Technical Companion for "Inpatient Overflow Management with Proximal Policy Optimization"

1 Proof of Proposition 1

Following Dai and Gluzman (2022), the average cost gap between the two policies can be written as

$$
\begin{aligned}
&(\mu_\theta^h)^T \tilde{\mathbf{g}}_\theta^h - (\mu_\eta^h)^T \tilde{\mathbf{g}}_\eta^h \\
=&(\mu_\theta^h)^T (\tilde{\mathbf{g}}_\theta^h + (\tilde{\mathbf{P}}_\theta^h - I)\mathbf{v}_\eta^h)) - (\mu_\eta^h)^T \tilde{\mathbf{g}}_\eta^h \\
=&(\mu_\eta^h)^T (\tilde{\mathbf{g}}_\theta^h - (\mu_\eta^h)^T \tilde{\mathbf{g}}_\eta^h \mathbf{e} + (\tilde{\mathbf{P}}_\theta^h - I)\mathbf{v}_\eta^h) + (\mu_\theta^h - \mu_\eta^h)^T (\tilde{\mathbf{g}}_\theta^h + (\tilde{\mathbf{P}}_\theta^h - I)\mathbf{v}_\eta^h) \\
=&N_1^h(\theta,\eta) + N_2^h(\theta,\eta),
\end{aligned}
$$

where the first equation holds since the stationary distribution μ_θ^h satisfying $(\mu_\theta^h)^T(\tilde{\mathbf{P}}_\theta^h - I) = 0$. The third equation holds since for any constant α, $(\mu_\theta^h - \mu_\eta^h)^T \alpha \mathbf{e} = \alpha(\mu_\theta^h)^T \mathbf{e} - \alpha(\mu_\eta^h)^T \mathbf{e} = 0$, and we choose $\alpha = (\mu_\eta^h)^T \tilde{\mathbf{g}}_\eta^h$.Next, we characterize the decay rate of N_1^h and N_2^h.

For a given vector $\omega \in \mathcal{S}^h$, define two types of $\mathcal{V}$-norm as

$$
\|\omega\|_{\infty,\mathcal{V}} = \sum_{s\in\mathcal{S}^h} \frac{|\omega(s)|}{\mathcal{V}(s)}, \quad \|\omega\|_{1,\mathcal{V}} = \sum_{s\in\mathcal{S}^h} |\omega(s)|\mathcal{V}(s). \tag{1}
$$

Note that slightly different from the definition given in Dai and Gluzman (2022), we focus on each given h and the summation is taken over the states in the subspace $\mathcal{S}^h$. Similarly, we slightly adapt the $\mathcal{V}$-norm for a given matrix $\Omega \in \mathcal{S}^h \times \mathcal{S}^h$ as

$$
\|\Omega\|_{\mathcal{V}} = \sup_{s\in\mathcal{S}^h} \frac{1}{\mathcal{V}(s)} \sum_{s'\in\mathcal{S}^h} |\Omega(s,s')|\mathcal{V}(s'). \tag{2}
$$

We further denote

$$
\tilde{N}^h(\theta,\eta) = \tilde{\mathbf{g}}_\theta^h - (\mu_\eta^h)^T \tilde{\mathbf{g}}_\eta^h \mathbf{e} + (\tilde{\mathbf{P}}_\theta^h - I)\mathbf{v}_\eta^h,
$$

which is an $\mathcal{S}^h$-dimensional vector. Following Dai and Gluzman (2022), we can bound the absolute value of the two scalars N_1^h and N_2^h as

$$
\begin{aligned}
|N_1^h(\theta,\eta)| \le&(\mu_\eta^h)^T \mathcal{V} \cdot \|\tilde{N}^h(\theta,\eta)\|_{\infty,\mathcal{V}}, \\
|N_2^h(\theta,\eta)| \le&\|\mu_\theta^h - \mu_\eta^h\|_{1,\mathcal{V}} \cdot \|\tilde{N}^h(\theta,\eta)\|_{\infty,\mathcal{V}}.
\end{aligned} \tag{3}
$$

For the bound of N_1^h, only the term $\|\tilde{N}^h(\theta,\eta)\|_{\infty,\mathcal{V}}$ relates to the new parameter θ, while for N_2^h, it contains an additional term $\|\mu_\theta^h - \mu_\eta^h\|_{1,\mathcal{V}}$ which also relates to θ. According to Theorem 1 and Lemma 5 in Dai and Gluzman (2022), we have $\|\mu_\theta^h - \mu_\eta^h\|_{1,\mathcal{V}} = O(\|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)Z_\eta\|_\mathcal{V})$ and $\|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)Z_\eta\|_\mathcal{V} \to 0$ as $\theta \to \eta$, where the matrix Z_η is defined as

$$Z_\eta := \sum_{n=0}^{\infty} (\tilde{\mathbf{P}}_\eta^h - \Pi_\eta^h)^n. \tag{4}$$

To get the exact order of N_1^h, N_2^h with respect to $\|r_{\theta,\eta}^h - 1\|$, we need to further check the order of $\|\tilde{N}^h(\theta,\eta)\|_{\infty,\mathcal{V}}$ and $\|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)Z_\eta\|_\mathcal{V}$.

