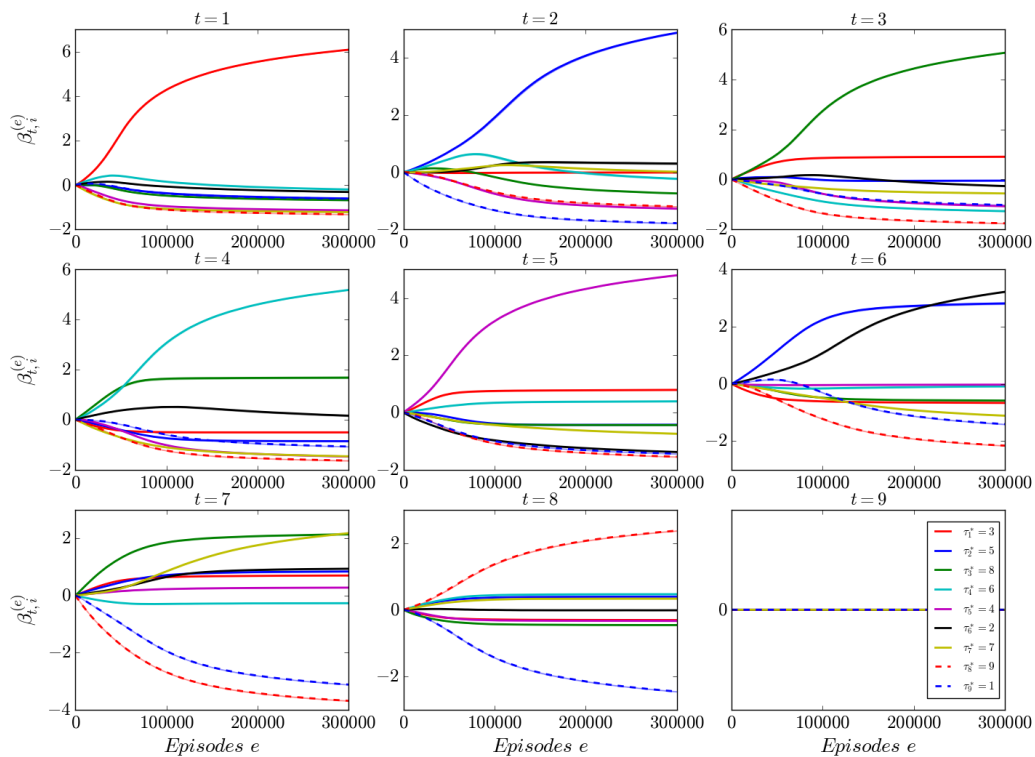


(b)

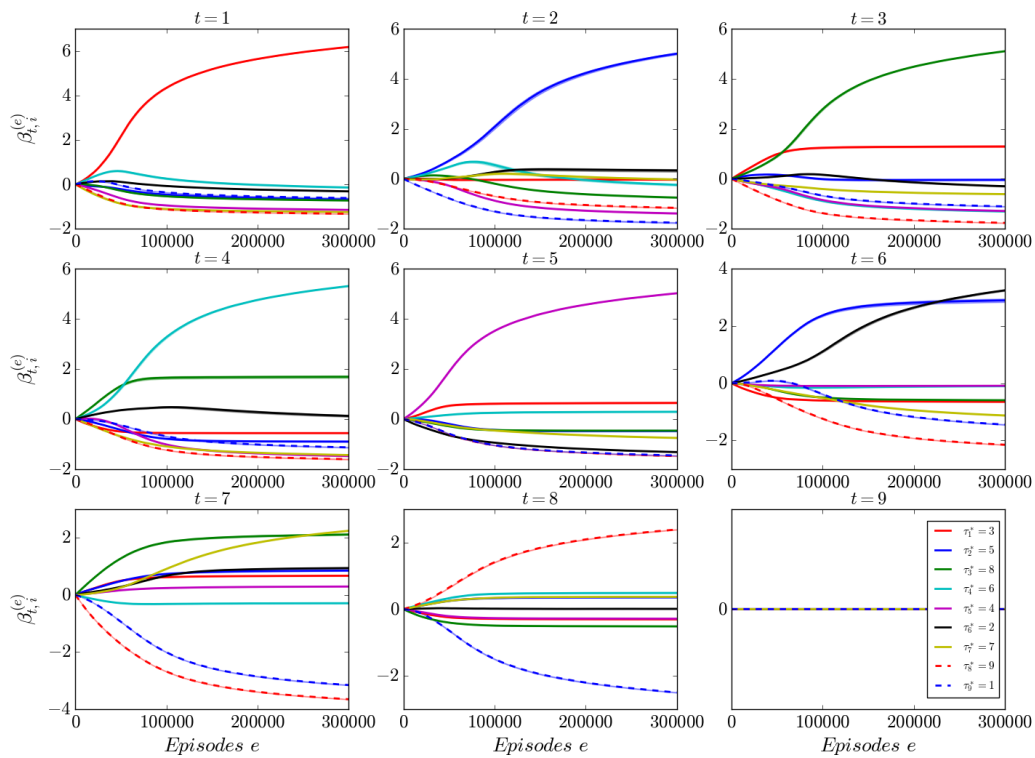
Figure S.4: $\pi_{\beta(e)}^u(\sigma_t = \hat{\tau})$ for model 1 in (a) and model 2 in (b). $\hat{\tau}$ is obtained as in Eq. 8 for each model.

