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 Lehéricy
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 $e \geqslant 1$, produces $Y^{(e),[r]}$ for any variable Y. Therefore, the mean and the standard deviation of Y over the realizations is given as

$$\forall e \geqslant 1 \longmapsto \quad \overline{Y}^{(e)} = \frac{1}{N}\sum_{r=1}^{N} Y^{(e),[r]} = \text{mean}\Big((Y^{(e),[r]})_{r\in[N]}\Big),$$

$$\forall e \geqslant 1 \longmapsto \quad \sigma_Y^{(e)} = \text{SD}\Big((Y^{(e),[r]})_{r\in[N]}\Big).$$

The confidence interval at episode e is given as follow

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

where $c \approx 2.045$ is the 97.5^{th} 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 $k_i \in [\![1; 7]\!]$ 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: $\alpha_\beta = \dfrac{\alpha_\theta}{n}$. 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 $\mathcal{N}(\mu_{i,t}, 1)$ 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

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

where $\lambda_{\tau,\mathbf{a}} = \mathbb{E}\big[R^{(1)} \,|\, \sigma^{(1)} = \tau, A^{(1)} = \mathbf{a}\big]$.
Therefore, for any $u \in \{1, 2\}$ and $v \in \{A, B\}$,

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

Model 1. Recall that

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

So that

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

Since

$$\begin{aligned}(\star) &= \frac{\partial \Phi_{\tau_{t:n}}(\beta, \tau_t)}{\partial \beta_i},\\ &= \mathbb{1}_{t \leqslant \tau^{-1}(i)} \frac{\partial}{\partial \beta_i}\left(\frac{e^{\beta_{\tau_t}}}{\sum\limits_{j \in \tau_{t:n}} e^{\beta_j}}\right),\\ &= \mathbb{1}_{t \leqslant \tau^{-1}(i)} \frac{\mathbb{1}_{\tau_t = i} e^{\beta_{\tau_t}}}{\sum\limits_{j \in \tau_{t:n}} e^{\beta_j}} - \mathbb{1}_{t \leqslant \tau^{-1}(i)} \frac{e^{\beta_{\tau_t}}}{\sum\limits_{j \in \tau_{t:n}} e^{\beta_j}} \frac{e^{\beta_i}}{\sum\limits_{j \in \tau_{t:n}} e^{\beta_j}},\\ &= \mathbb{1}_{t \leqslant \tau^{-1}(i)} \Phi_{\tau_{t:n}}(\beta, \tau_t)\big(\mathbb{1}_{\tau_t = i} - \Phi_{\tau_{t:n}}(\beta, i)\big).\end{aligned}$$

Eq. (S.1) becomes

$$\frac{\partial \pi^1_\beta(\sigma^{(1)}=\tau)}{\partial \beta_i} = \pi^1_\beta(\sigma^{(1)}=\tau)\Bigg(\sum_{t\in[n]} \mathbb{1}_{t\leqslant \tau^{-1}(i)}\big(\mathbb{1}_{\tau_t=i} - \Phi_{\tau_{t:n}}(\beta,i)\big)\Bigg),$$
$$= \pi^1_\beta(\sigma^{(1)}=\tau)\Bigg(1 - \sum_{t=1}^{\tau^{-1}(i)} \Phi_{\tau_{t:n}}(\beta,i)\Bigg).$$

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

$$\frac{\partial \mathbb{E}^{1,v}_{\beta,\theta}\big[R^{(1)}\big]}{\partial \beta_i} = \sum_{\tau\in\mathfrak{S}(n)} \pi^1_\beta(\sigma^{(1)}=\tau)\Bigg(1 - \sum_{t=1}^{\tau^{-1}(i)} \Phi_{\tau_{t:n}}(\beta,\tau_t)\Bigg)\mathbb{E}^v_\theta\big[R^{(1)}\,|\,\sigma^{(1)}=\tau\big],$$
$$= \mathbb{E}^{1,v}_{\beta,\theta}\Bigg[\Bigg(1 - \sum_{t=1}^{\sigma^{(1)^{-1}}(i)} \Phi_{\sigma^{(1)}_{t:n}}(\beta,i)\Bigg)\big(R^{(1)} - B^{(1)}\big)\Bigg],$$

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

Model 2. In this case,

$$\pi^2_\beta(\sigma^{(1)}=\tau) = \prod_{t\in[n]} \pi^2_\beta\big(\sigma^{(1)}_t = \tau_t\,|\,\sigma^{(1)}_{1:(t-1)} = \tau_{1:(t-1)}\big),$$
$$= \prod_{t\in[n]} \Phi_{\tau_{t:n}}(\beta_t,\tau_t).$$

Hence,

$$\frac{\partial \pi^1_\beta(\sigma^{(1)}=\tau)}{\partial \beta_{t,i}} = \frac{\partial}{\partial \beta_{t,i}}\Bigg(\prod_{l\in[n]} \Phi_{\tau_{l:n}}(\beta_l,\tau_l)\Bigg),$$
$$= \prod_{\substack{l\in[n]\\ l\neq t}} \Phi_{\tau_{l:n}}(\beta_l,\tau_l)\,\underbrace{\frac{\partial}{\partial \beta_{t,i}}\Phi_{\tau_{t:n}}(\beta_t,\tau_t)}_{(\star)}. \tag{S.2}$$

Since

$$(\star) = \frac{\partial}{\partial \beta_{t,i}}\Phi_{\tau_{t:n}}(\beta_t,\tau_t),$$
$$= \frac{\partial}{\partial \beta_{t,i}}\Bigg(\mathbb{1}_{i\in\tau_{t:n}}\frac{e^{\beta_{t,\tau_t}}}{\sum_{j\in\tau_{t:n}} e^{\beta_{t,j}}}\Bigg),$$
$$= \mathbb{1}_{i\in\tau_{t:n}}\Bigg[\frac{\mathbb{1}_{i=\tau_t}e^{\beta_{t,\tau_t}}}{\sum_{j\in\tau_{t:n}} e^{\beta_{t,j}}} - \frac{e^{\beta_{t,\tau_t}}e^{\beta_{t,i}}}{\big(\sum_{j\in\tau_{t:n}} e^{\beta_{t,j}}\big)^2}\Bigg],$$
$$= \Phi_{\tau_{t:n}}(\beta_t,\tau_t)\big(\mathbb{1}_{i=\tau_t} - \Phi_{\tau_{t:n}}(\beta_t,i)\big).$$

Eq. (S.2) becomes

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

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

$$\frac{\partial \mathbb{E}^{2,v}_{\beta,\theta}[R^{(1)}]}{\partial \beta_{t,i}} = \sum_{\tau\in\mathfrak{S}(n)} \pi^2_\beta(\sigma^{(1)}=\tau)\big(\mathbb{1}_{i=\tau_t} - \Phi_{\tau_{t:n}}(\beta_t,i)\big)\mathbb{E}^v_\theta\big[R^{(1)}\,|\,\sigma^{(1)}=\tau\big],$$
$$= \mathbb{E}^{2,v}_{\beta,\theta}\Big[\big(\mathbb{1}_{i=\sigma^{(1)}_t} - \Phi_{\sigma^{(1)}_{t:n}}(\beta_t,i)\big)\big(R^{(1)} - B^{(1)}\big)\Big],$$

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

2.2 Proof of Theorem. 2

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

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

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

$$\begin{aligned}\frac{\partial \pi^A_\theta(A^{(1)}=\mathbf{a}\,|\,\sigma^{(1)}=\tau)}{\partial\theta_{i,a}} &= \frac{\partial}{\partial\theta_{i,a}}\Bigg(\prod_{j\in[n]}\pi^A_\theta\big(A^{(1)}_j=a_j\,|\,\sigma^{(1)}=\tau\big)\Bigg),\\ &= \Bigg(\prod_{\substack{j\in[n]\\ j\neq i}}\pi^A_\theta\big(A^{(1)}_j=a_j\,|\,\sigma^{(1)}=\tau\big)\Bigg)\underbrace{\frac{\partial}{\partial\theta_{i,a}}\Bigg(\frac{e^{\theta_{i,a_i}}}{\sum\limits_{b\in[k_i]}e^{\theta_{i,b}}}\Bigg)}_{(\star)}.\end{aligned}$$