For $\|\tilde{N}^h(\theta,\eta)\|_{\infty,\mathcal{V}}$, we can rewrite and bound it as follows:

$$\|\tilde{N}^h(\theta,\eta)\|_{\infty,\mathcal{V}} \le \|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)\mathbf{v}_\eta^h\|_{\infty,\mathcal{V}} + \|\tilde{\mathbf{g}}_\theta^h - \tilde{\mathbf{g}}_\eta^h\|_{\infty,\mathcal{V}}. \tag{5}$$

Our remaining task is to analyze the order of $\|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)\mathbf{v}_\eta^h\|_{\infty,\mathcal{V}}$ and $\|\tilde{\mathbf{g}}_\theta^h - \tilde{\mathbf{g}}_\eta^h\|_{\infty,\mathcal{V}}$. Note that the second term on the right-hand side does not exist in Dai and Gluzman (2022) since they have action-independent cost. Also note that they characterize the decay rates of $N^h(\theta,\eta)$ in terms of $D_{\theta,\eta}^h := \|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)Z_\eta\|_\mathcal{V}$. However, since we cannot directly bound $\|\tilde{\mathbf{g}}_\theta^h - \tilde{\mathbf{g}}_\eta^h\|_{\infty,\mathcal{V}}$ with respect to $D_{\theta,\eta}^h$, we need to take one step further to bound N_1, N_2 directly to $\|r_{\theta,\eta}^h - 1\|_\infty$ instead of $D_{\theta,\eta}^h$.

First, we consider $\|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)\mathbf{v}_\eta^h\|_{\infty,\mathcal{V}}$. According to Lemma 3 of Dai and Gluzman (2022), the relative value function $\mathbf{v}_\eta^h$ can be rewritten as

$$\mathbf{v}_\eta^h = Z_\eta\big(\tilde{\mathbf{g}}_\eta^h - (\mu_\eta^h)^T\tilde{\mathbf{g}}_\eta^h\mathbf{e}\big).$$

Therefore, we can bound

$$\|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)\mathbf{v}_\eta^h\|_{\infty,\mathcal{V}} \le \|\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h\|_\mathcal{V}\|Z_\eta\|_\mathcal{V} \cdot \Big(\|\tilde{\mathbf{g}}_\eta^h\|_{\infty,\mathcal{V}} + (\mu_\eta^h)^T\tilde{\mathbf{g}}_\eta^h\Big). \tag{6}$$

Then, to bound the term $\|\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h\|_\mathcal{V}$, we recall that the one-day transition matrices are specified as:

$$\begin{aligned} \tilde{\mathbf{P}}_\eta^h &= \mathbf{P}_\eta^{h,h+1}\mathbf{P}_\eta^{h+1,h+2}\cdots\mathbf{P}_\eta^{h-1,h}, \\ \tilde{\mathbf{P}}_\theta^h &= \mathbf{P}_\theta^{h,h+1}\mathbf{P}_\eta^{h+1,h+2}\cdots\mathbf{P}_\eta^{h-1,h}. \end{aligned} \tag{7}$$

We denote the elements of these one-day transition matrices as $\{\tilde{p}_\eta^h(s'|s), s, s' \in \mathcal{S}^h\}$ and $\{\tilde{p}_\theta^h(s'|s), s, s' \in \mathcal{S}^h\}$, respectively. The probability $\tilde{p}_\theta^h(s'|s)$ follows

$$\begin{aligned} \tilde{p}_\theta^h(s'|s) &= \sum_{s^{h+1}\in\mathcal{S}^{h+1},\cdots,s^{h-1}\in\mathcal{S}^{h-1}} p_\theta^{h,h+1}(s^{h+1}|s)p_\eta^{h+1,h+2}(s^{h+2}|s^{h+1})\cdots p_\eta^{h-1,h}(s'|s^{h-1}) \\ &= \sum_{f\in\mathcal{A}(s)} \pi_\theta(f|s) \sum_{s^{h+1}\in\mathcal{S}^{h+1},\cdots,s^{h-1}\in\mathcal{S}^{h-1}} p^{h,h+1}(s^{h+1}|s,f)p_\eta^{h+1,h+2}(s^{h+2}|s^{h+1})\cdots p_\eta^{h-1,h}(s'|s^{h-1}), \end{aligned}$$