3.2 Sampling permutation and actions

Case 1

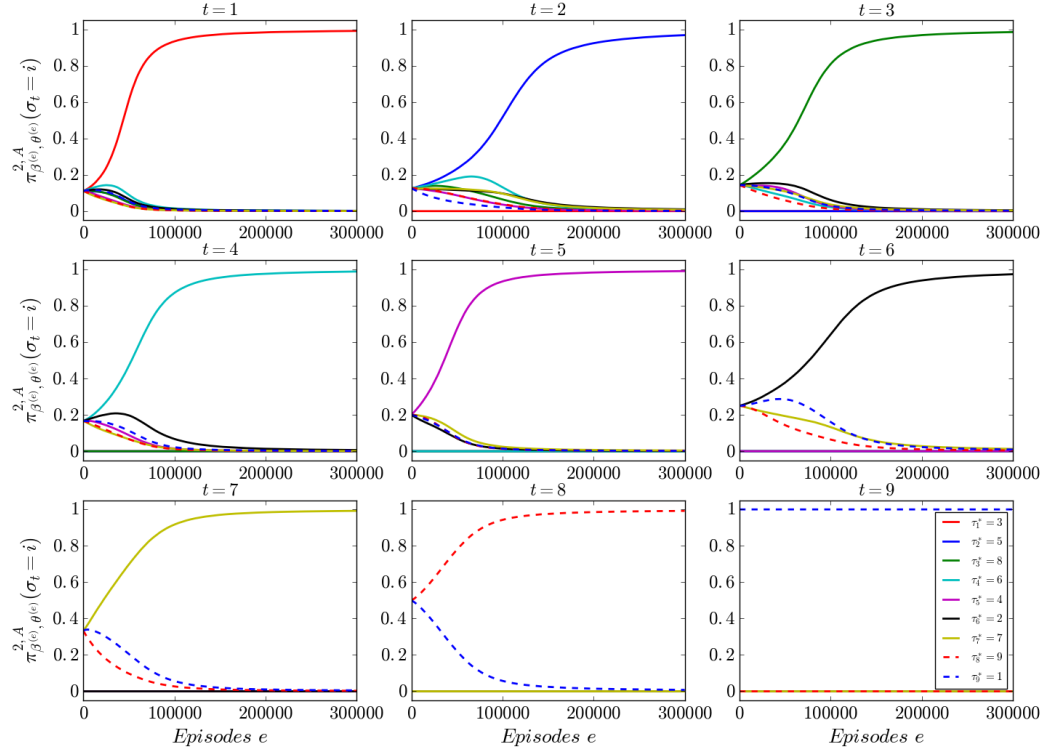


(a)

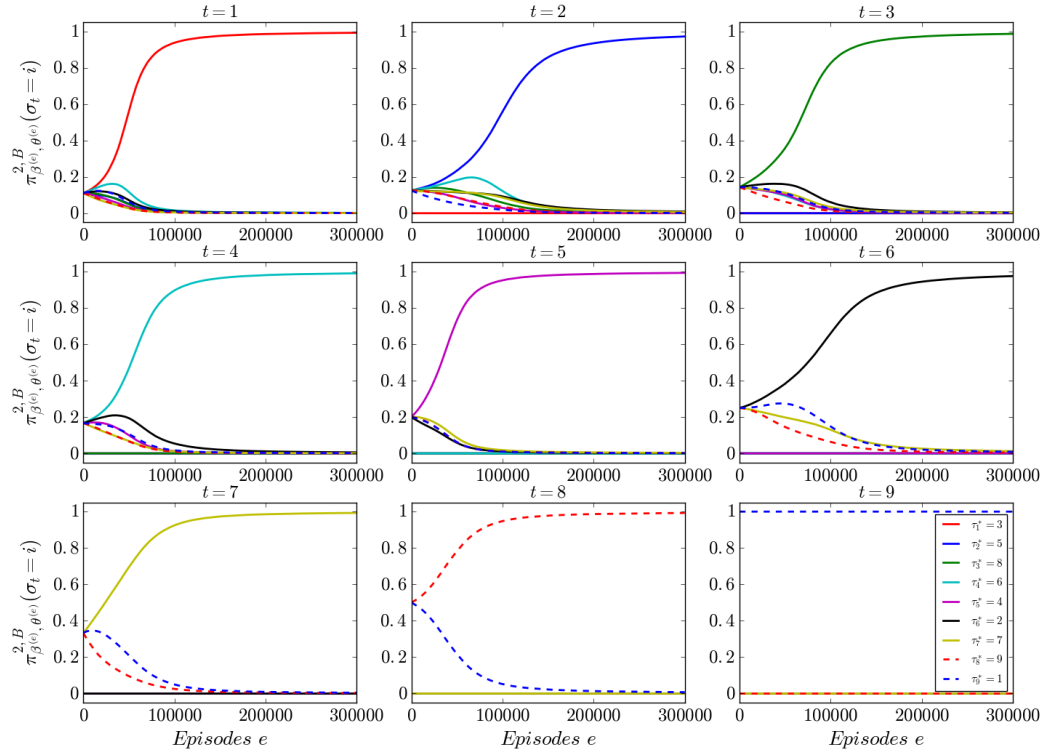


(b)

Figure S.5: The average parameter $(\beta)_i$ on model 2A in (a) and model 2B in (b) for each agent t in Sec. 5.2.1. Each color/shape corresponds to a machine i .