Since

$$\begin{aligned}(\star) &= \frac{\mathbb{1}_{a_i=a}e^{\theta_{i,a_i}}}{\sum\limits_{b\in[k_i]}e^{\theta_{i,b}}} - \frac{e^{\theta_{i,a_i}}e^{\theta_{i,a}}}{\big(\sum\limits_{b\in[k_i]}e^{\theta_{i,b}}\big)^2},\\ &= \mathbb{1}_{a_i=a}\,\Phi(\theta_i,a_i) - \Phi(\theta_i,a_i)\Phi(\theta_i,a),\\ &= \pi^A_\theta(A^{(1)}_i=a_i\,|\,\sigma^{(1)}=\tau)\big(\mathbb{1}_{a_i=a}-\Phi(\theta_i,a)\big).\end{aligned}$$

Thus,

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

Therefore, the gradient of the objective function is

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

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

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

$$\begin{aligned}\frac{\partial \pi^B_\theta(A^{(1)}=\mathbf{a}\,|\,\sigma^{(1)}=\tau)}{\partial\theta_{i,t,a}} &= \frac{\partial}{\partial\theta_{i,t,a}}\Bigg(\prod_{j\in[n]}\pi^B_\theta\big(A^{(1)}_j=a_j\,|\,\sigma^{(1)}=\tau\big)\Bigg),\\ &= \frac{\partial}{\partial\theta_{i,t,a}}\Bigg(\prod_{s\in[n]}\Phi(\theta_{\tau_s,s},a_{\tau_s})\Bigg),\\ &= \Bigg(\prod_{\substack{s\in[n]\\ s\neq t}}\pi^B_\theta\big(A^{(1)}_{\tau_s}=a_{\tau_s}\,|\,\sigma^{(1)}=\tau\big)\Bigg)\mathbb{1}_{\tau_t=i}\underbrace{\frac{\partial}{\partial\theta_{i,t,a}}\Bigg(\frac{e^{\theta_{i,t,a_i}}}{\sum\limits_{b\in[k_i]}e^{\theta_{i,t,b}}}\Bigg)}_{(\star)}.\end{aligned}$$

Since

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

Thus,

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

Therefore, the gradient of the objective function is

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

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

3 Additional figures

3.1 Permutation sampling

Case 1

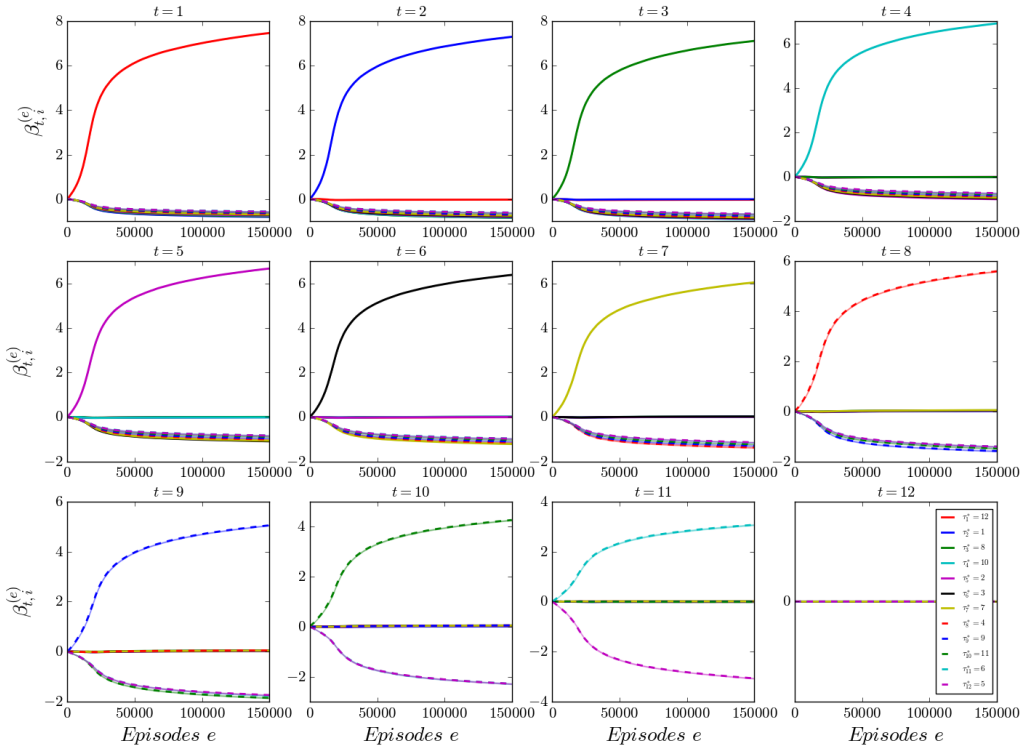
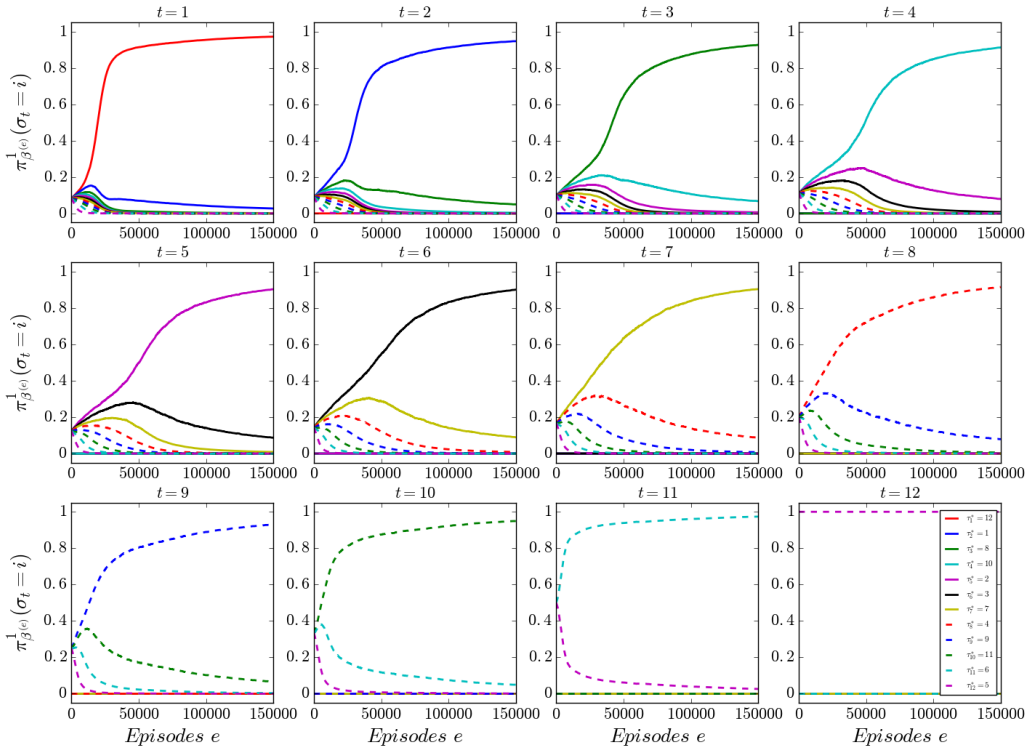
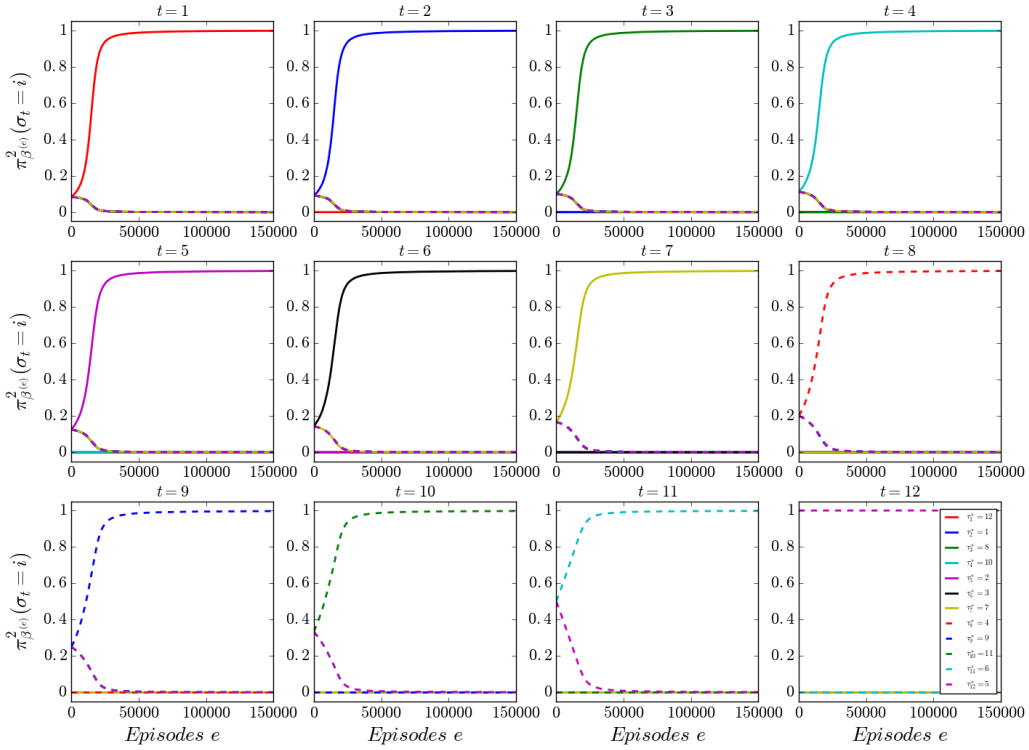


Figure S.1: The average parameter $(\beta)_i$ on model 2 for each agent t in Sec. 5.1.1. Each color/shape corresponds to a machine i.