where $\{p_\eta^{\ell,\ell+1}(s'|s), s \in \mathcal{S}^\ell, s' \in \mathcal{S}^{\ell+1}\}$ is the set of elements of one-epoch transition matrix $\mathbf{P}_\eta^{\ell,\ell+1}$, and $p^{h,h+1}(s^{h+1}|s,f)$ is the one-epoch transition probability from state $s \in \mathcal{S}^h$ to $s^{h+1} \in \mathcal{S}^{h+1}$ given action f. Here we use the fact that, after the action f is fixed, the transition only depends on the arrivals and departures during this epoch and no longer depends on the action or policy. We denote

$$\tilde{p}_\eta^h(s'|s,f) = \sum_{s^{h+1}\in\mathcal{S}^{h+1},\cdots,s^{h-1}\in\mathcal{S}^{h-1}} p^{h,h+1}(s^{h+1}|s,f)p_\eta^{h+1,h+2}(s^{h+2}|s^{h+1})\cdots p_\eta^{h-1,h}(s'|s^{h-1}),$$

which was also used in Equation (23) and introduced there. This term is independent of the new policy parameter θ. Using this term, we can write that $\tilde{p}_\theta^h(s'|s) = \sum_{f\in\mathcal{A}(s)} \pi_\theta(f|s)\tilde{p}_\eta^h(s'|s,f)$. Similarly, we have $\tilde{p}_\eta^h(s'|s) = \sum_{f\in\mathcal{A}(s)} \pi_\eta(f|s)\tilde{p}_\eta^h(s'|s,f)$.

Therefore, the term $\|\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h\|_\mathcal{V}$ can be bounded as

$$\begin{aligned}
\|\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h\|_\mathcal{V} &= \sup_{s\in\mathcal{S}^h} \frac{1}{\mathcal{V}(s)} \sum_{s'\in\mathcal{S}^h} |\tilde{p}_\theta^h(s'|s) - \tilde{p}_\eta^h(s'|s)|\mathcal{V}(s') \\
&= \sup_{s\in\mathcal{S}^h} \frac{1}{\mathcal{V}(s)} \sum_{s'\in\mathcal{S}^h} \left| \sum_{f\in\mathcal{A}(s)} \pi_\theta(f|s)\tilde{p}_\eta^h(s'|s,f) - \sum_{f\in\mathcal{A}(s)} \pi_\eta(f|s)\tilde{p}_\eta^h(s'|s,f) \right| \mathcal{V}(s') \\
&= \sup_{s\in\mathcal{S}^h} \frac{1}{\mathcal{V}(s)} \sum_{s'\in\mathcal{S}^h} | \sum_{f\in\mathcal{A}(s)} (r_{\theta,\eta}(f|s)-1)\pi_\eta(f|s)\tilde{p}_\eta^h(s'|s,f)|\mathcal{V}(s') \\
&\leq \|\mathbf{r}_{\theta,\eta}^h - 1\|_\infty \sup_{s\in\mathcal{S}^h} \frac{1}{\mathcal{V}(s)} \sum_{s'\in\mathcal{S}^h} \tilde{p}_\eta(s'|s)\mathcal{V}(s') \\
&< \|\mathbf{r}_{\theta,\eta}^h - 1\|_\infty \sup_{s\in\mathcal{S}^h} \frac{1}{\mathcal{V}(s)} (b\mathcal{V}(s) + d\mathbf{1}_C(s)) \\
&\leq \|\mathbf{r}_{\theta,\eta}^h - 1\|_\infty \sup_{s\in\mathcal{S}^h} (b + \frac{d}{\mathcal{V}(s)}) \\
&\leq \|\mathbf{r}_{\theta,\eta}^h - 1\|_\infty (b+d),
\end{aligned}$$

where the second inequality holds because of the drift condition in Assumption 1, and the last inequality holds since we assume that $\mathcal{V} \geq 1$.

By plugging this upper bound into Equation (6), we have

$$\|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)\mathbf{v}_\eta^h\|_{\infty,\mathcal{V}} \leq \|\mathbf{r}_{\theta,\eta}^h - 1\|_\infty (b+d)\|Z_\eta\|_\mathcal{V} \cdot (\|\tilde{\mathbf{g}}_\eta^h\|_{\infty,\mathcal{V}} + (\mu_\eta^h)^T\tilde{\mathbf{g}}_\eta^h),$$

where $\|Z_\eta\|_\mathcal{V} < \infty$ from Theorem 16.1.2 in Meyn and Tweedie (2012), $\|\tilde{\mathbf{g}}_\eta^h\|_{\infty,\mathcal{V}} \leq 1$ since $\mathcal{V} \geq \tilde{\mathbf{g}}_\eta^h$ according to Assumption 1, and $(\mu_\eta^h)^T\tilde{\mathbf{g}}_\eta^h < \infty$ according to Lemma 1 of Dai and Gluzman (2022). As a result, we have $\|(\tilde{\mathbf{P}}_\theta^h - \tilde{\mathbf{P}}_\eta^h)\mathbf{v}_\eta^h\|_{\infty,\mathcal{V}} = O(\|\mathbf{r}_{\theta,\eta}^h - 1\|_\infty)$.