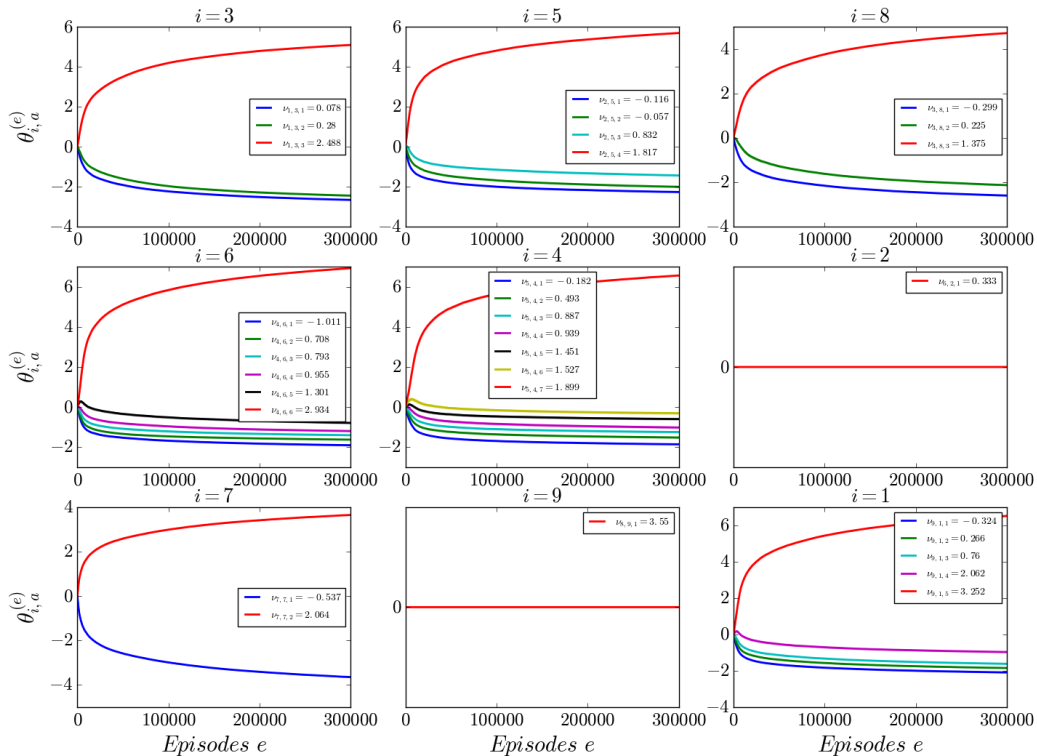


(a)

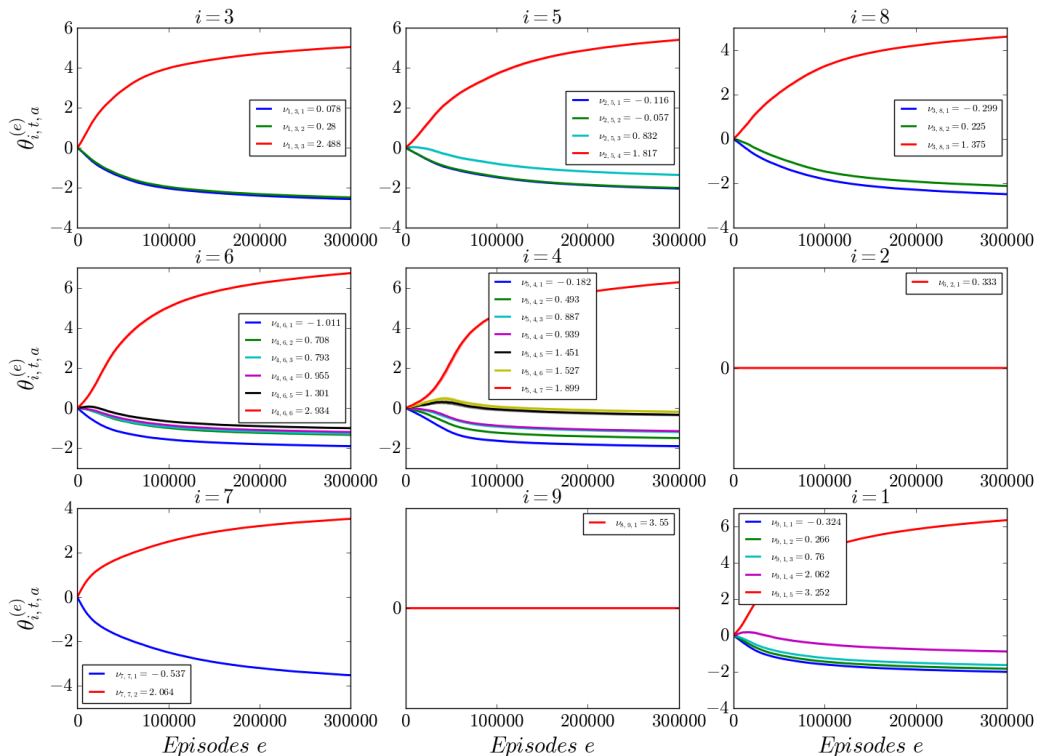


(b)

Figure S.6: $\pi_{\beta^{(e)}, \theta^{(e)}}^u(\sigma_t = \hat{\tau})$ for model 2A in (a) and model 2B in (b). $\hat{\tau}$ is obtained as in Eq. 8 for each model.