(a)



(b)

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

Case 2

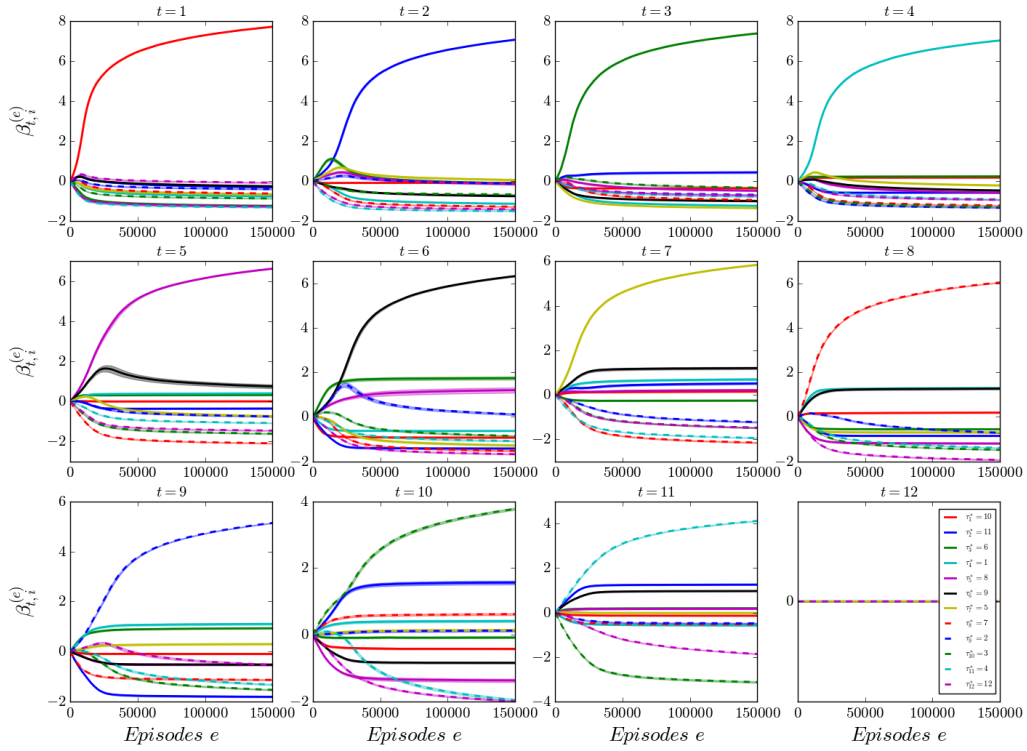
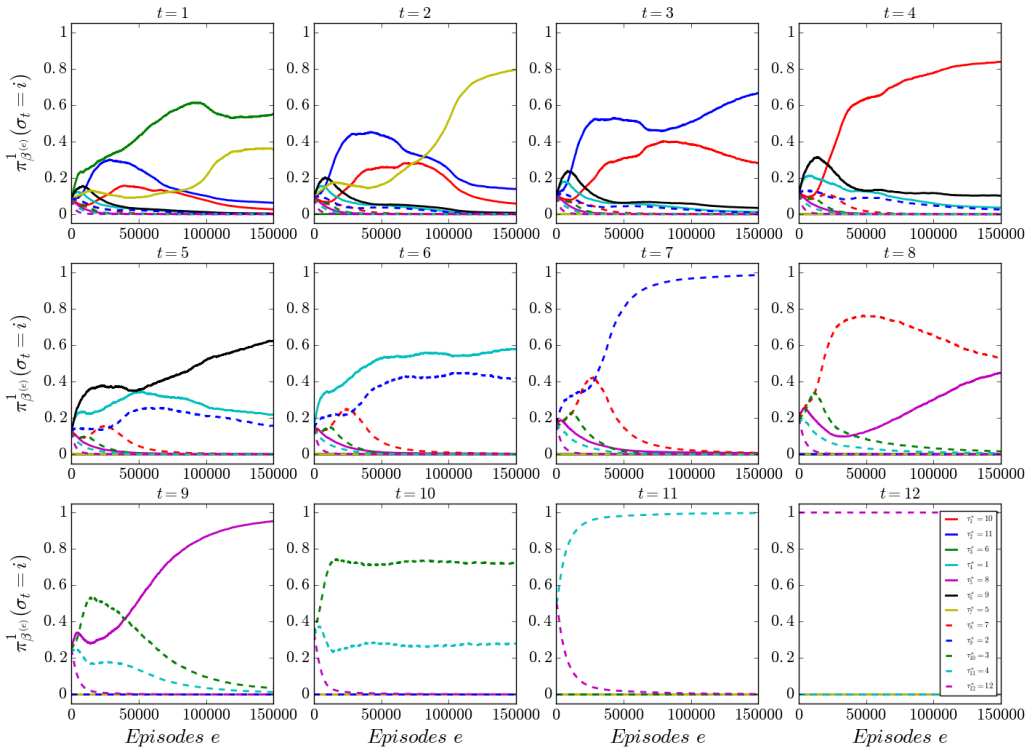
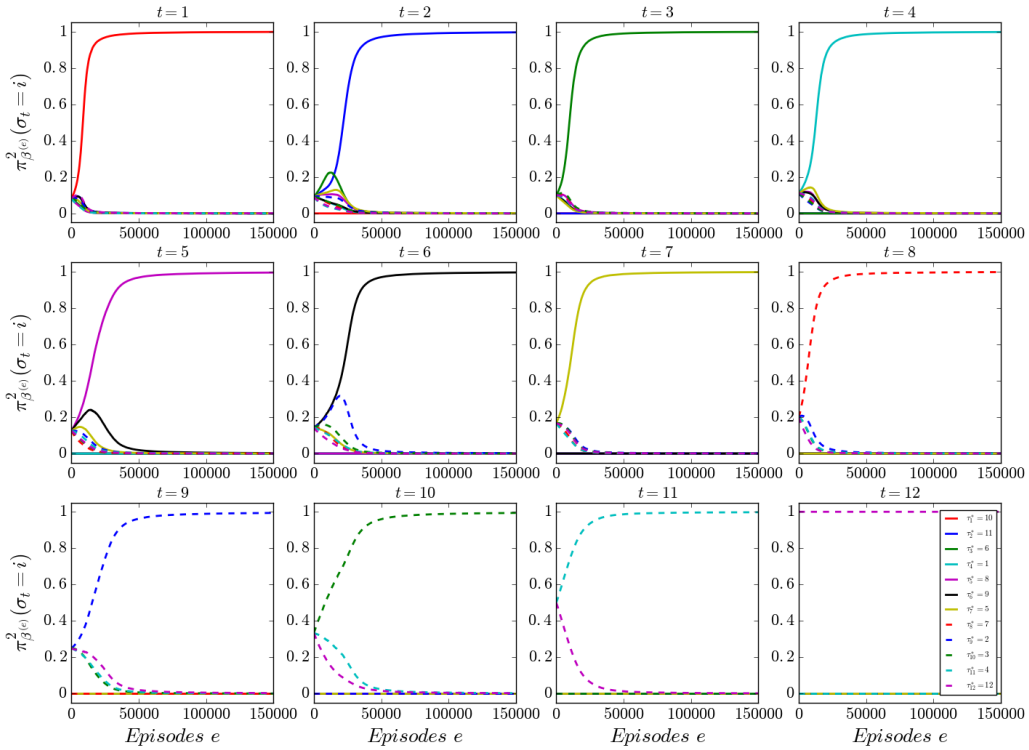


Figure S.3: The average parameter $(\beta)_i$ on model 2 for each agent t in Sec. 5.1.2. Each color/shape corresponds to a machine i.