Next, we try to bound $\|\tilde{\mathbf{g}}_\theta^h - \tilde{\mathbf{g}}_\eta^h\|_{\infty,\mathcal{V}}$. Recall that the expected one-day cost vectors are:

$$\begin{aligned}
\tilde{\mathbf{g}}_\theta^h &= \mathbf{g}_\theta^h + \mathbf{P}_\theta^{h,h+1}\mathbf{g}_\eta^{h+1} + (\mathbf{P}_\theta^{h,h+1}\mathbf{P}_\eta^{h+1,h+2})\mathbf{g}_\eta^{h+2} + \cdots + (\mathbf{P}_\theta^{h,h+1}\mathbf{P}_\eta^{h+1,h+2}\cdots\mathbf{P}_\eta^{h-2,h-1})\mathbf{g}_\eta^{h-1} \\
\tilde{\mathbf{g}}_\eta^h &= \mathbf{g}_\eta^h + \mathbf{P}_\eta^{h,h+1}\mathbf{g}_\eta^{h+1} + (\mathbf{P}_\eta^{h,h+1}\mathbf{P}_\eta^{h+1,h+2})\mathbf{g}_\eta^{h+2} + \cdots + (\mathbf{P}_\eta^{h,h+1}\mathbf{P}_\eta^{h+1,h+2}\cdots\mathbf{P}_\eta^{h-2,h-1})\mathbf{g}_\eta^{h-1}.
\end{aligned} \tag{8}$$

We denote the elements of the one-day cost vector $\tilde{\mathbf{g}}^h_\eta$ and $\tilde{\mathbf{g}}^h_\theta$ as $\{\tilde{g}^h_\eta(s), s \in \mathcal{S}^h\}$ and $\{\tilde{g}^h_\theta(s), s \in \mathcal{S}^h\}$, respectively. Moreover, we denote

$$\begin{aligned}\tilde{g}^h(s,f) =& g^h(s,f) + \sum_{s^{h+1}\in\mathcal{S}^{h+1}} p^{h,h+1}(s^{h+1}|s,f)\Big(g^{h+1}_\eta(s^{h+1}) + \sum_{s^{h+1}\in\mathcal{S}^{h+1},s^{h+2}\in\mathcal{S}^{h+2}} p^{h+1,h+2}_\eta(s^{h+2}|s^{h+1})g^{h+2}_\eta(s^{h+2}\\ &+\cdots+\sum_{s^{h+1}\in\mathcal{S}^{h+1},\cdots,s^{h-1}\in\mathcal{S}^{h-1}} p^{h+1,h+2}_\eta(s^{h+2}|s^{h+1})\cdots p^{h-2,h-1}_\eta(s^{h-1}|s^{h-2})g^{h-1}_\eta(s^{h-1})\Big)\end{aligned}$$

which was also used in Equation (23). Following the same argument as for the transition probabilities, this term is independent of the new policy parameter θ, which implies that

$$\tilde{g}^h_\theta(s) = \sum_{f\in\mathcal{A}(s)} \pi_\theta(f|s)\tilde{g}^h_\eta(s,f), \quad \tilde{g}^h_\eta(s) = \sum_{f\in\mathcal{A}(s)} \pi_\eta(f|s)\tilde{g}^h_\eta(s,f).$$

Then, the term $\|\tilde{\mathbf{g}}^h_\theta - \tilde{\mathbf{g}}^h_\eta\|_{\infty,\mathcal{V}}$ can be bounded as

$$\begin{aligned}\|\tilde{\mathbf{g}}^h_\theta - \tilde{\mathbf{g}}^h_\eta\|_{\infty,\mathcal{V}} &= \sup_{s\in\mathcal{S}^h}\frac{|\tilde{g}^h_\theta(s)-\tilde{g}^h_\eta(s)|}{\mathcal{V}(s)}\\ &= \sup_{s\in\mathcal{S}^h}\frac{|\sum_{f\in\mathcal{A}(s)}\pi_\theta(f|s)\tilde{g}^h_\eta(s,f)-\sum_{f\in\mathcal{A}(s)}\pi_\eta(f|s)\tilde{g}^h_\eta(s,f)|}{\mathcal{V}(s)}\\ &= \sup_{s\in\mathcal{S}^h}\frac{|\sum_{f\in\mathcal{A}(s)}(r_{\theta,\eta}(f|s)-1)\pi_\eta(f|s)\tilde{g}^h_\eta(s,f)|}{\mathcal{V}(s)}\\ &\le \sup_{s\in\mathcal{S}^h}\frac{\|\mathbf{r}^h_{\theta,\eta}-1\|_\infty|\sum_{f\in\mathcal{A}(s)}\pi_\eta(f|s)\tilde{g}^h_\eta(s,f)|}{\mathcal{V}(s)}\\ &= \|\mathbf{r}^h_{\theta,\eta}-1\|_\infty\|\tilde{\mathbf{g}}^h_\eta\|_{\infty,\mathcal{V}}\\ &\le \|\mathbf{r}^h_{\theta,\eta}-1\|_\infty,\end{aligned} \tag{9}$$