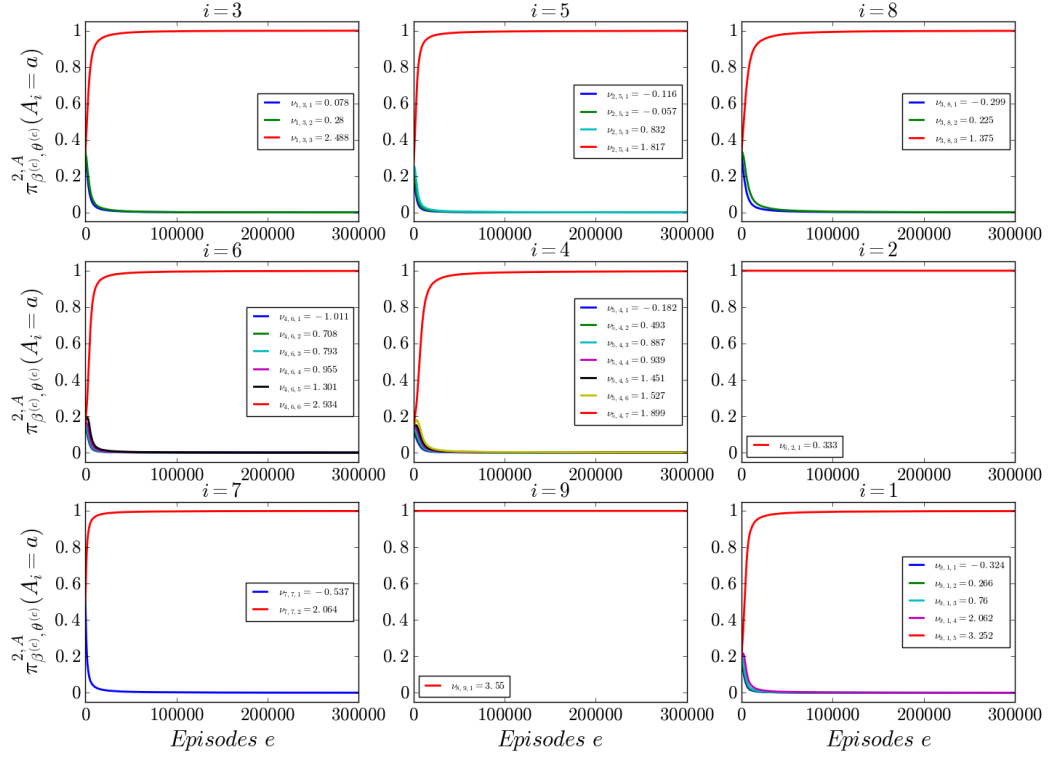


(a)

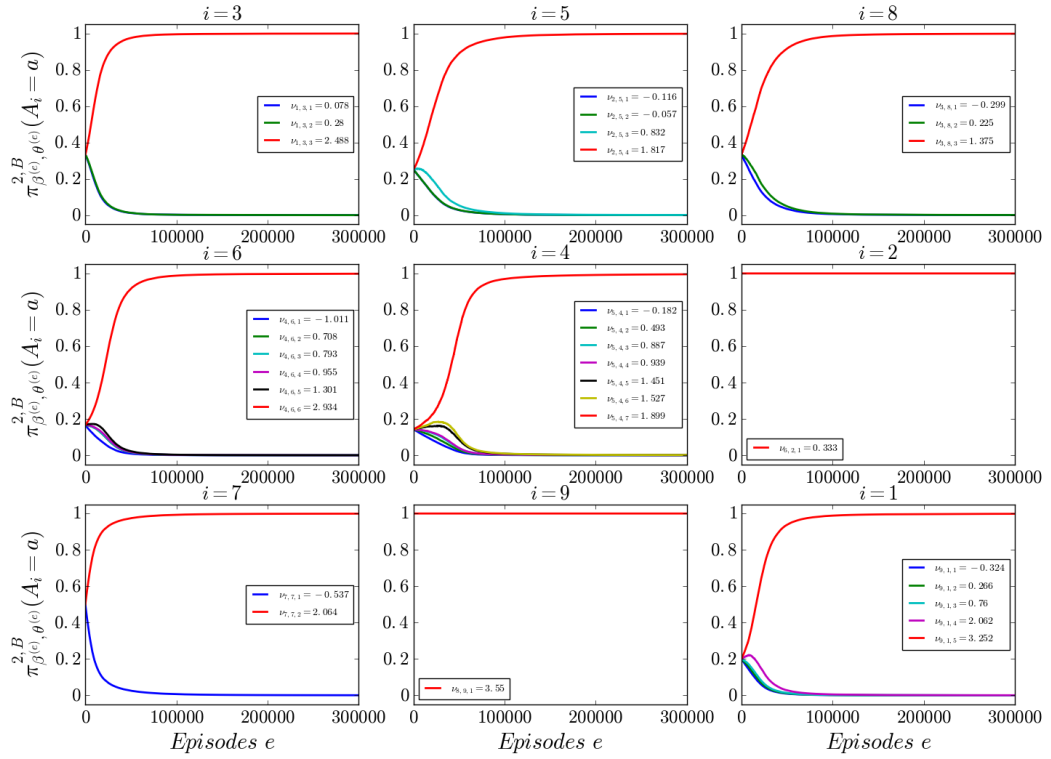


(b)

Figure S.7: The average parameter $(\theta)_i$ on model 2A in (a) and model 2B in (b) for each machine i in Sec. 5.2.1 (sorted according to the optimal allocation order τ^*). Each color corresponds to an action $a \in [k_i]$. The red curve represents the best action to be chosen.