(a)

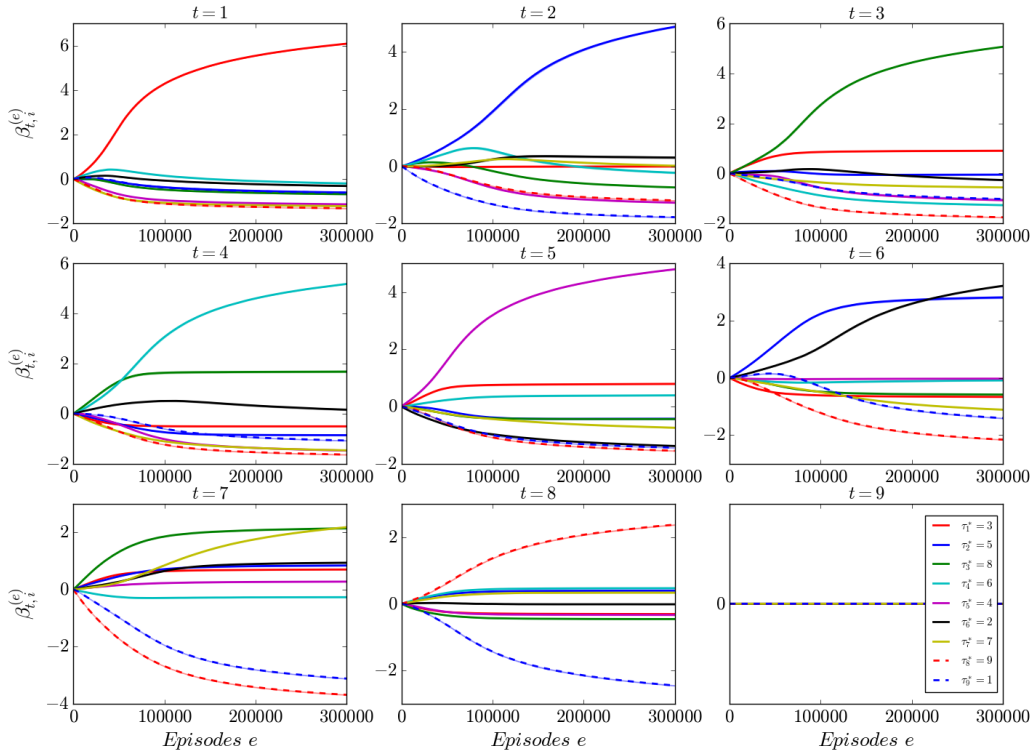


(b)

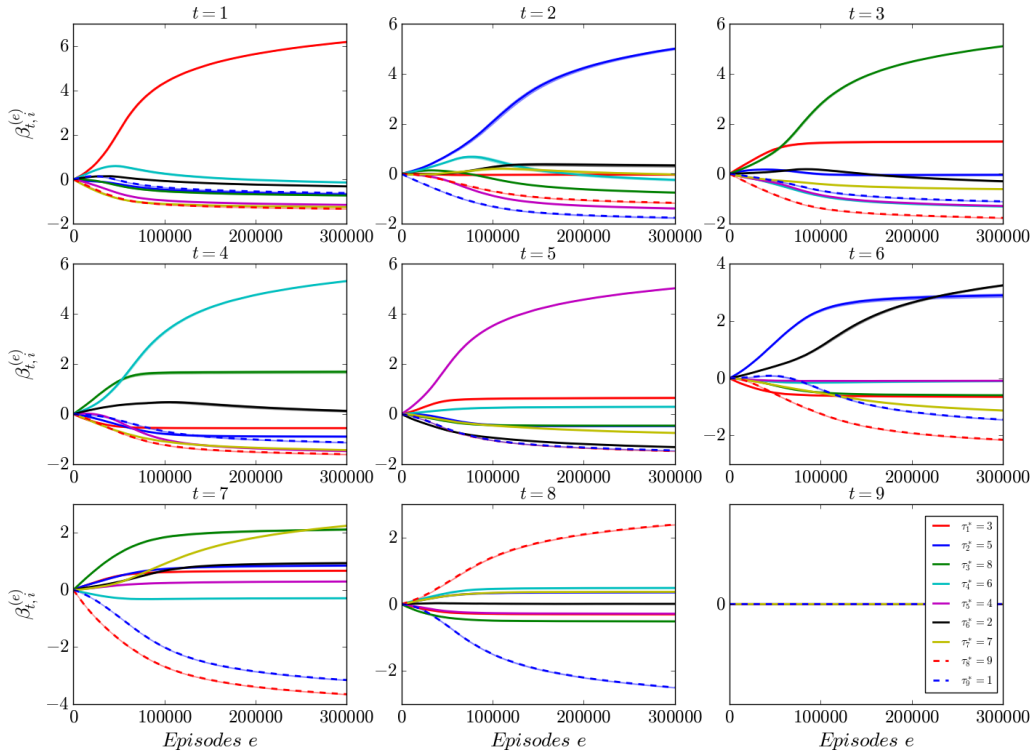
Figure S.4: $\pi^u_{\beta^{(e)}}(\sigma_t = \hat{\tau})$ for model 1 in (a) and model 2 in (b). $\hat{\tau}$ is obtained as in Eq. 8 for each model.

3.2 Sampling permutation and actions

Case 1

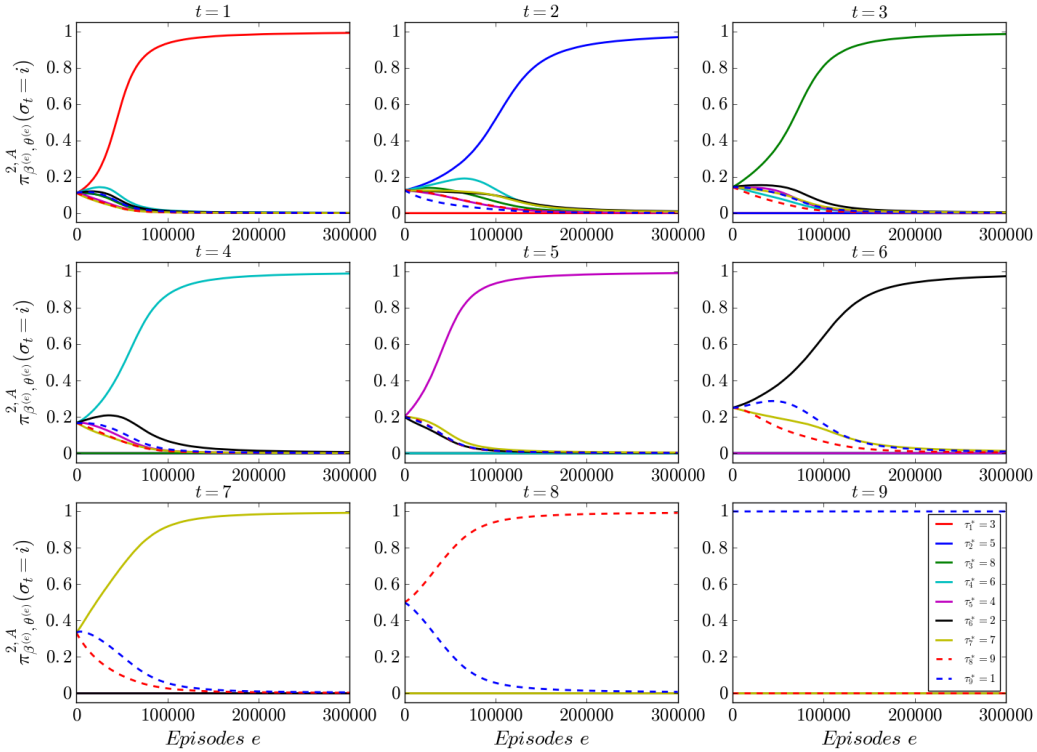


(a)