which implies $\|\tilde{\mathbf{g}}^h_\theta - \tilde{\mathbf{g}}^h_\eta\|_{\infty,\mathcal{V}} = O(\|\mathbf{r}^h_{\theta,\eta}-1\|_\infty)$ as well. The second-to-last inequality holds since in our setting $\tilde{g}^h_\eta(s,f) \ge 0, \forall (s,f)$, and the last inequality holds because $\mathcal{V} \ge \tilde{\mathbf{g}}^h_\eta$ implies that $\|\tilde{\mathbf{g}}^h_\eta\|_{\infty,\mathcal{V}} \le 1$.

By far, we have shown that both terms in the bound for $\|\tilde{N}^h(\theta,\eta)\|_{\infty,\mathcal{V}}$ has the same order $O(\|\mathbf{r}^h_{\theta,\eta}-1\|_\infty)$, so we also have $\|\tilde{N}^h(\theta,\eta)\| = O(\|\mathbf{r}^h_{\theta,\eta}-1\|_\infty)$. As a result, according to (3) and the fact that $\|\mu^h_\theta - \mu^h_\eta\|_{1,\mathcal{V}} = O(\|\tilde{\mathbf{P}}^h_\theta - \tilde{\mathbf{P}}^h_\eta\|_\mathcal{V}\|Z_\eta\|_\mathcal{V}) = O(\|\mathbf{r}^h_{\theta,\eta}-1\|_\infty)$, we have $N_1(\theta,\eta) = O(\|\mathbf{r}^h_{\theta,\eta}-1\|_\infty)$, and $N_2(\theta,\eta) = O(\|\mathbf{r}^h_{\theta,\eta}-1\|^2_\infty)$, which completes the proof. □

2 Illustration of PPO in Two-pool Setting

In this section, we present an illustration of the mechanism behind PPO in our specific context – the overflow assignment for inpatients. For illustration purpose, we focus on a simple two-pool midnight model with randomized atomic action. Furthermore, we focus on illustrating the updates for one given state s and assume the policy at other states remain unchanged. That is, the objective in this showcase example is to minimize

$$\hat{N}_1(\theta, s) := \mathop{\mathbb{E}}_{f \sim \pi_\eta(\cdot|s)}[r_{\theta,\eta}(f|s)\hat{A}_\eta(s,f)] = \mathop{\mathbb{E}}_{f \sim \pi_\theta(\cdot|s)}[\hat{A}_\eta(s,f)] \tag{10}$$

for the given state s. The clipping function can be easily added to this. By analyzing the gradient of $\hat{N}_1(\theta, s)$, we showcase how the overflow policy will change with different model parameters (B, C, μ, λ), providing some *explainability* of the mechanism behind PPO.

We consider a simple two-pool midnight MDP with one decision epoch each day ($m = 1$). The state is simplified as $s = (x_1, x_2) \in \mathbb{R}^2$, since we do not need to track the to-be-discharged counts when $m = 1$. Correspondingly, the transition dynamics from current state to the state of the next day given overflow action $f = \{f_{i,j}\}$ can be specified as

$$x'_j = x_j + a_j - d_j + \sum_{i=1, i\neq j}^{2} f_{i,j} - \sum_{\ell=1,\ell\neq j}^{2} f_{j,\ell}, \quad j = 1, 2.$$

where a_j, d_j denote the number of new arrivals and departures within a day. Here, a_j is a realization of the random variable A_j which follows $Pisson(\Lambda_j)$, and d_j is a realization of the random variable D_j which follows $Bin(q_j, \mu_j)$.

2.1 Policy Gradient

To facilitate the gradient analysis, we make additional assumptions.

Assumption 1. *(Symmetric Two-pool Midnight MDP) The two pools have* $(N_j, \lambda_j, \mu_j) = (N, \lambda, \mu)$ *for* $j = 1, 2$*;* $C_1 = C_2 = C$*,* $B_{12} = B_{21} = B$*.*

Under Assumption 2, we can define two state subspaces according to the feasible actions:

$$\mathcal{S}_1 = \{(x_1, x_2) \in \mathcal{S} : x_1 > N, x_2 < N\}; \quad \mathcal{S}_2 = \{(x_1, x_2) \in \mathcal{S} : x_1 < N, x_2 > N\}.$$