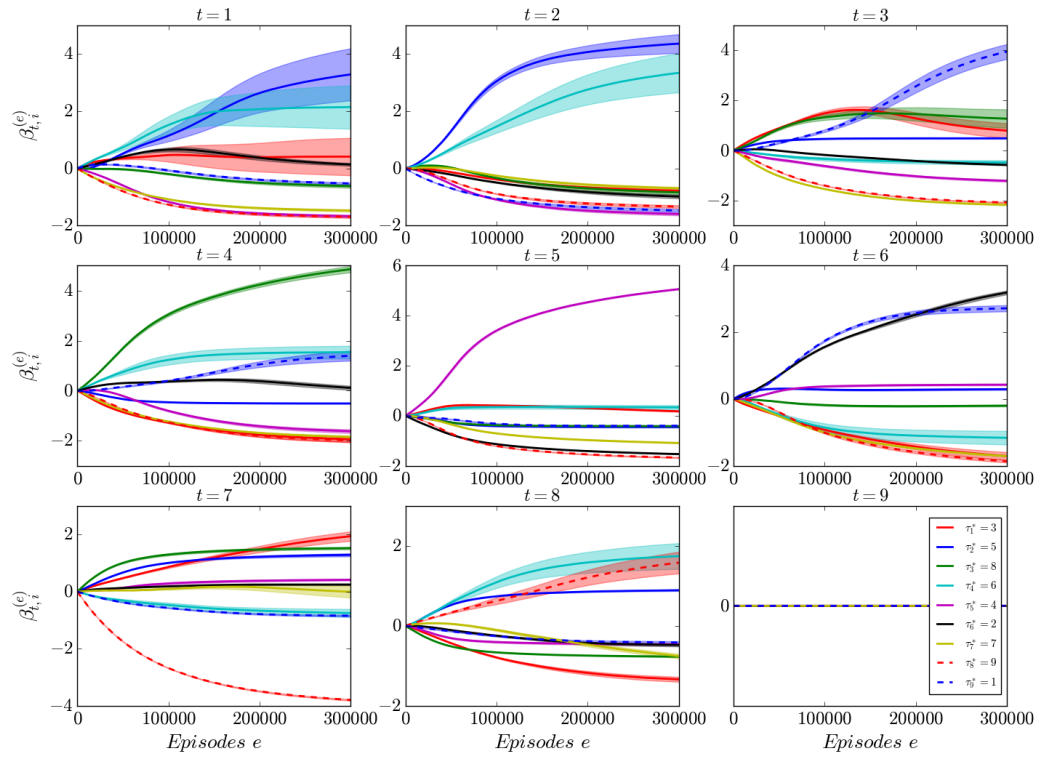
(a)



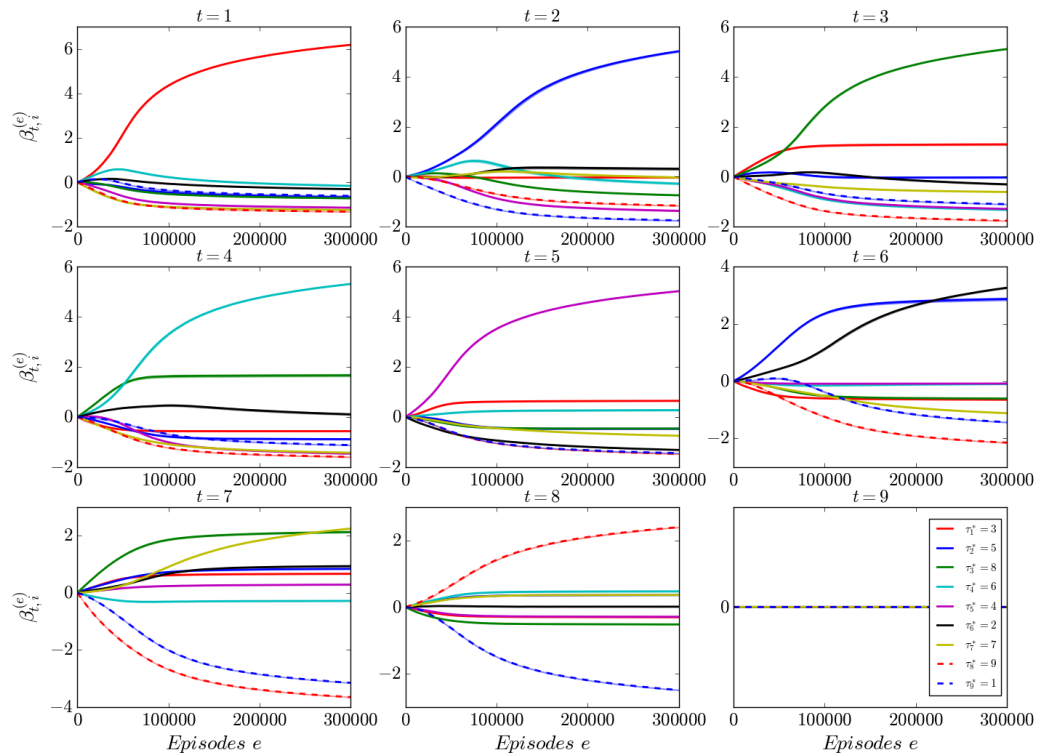
(b)

Figure S.8: $\pi_{\beta^{(e)}, \theta^{(e)}}^u(A_i = a)$ for model 2A in (a) and model 2B in (b).