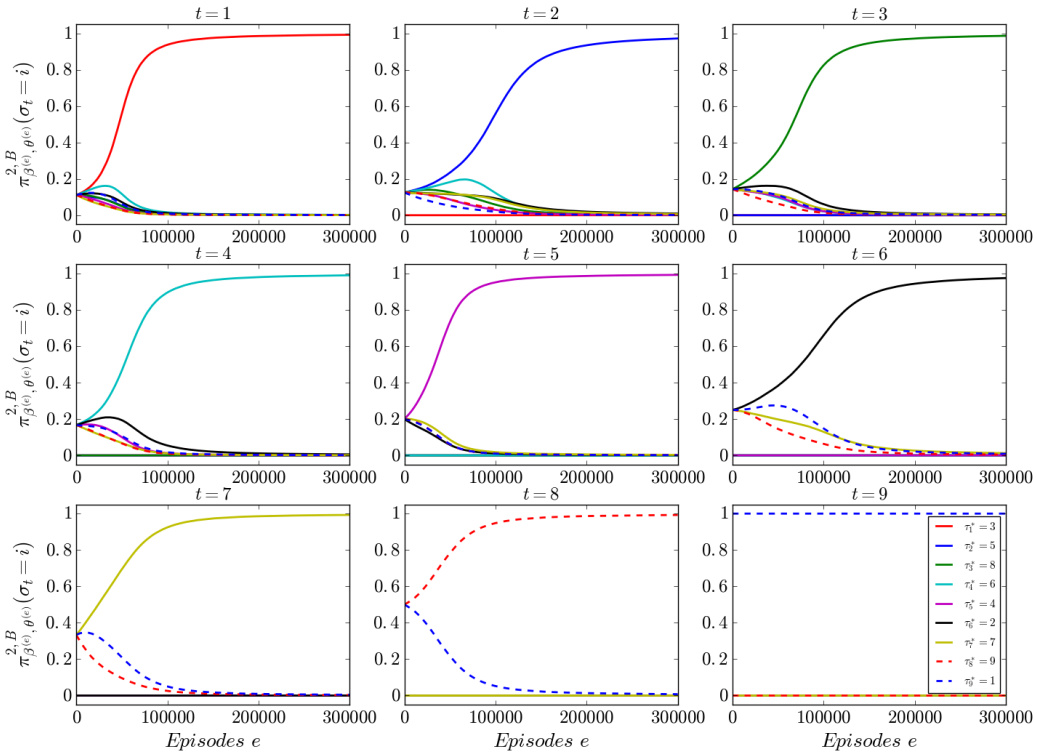


(b)

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.

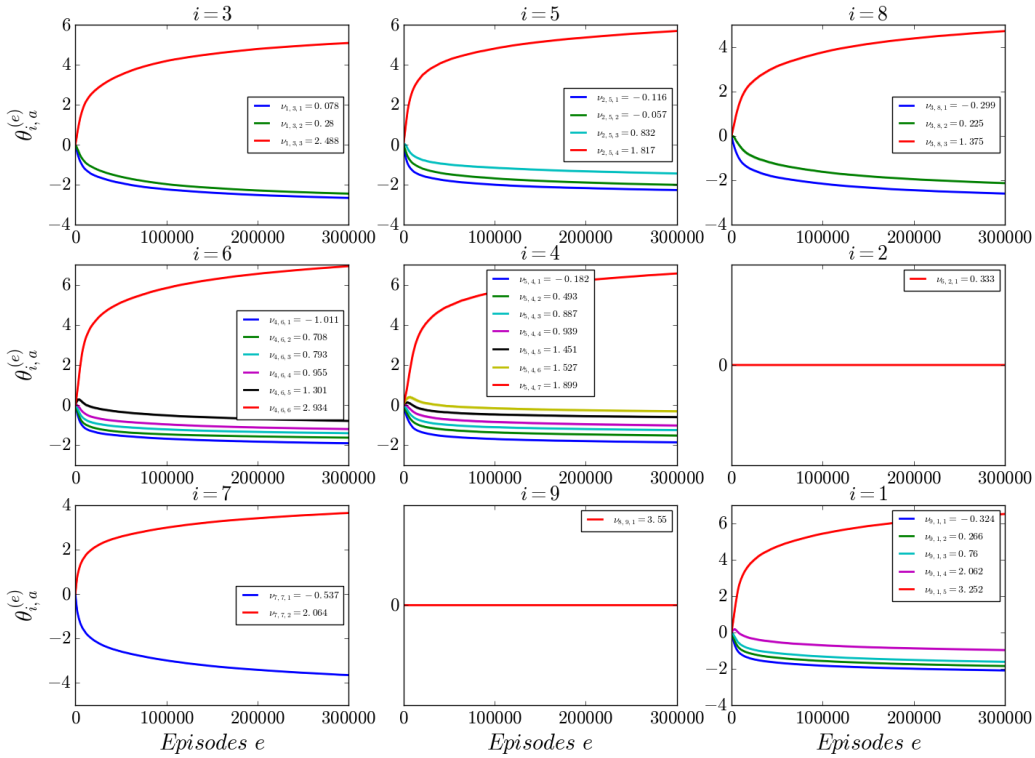


(a)

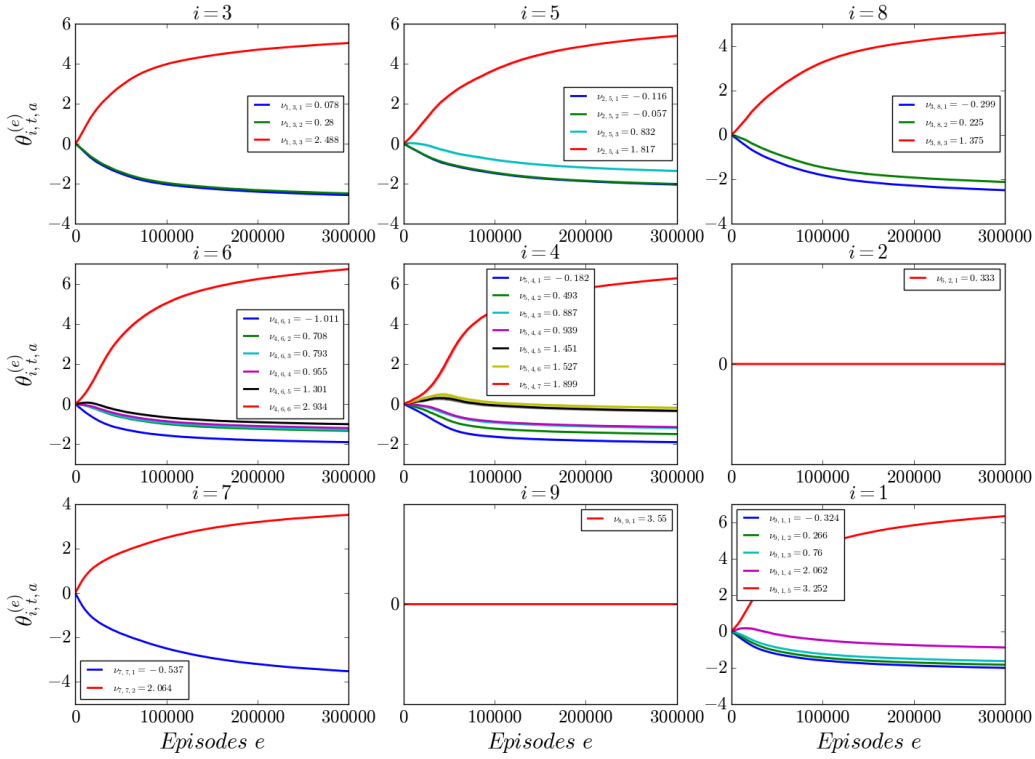


(b)