Recall that the system-level action takes the form $f = \{f_{i,j}, i, j = 1, 2\}$, where $f_{i,j}$ represents the number of assignments from class i to pool j. According to the definition of feasible action defined in Equation (1), for $s \in \mathcal{S}_1$, the feasible action space is $\{\{q_1 - f_{1,2}, f_{1,2}, 0, 0\} : f_{1,2} = 0, 1, \ldots, \min\{q_1, N - x_2\}\}$, where $q_1 = (x_1 - N) \vee 0$ denotes the queue length of class 1. Similarly, for $s \in \mathcal{S}_2$, the feasible action space is $\{\{0, 0, f_{2,1}, q_2 - f_{2,1}\} : f_{2,1} = 0, 1, \ldots, \min\{q_2, N - x_1\}\}$, where $q_2 = (x_2 - N) \vee 0$ denotes the queue length of class 2. For any state $s \in \mathcal{S} \setminus (\mathcal{S}_1 \cup \mathcal{S}_2)$, the only feasible action is no-overflow (action $\{0, 0, 0, 0\}$). Without loss of generality, we focus on $\mathcal{S}_1$ in the following analysis, as the results can be easily extend to $\mathcal{S}_2$ due to symmetry in Assumption 1.

Assumption 2. *(Parametric Randomized Atomic Action)*

(i) Batched setting: For a given pre-action state s*, each atomic action* a^n *depends on* s*, i.e., not affected by the previous atomic assignment.*

(ii) Parametric logistic model: The routing probability for the atomic action of a class 1 customer is parameterized as

$$\kappa_\theta(1|s,1) = \frac{1}{1+\exp(\theta_1 x_1 + \theta_2 x_2 + \theta_0)}, \quad \kappa_\theta(2|s,1) = 1 - \kappa_\theta(1|s,1).$$

Under a randomized policy π_θ satisfying Assumption 2, for a given pre-action state $s \in \mathcal{S}_1$ and an associated feasible action $f = (q_1 - f_{1,2}, f_{1,2}, 0, 0)$, the overflow quantity $f_{1,2}$ follows a binomial distribution $Bin(q_1, \kappa_\theta(2|s,1))$ (note that we allow overflow assignments to a full server here). Therefore, the aim of PPO is to update the parameters $\theta = (\theta_0, \theta_1, \theta_2)' \in \mathbb{R}^3$ to minimize $\hat{N}_1(\theta, s) = \mathbb{E}_{f\sim\pi_\theta(\cdot|s)}[\hat{A}_\eta(s,f)]$ in (10) through the policy gradient approach.

Assumption 3. *(Advantage function approximation)*

(i) Linear approximation: The value function v_η is approximated with linear combinations of a set of linear and quadratic basis functions, i.e.,

$$\hat{v}_\eta(s) = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1^2 + \hat{\beta}_4 x_2^2,$$

where $\hat{\beta}_k, k = 1,2,3,4$ are the coefficient parameters.

(ii) Transition probability approximation: The waiting customers in buffers can be served with the same service time distribution as in server pools, and leave the system after service.

In Section 5.2 of the main paper, we approximate the relative value function with the linear combinations of a set linear/quadratic basis as well as some queueing-based basis. Here for simplicity, in Assumption 3(i), we focus on the simpler linear and quadratic basis. Assumption 3(ii) is made to simplify the evaluation of the advantage function $\hat{A}_\eta$, which can be computed according to

$$\hat{A}_\eta(s,f) = g(s,f) - \gamma_\eta + \mathbb{E}_{s'\sim p(\cdot|s,f)}[\hat{v}_\eta(s')] - \hat{v}_\eta(s). \tag{11}$$

2.2 Policy gradient

We state the policy gradient result in the following lemma, with its detailed proof in Section 2.3.

Lemma 1. *Under Assumptions 1-3, for any $s = (x_1, x_2) \in \mathcal{S}_1$ and $f = (q_1 - f_{1,2}, f_{1,2}, 0, 0)$,*

$$\frac{\partial \hat{N}_1(\theta, s)}{\partial \theta_0} = \nabla_0 \hat{N}_1(\theta, s),$$
$$\frac{\partial \hat{N}_1(\theta)}{\partial \theta_k} = \nabla_0 \hat{N}_1(\theta, s) \cdot x_k, \quad k = 1, 2.$$

Here,

$$\nabla_0 \hat{N}_1(\theta, s) = \sum_{f_{1,2}=0}^{q_1} \pi_\theta(f|s)\big(f_{1,2} - q_1\kappa_\theta(2|s,1)\big)\hat{A}_\eta(s,f)$$
$$= q_1\kappa_\theta(2|s,1)\big(1-\kappa_\theta(2|s,1)\big)\Big(2\hat{\beta}_3(1-\mu)^2\big(2(q_1-1)\kappa_\theta(2|s,1) + x_2 - x_1 + 1\big) + B - C\Big). \tag{12}$$

This closed form for policy gradient allows us to examine the optimal action that minimizes $\hat{N}_1(\theta, s)$. Through this examination, we generate insights into how the policy gradient approach is guiding us to find a good action under different model and cost parameters (λ, μ, B, C).