Case 2

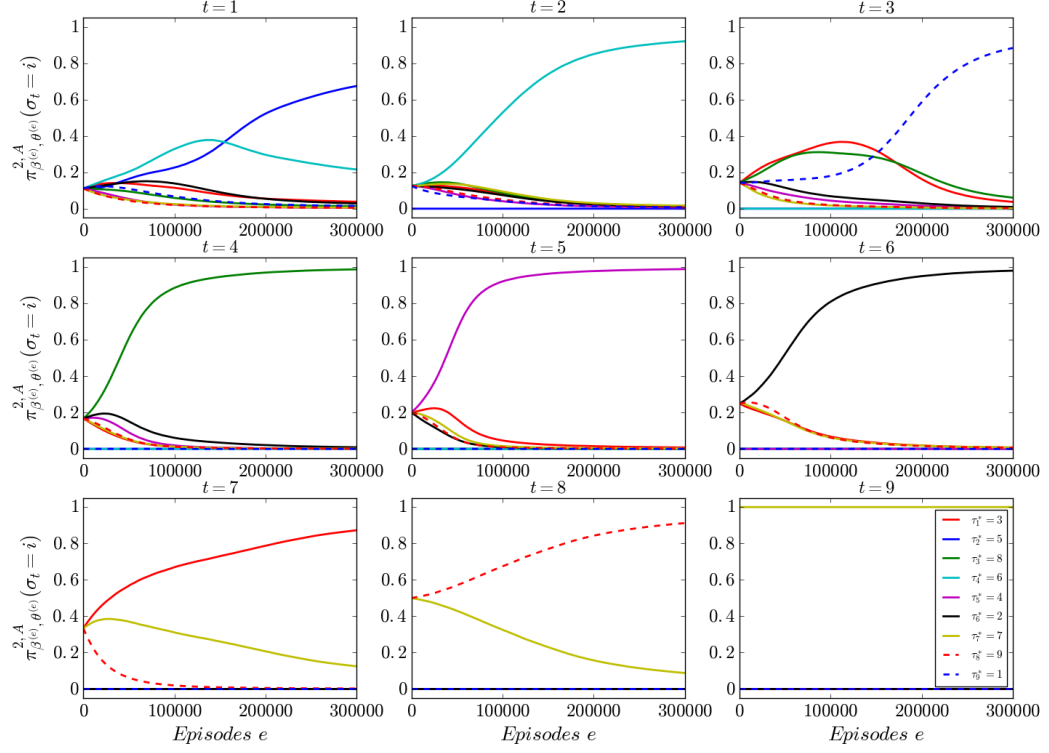


(a)

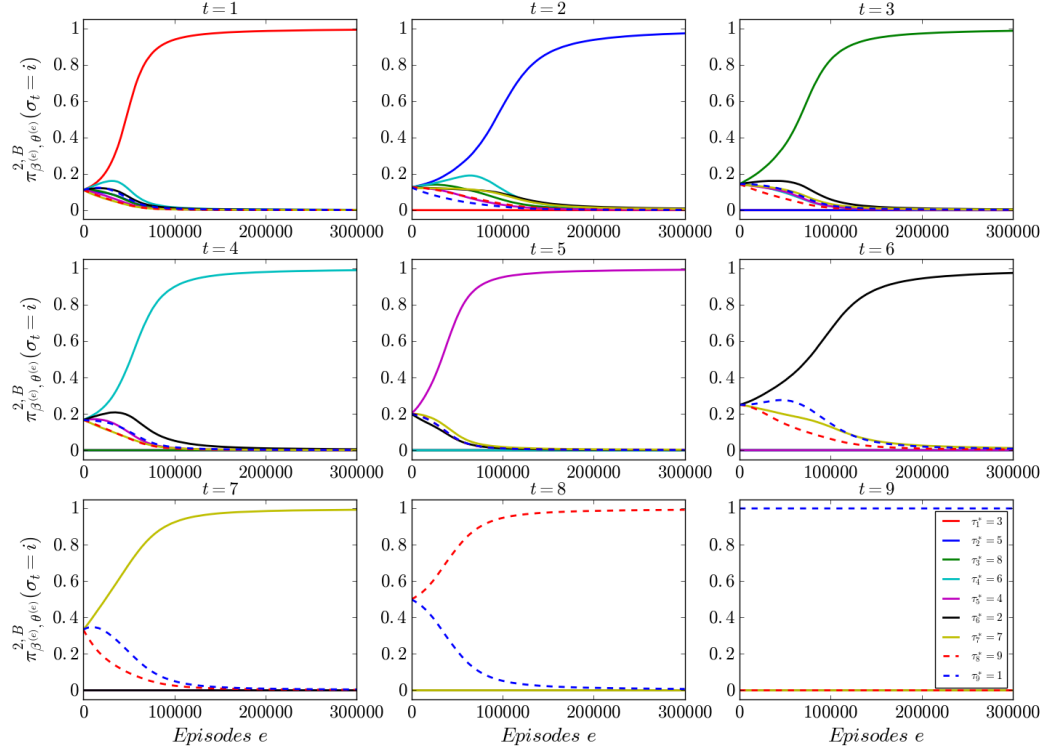


(b)

Figure S.9: The average parameter $(\beta)_i$ on model 2A in (a) and model 2B in (b) for each agent t in Sec. 5.2.2. Each color/shape corresponds to a machine i .