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

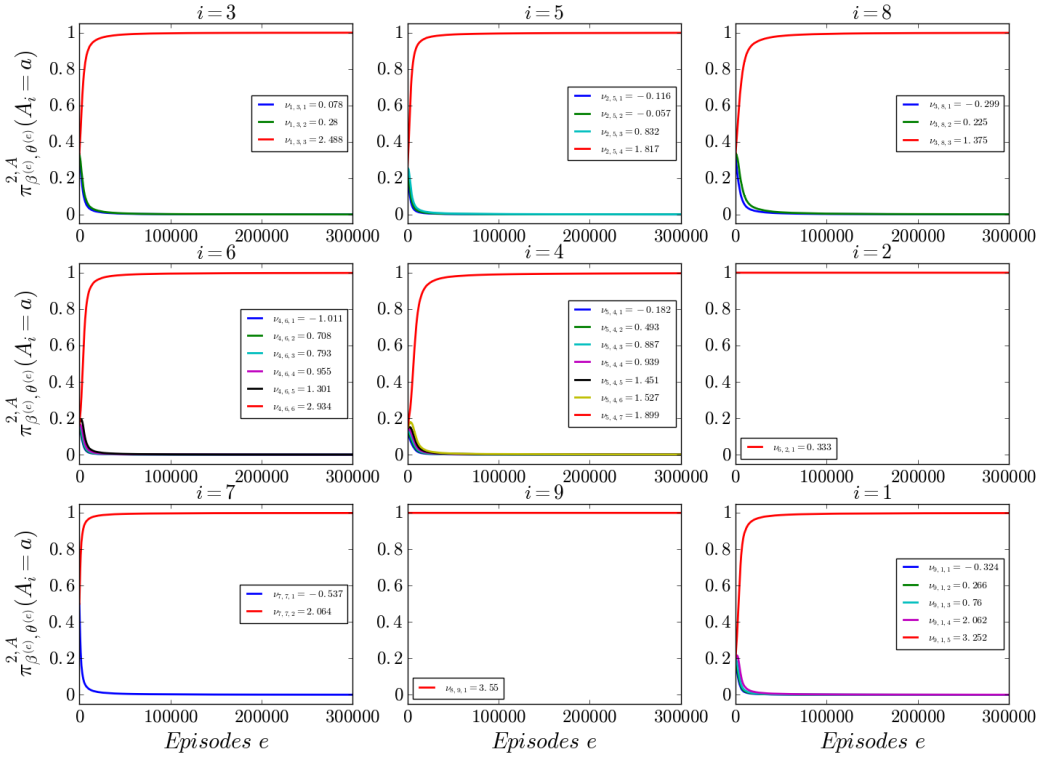


(a)

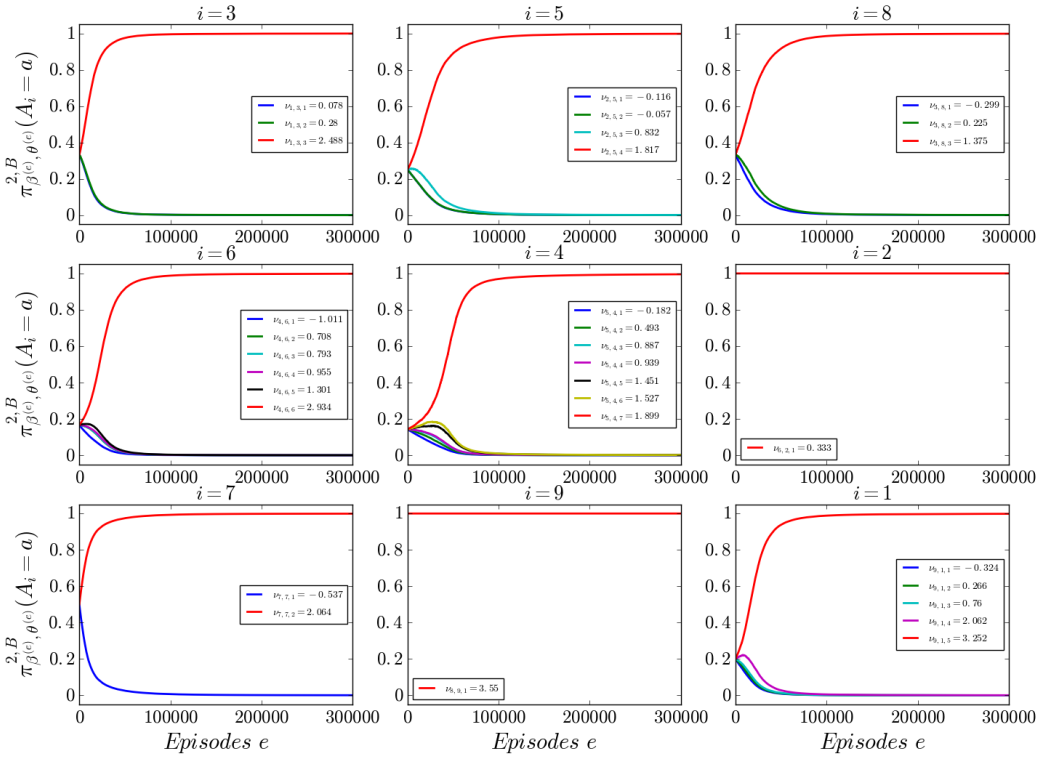


(b)

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.



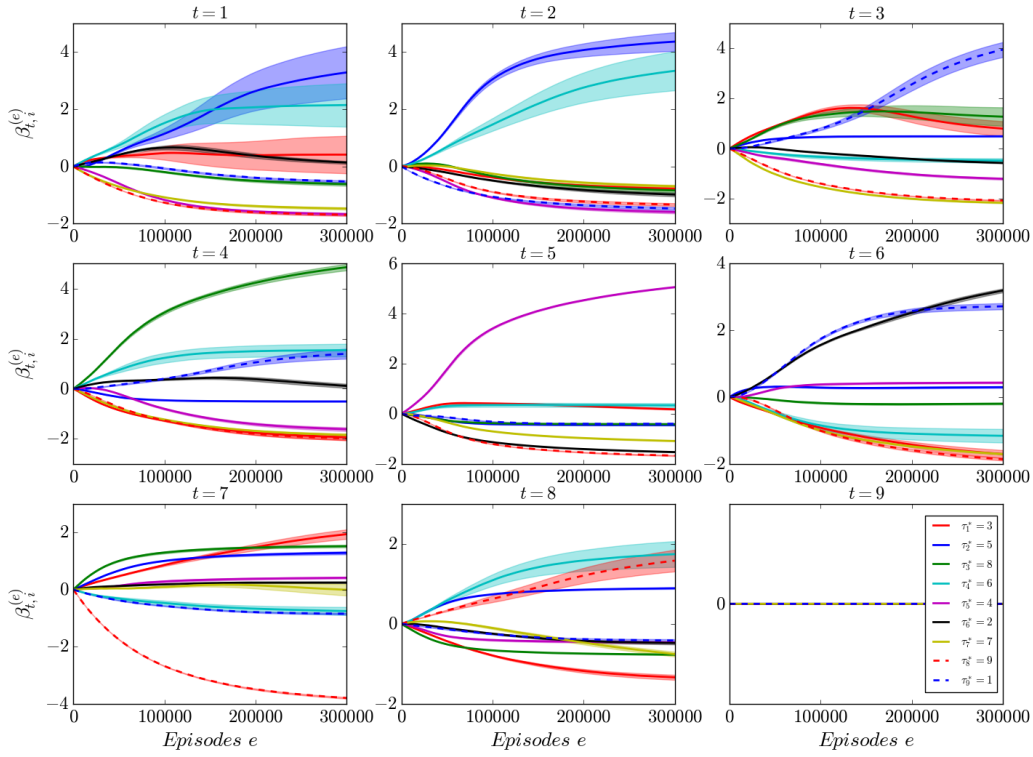
(a)



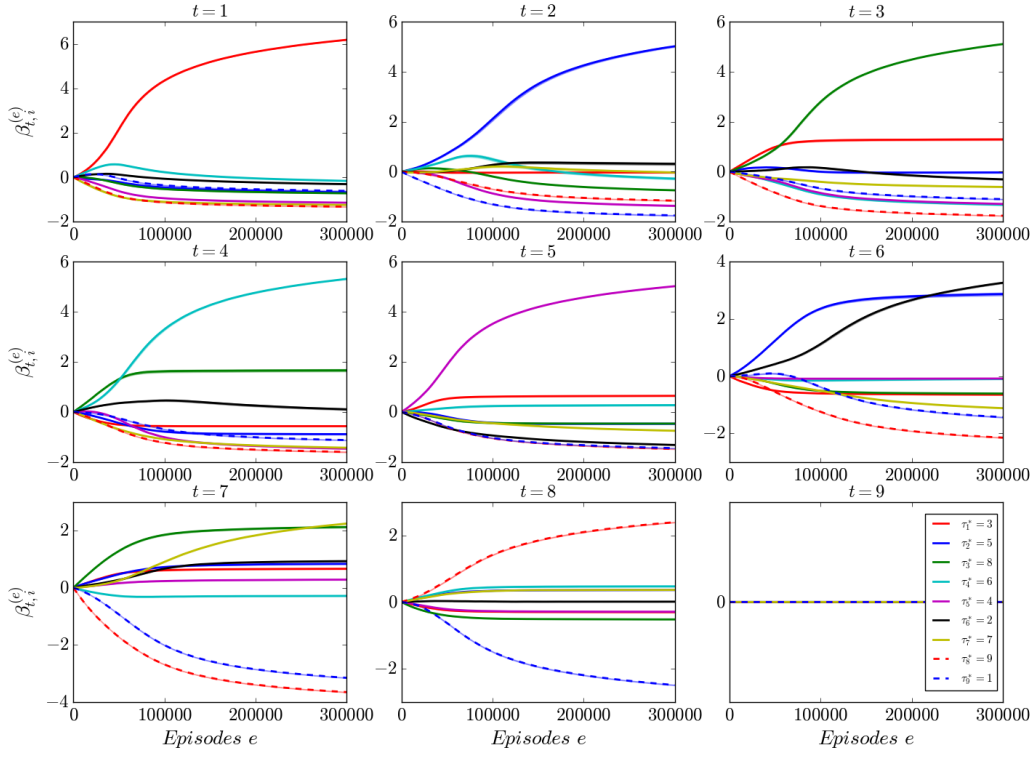
(b)