We start by analyzing the monotonicity of $\nabla_0 \hat{N}_1(\theta, s)$ w.r.t. $\kappa_\theta(2|s,1)$, which depends on the sign of $\hat{\beta}_3$. In the rest of the analysis, we focus on the case where $\hat{\beta}_3 > 0$ since it leads to non-trivial policy updates. In this case, the new overflow probability obtained by minimizing $\hat{N}_1(\theta, s)$ should either equals to 0 or 1, or make $\nabla_0 \hat{N}_1(\theta^*, s) = 0$ hold. The latter first-order condition gives us $\kappa_{\theta^*}(2|s,1) = \max\big(0, \min(1, \kappa^*(s))\big)$, where

$$\kappa^*(s) = \frac{N - x_2 - (B-C)/2\hat{\beta}_3(1-\mu)^2}{2(q_1-1)} + \frac{1}{2}.$$

We discuss the property of $\kappa^*(s)$ separately when $B \le C$ or $B > C$.

When $B \le C$, $\kappa^*(s) \ge \frac{N-x_2+q_1-1}{2(q_1-1)} \ge \frac{N-x_2}{q_1}$. If $\kappa_\theta(2|s,1) = \frac{N-x_2}{q_1}$, the expected number of overflow assignments equals to the number of idle servers in pool 2. This essentially corresponds to the complete-overflow policy, which is expected to be optimal when the overflow cost is cheap.

When $B > C$, $\kappa^*(s)$ decreases in x_2, which follows the intuition about a "good" policy, i.e., the more crowded pool 2 gets, the less overflow should be assigned from class 1 t pool 2. For how this action changes with x_1, we focus on examining the mean of overflow assignment, i.e., $q_1\kappa^*(s)$. We have

$$q_1\kappa^*(s) = \frac{N - x_2 - (B-C)/2\hat{\beta}_3(1-\mu)^2 + 1}{2} + \frac{N - x_2 - (B-C)/2\hat{\beta}_3(1-\mu)^2}{q_1 - 1} + \frac{1}{2}(q_1 - 1),$$

which increase with q_1 when $q_1 \ge 2(N - x_2) - \frac{B-C}{\hat{\beta}_3(1-\mu^2)}$, but decrease with q_1 otherwise. Therefore, when x_2 is close to N, the critical point $2(N - x_2) - \frac{B-C}{\hat{\beta}_3(1-\mu^2)} \le 1$, so $q_1\kappa^*(s)$ increase with q_1, which also follows the intuition about a "good" policy since we need to overflow more to balance the load when there are more waiting patients. However, when x_2 is small, which means there are enough idle beds, the mean value of overflow assignments firstly decrease then increase with x_1. This policy is desired since when x_1 is very large, load balancing is the first-order issue, so we overflow more when x_1 is large; in contrast, when x_1 is relatively small, we need to trade-off between holding cost and undesirable overflow assignments, so when x_1 is larger, even with the same mean value of overflow assignments, there is a larger possibility that it will conduct a large number of overflow assignments and occupy too much class 2 servers in that case, causing a very large future cost according to "snowball effect". Therefore, mean value of overflow assignments should decrease.

In addition, a critical term in $q_1\kappa^*(s)$ is the term $\frac{(B-C)}{2\hat{\beta}_3(1-\mu)^2}$. Through some argument, we can show that when $B > C$, α in with $B - C$ and μ. It implies that the willingness of overflow decrease with $B - C$ and μ. These results also follow our intuition because when overflow cost is closer to holding cost (the gap $(B - C)$ is smaller), we prefer to overflow more to help balance the system; when the busy servers completes jobs faster (larger μ), we prefer to let customers wait since they can be admitted into primary ward within a shorter time.

2.3 Proof of Lemma 1

For a given $s = (x_1, x_2) \in \mathcal{S}_1$, a feasible action f takes the form of $(q_1 - f_{1,2}, f_{1,2}, 0, 0)$. Under the assumptions in Section 2.1, we get from (10) that $N_1(\theta, s) = \mathbb{E}_{f\sim\pi_\theta(\cdot|s)}[\hat{A}_\eta(s, f)] = \sum_{f_{1,2}=0}^{q_1} \pi_\theta(f|s)\hat{A}_\eta(s, f)$. Therefore, taking derivative of $\hat{N}_1(\theta, s)$ w.r.t. $\theta_k, k = 0, 1, 2$, respectively, we get

$$\frac{\partial}{\partial\theta_k}\hat{N}_1(\theta, s) = \sum_{f_{1,2}=0}^{q_1} \frac{\partial\pi_\theta(f|s)}{\partial\theta_k}\hat{A}_\eta(s, f). \tag{13}$$

From Assumption 2(i), the number of overflow quantity from class 1 to pool 2, f_{12} follows $Bin(q_1, \kappa_\theta(2|s, 1))$ under policy π_θ, we can rewrite $\pi_\theta(f|s)$ as

$$\pi_\theta(f|s) = \begin{pmatrix} q_1 \\ f_{1,2} \end{pmatrix} \kappa_\theta(2|s, 1)^{f_{1,2}}(1 - \kappa_\theta(2|s, 1))^{q_1-f_{1,2}}. \tag{14}$$