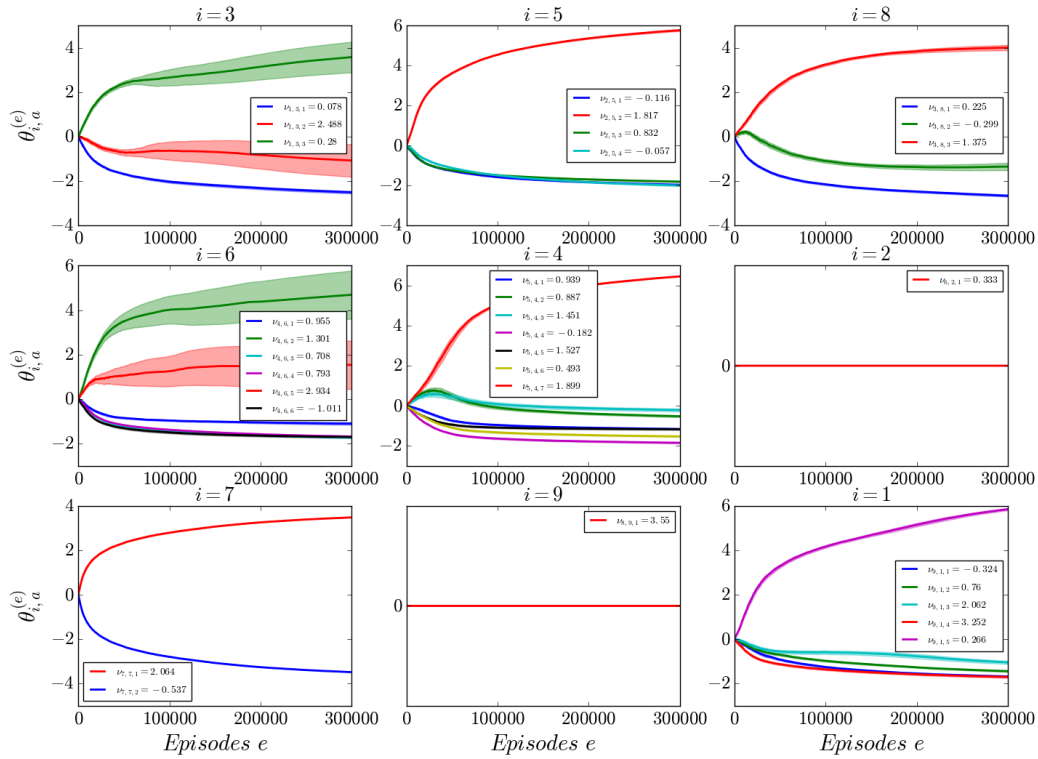


(a)

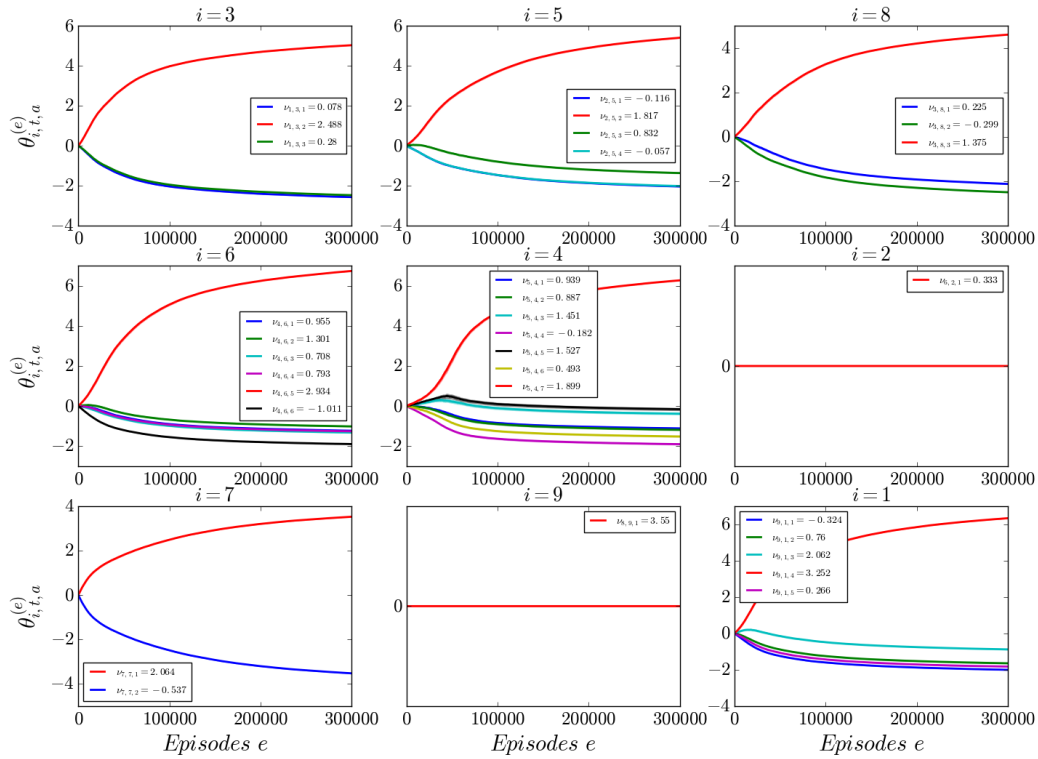


(b)

Figure S.10: $\pi_{\beta^{(\epsilon)}, \theta^{(\epsilon)}}^u(\sigma_t = \hat{\tau}_t)$ for model 2A in (a) and model 2B in (b). $\hat{\tau}$ is obtained as in Eq. 8 for each model.