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

Case 2

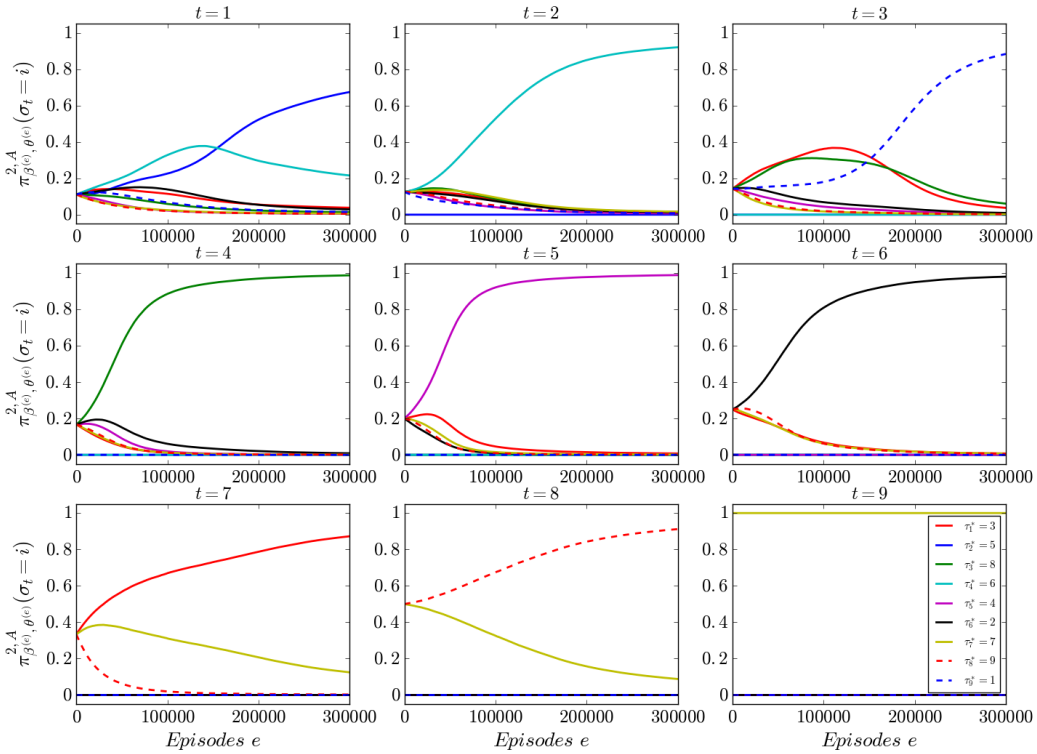


(a)

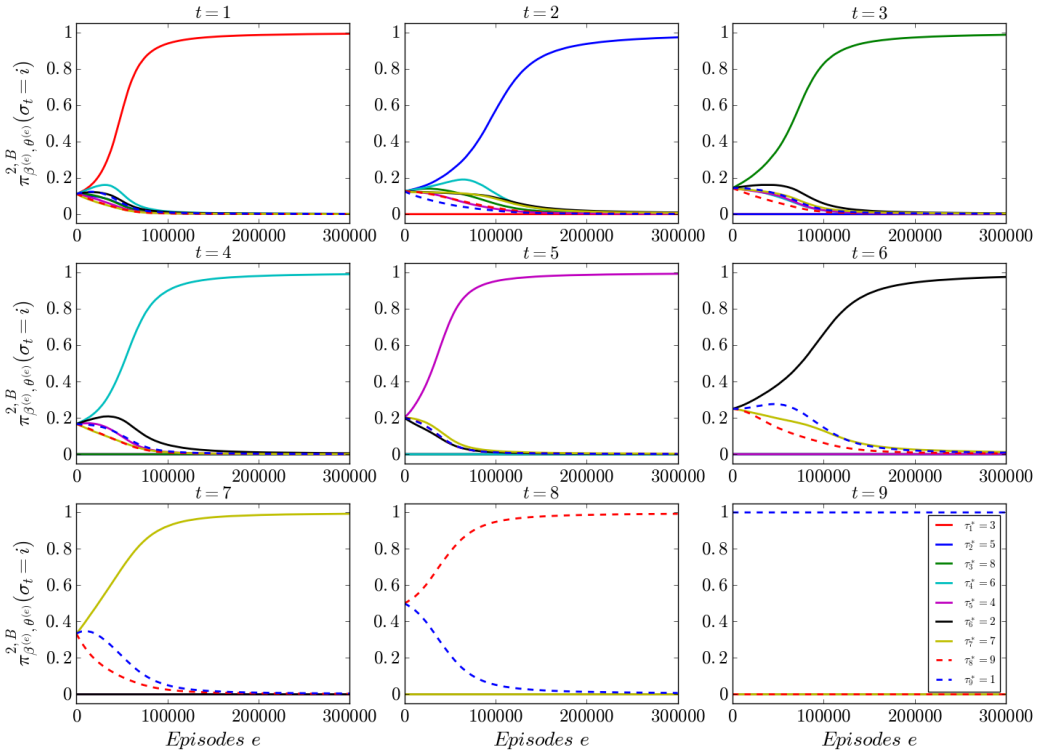


(b)

Figure S.9: he 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.

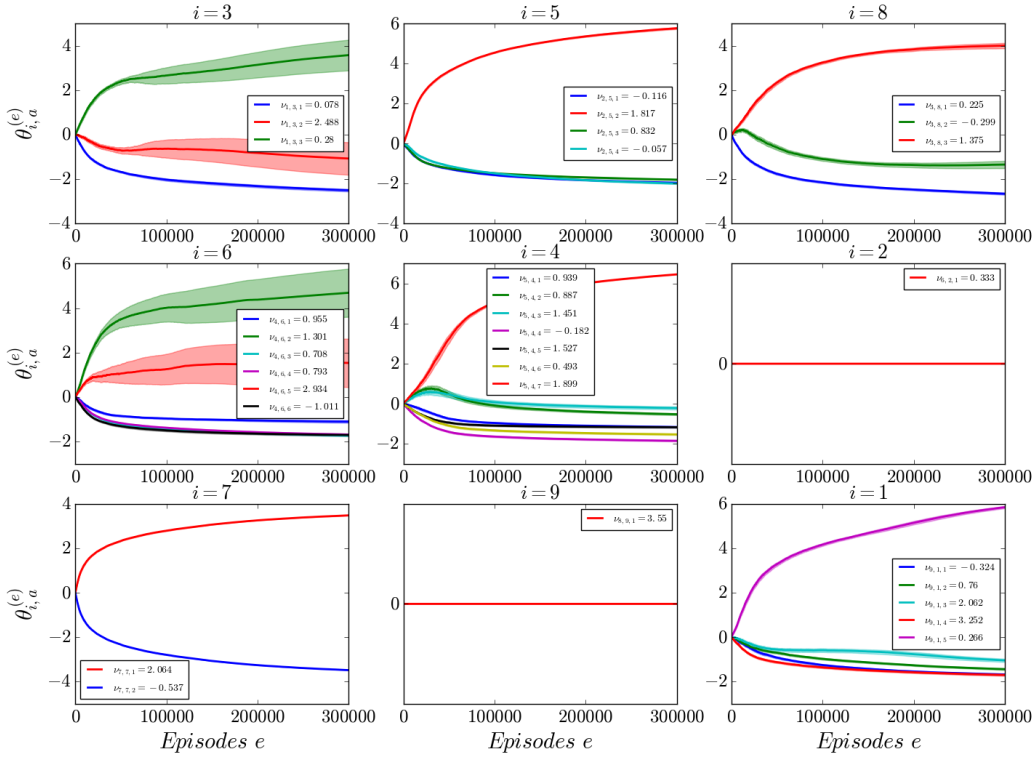


(a)

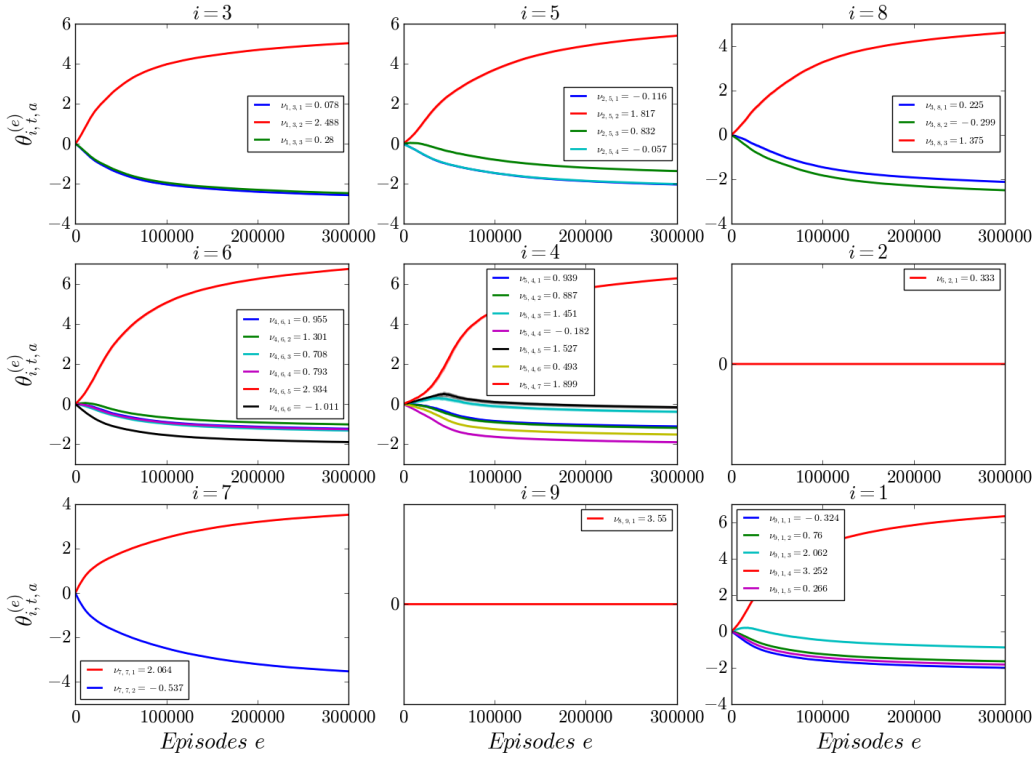


(b)

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

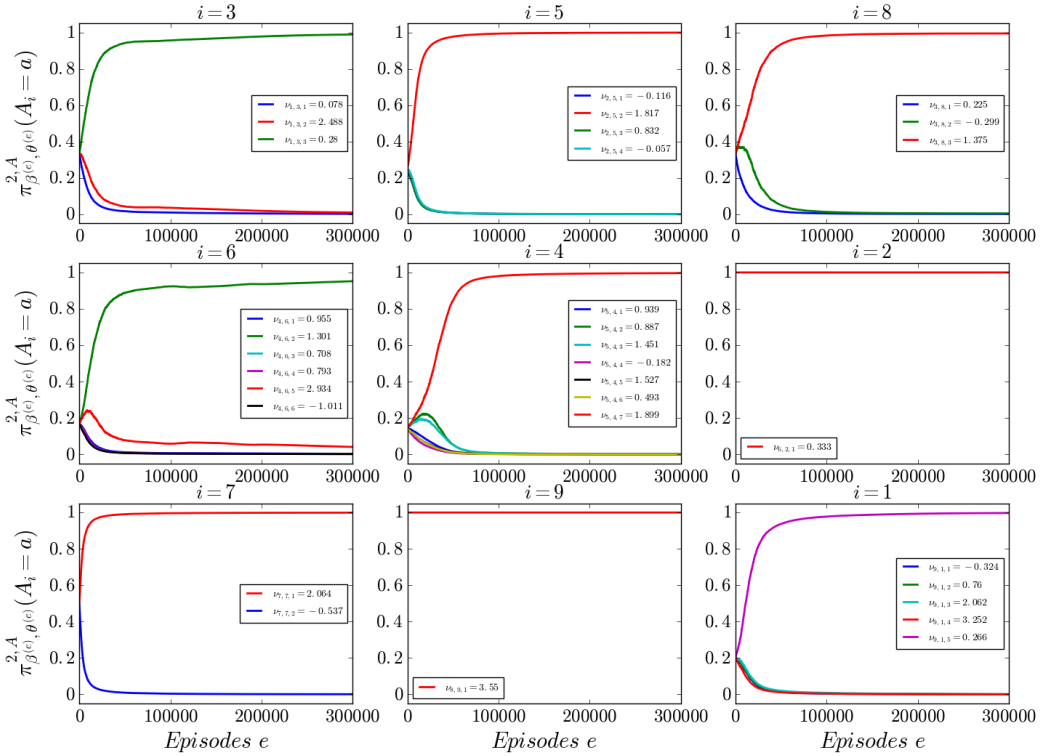


(a)

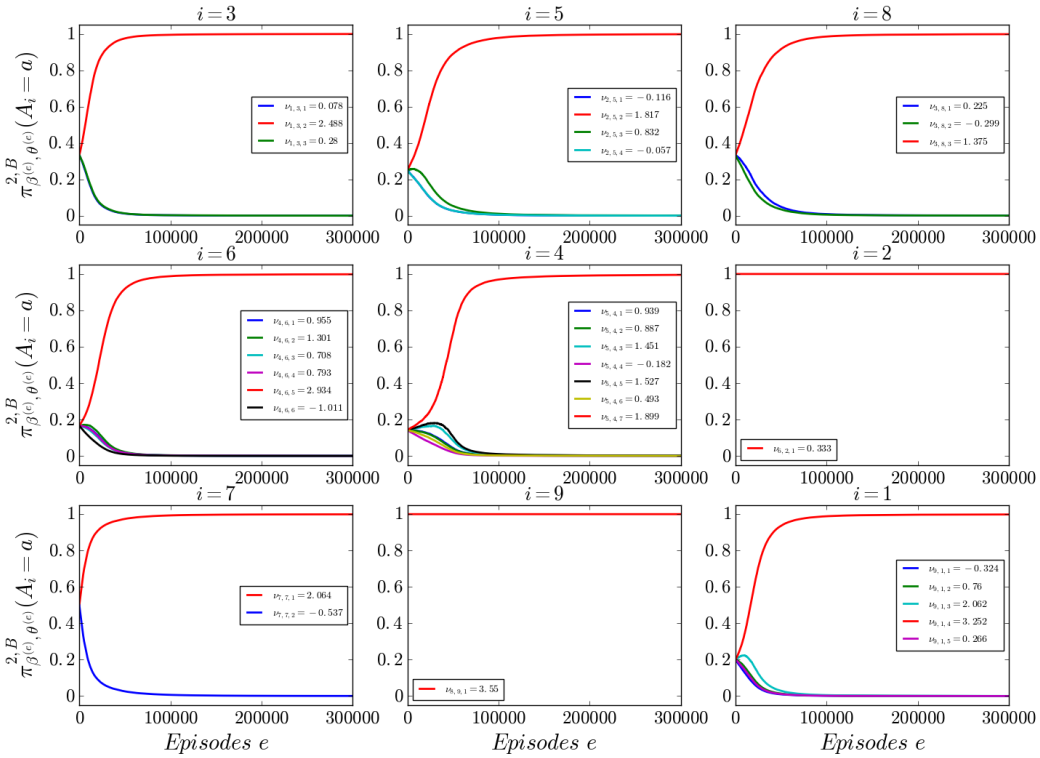


(b)

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