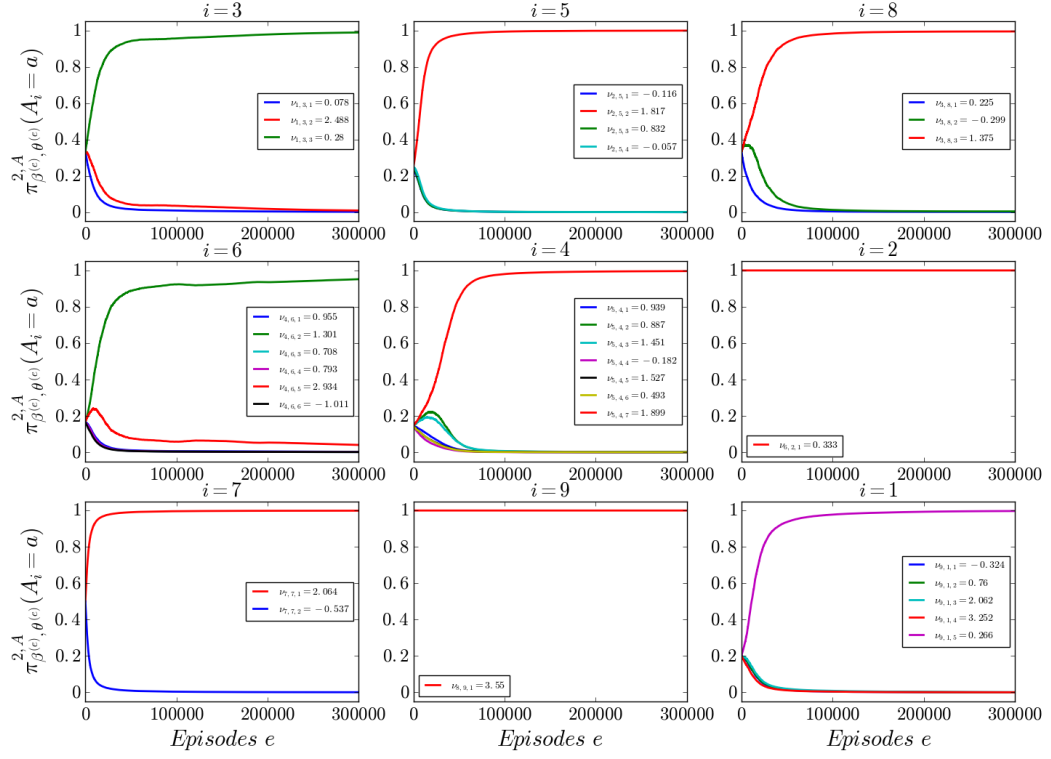


(a)

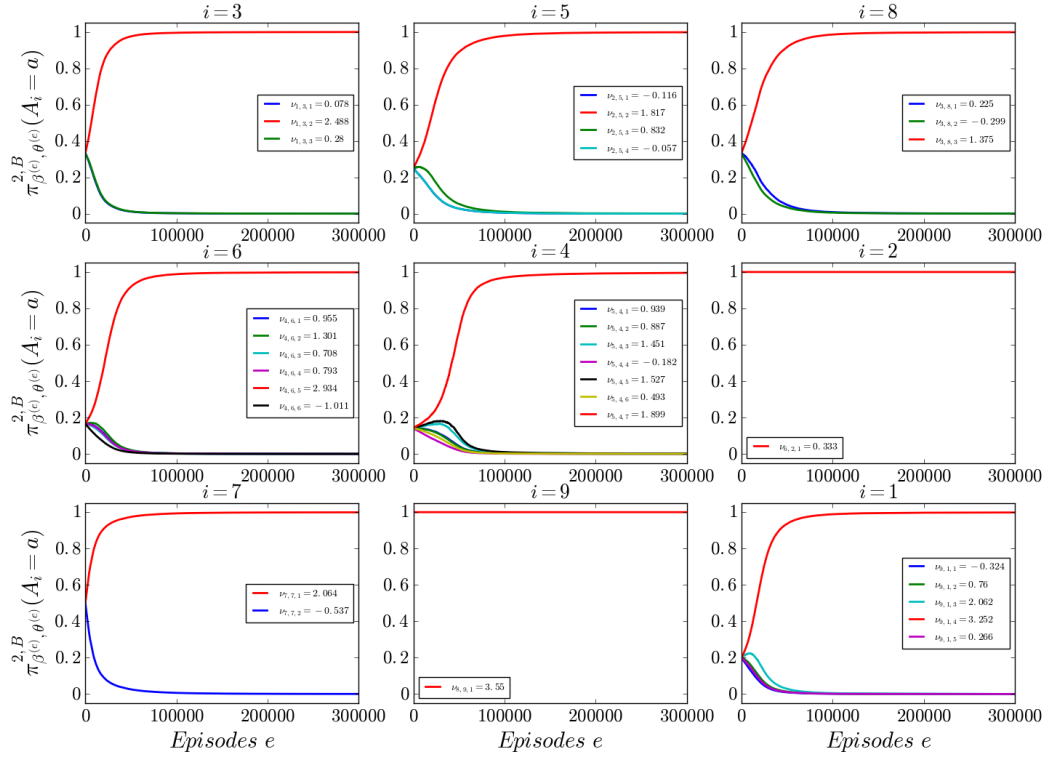


(b)

Figure S.11: The average parameter $(\theta)_i$ on model 2A in (a) and model 2B in (b) for each machine i in Sec. 5.2.2 (sorted according to the optimal allocation order τ^*). Each color corresponds to an action $a \in [k_i]$. The red curve represents the best action to be chosen.



(a)



(b)

Figure S.12: $\pi_{\beta^{(e)}, \theta^{(e)}}^u(A_i = a)$ for model 2A in (a) and model 2B in (b).