Then, by using some algebra, we have

$$\frac{\partial}{\partial\theta_k}\pi_\theta(f|s) =\pi_\theta(f|s)\left(\frac{f_{1,2}}{\kappa_\theta(2|s, 1)} - \frac{q_1 - f_{1,2}}{1 - \kappa_\theta(2|s, 1)}\right)\frac{\partial\kappa_\theta(2|s, 1)}{\partial\theta_k}, \quad k = 0, 1, 2. \tag{15}$$

Furthermore, recall that from Assumption 2(ii), $\kappa_\theta(2|s, 1)$ is parameterized as a logistic function. Therefore, we can further write out the following form for the gradients of $\kappa_\theta(2|s, 1)$. For the gradient w.r.t. θ_0, we have

$$\begin{aligned}\frac{\partial\kappa_\theta(2|s, 1)}{\partial\theta_0} &=- \frac{\exp(-(\theta_1x_1 + \theta_2x_2 + \theta_0)) \cdot (-1)}{(1 + \exp(-(\theta_1x_1 + \theta_2x_2 + \theta_0)))^2} \\ &=\kappa_\theta(2|s, 1)(1 - \kappa_\theta(2|s, 1)).\end{aligned} \tag{16}$$

Similarly, for θ_1, θ_2, we get

$$\frac{\partial\kappa_\theta(2|s, 1)}{\partial\theta_k} = \kappa_\theta(2|s, 1)(1 - \kappa_\theta(2|s, 1))x_k. \tag{17}$$

Combining Equations (15) through (17) and plugging them back to (13), we get the final results of policy gradient as follows.

$$\frac{\partial \hat{N}_1(\theta, s)}{\partial \theta_0} = \sum_{f_{1,2}=0}^{q_1} \pi_\theta(f|s)\left(f_{1,2} - q_1\kappa_\theta(2|s,1)\right)\hat{A}_\eta(s,f),$$
$$\frac{\partial \hat{N}_1(\theta, s)}{\partial \theta_k} = \sum_{f_{1,2}=0}^{q_1} \pi_\theta(f|s)\left(f_{1,2} - q_1\kappa_\theta(2|s,1)\right)x_k\hat{A}_\eta(s,f), \quad k=1,2.$$

For simplicity, we use $\nabla_0 \hat{N}_1(\theta, s)$ to denote

$$\sum_{f_{1,2}=0}^{q_1} \pi_\theta(f|s)\left(f_{1,2} - q_1\kappa_\theta(2|s,1)\right)x_k\hat{A}_\eta(s,f). \tag{18}$$

As a result, the policy gradient can be rewritten as

$$\frac{\partial \hat{N}_1(\theta, s)}{\partial \theta_0} = \nabla_0 \hat{N}_1(\theta, s),$$
$$\frac{\partial \hat{N}_1(\theta)}{\partial \theta_k} = \nabla_0 \hat{N}_1(\theta, s) \cdot x_k, \quad k=1,2.$$

Next, to derive the closed form of the policy gradient, we need to derive the closed form of $\hat{A}_\eta$ and plug it into (18). Recall that given a pre-action state $s \in \mathcal{S}_1$ and a feasible action $f = (f_{1,2}, q_1 - f_{1,2}, 0, 0)$ with $0 \le f_{1,2} \le q_1$, the advantage function $\hat{A}(s,f)$ can be computed via

$$\hat{A}_\eta(s,f) = g(s,f) + \mathbb{E}_{s' \sim p(\cdot|s,f)}[\hat{v}_\eta(s')], \tag{19}$$

where the current cost follows

$$g(s,f) = C(q_1 - f_{1,2}) + Bf_{1,2},$$

and according to Assumption 3, the estimated value function follows

$$\hat{v}_\eta(s) = \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \hat{\beta}_3 x_1^2 + \hat{\beta}_4 x_2^2.$$

According to Assumption 1(i), the two-pool system is symmetric, so the parameters $\{\hat{\beta}_i, i = 1, ..., 4\}$ for estimating $\hat{v}_\eta$ should also be symmetric, i.e.,

$$\hat{v}_\eta = \hat{\beta}_1(x_1 + x_2) + \hat{\beta}_3(x_1^2 + x_2^2).$$

To compute the closed form of cost-to-go $\mathbb{E}_{s'}[\hat{v}_\eta(s')]$, we need to specify the transition dynamics in our simplified two-pool setting. That is, given (s,f), the next state $s' = (x_1', x_2')$ follows

$$x_1' = x_1 - f_{1,2} + A_1 - D_1, \quad x_2' = x_2 + f_{1,2} + A_2 - D_2, \tag{20}$$

where from Assumption 1(i)(ii), the number of new arrivals A_1, A_2 both follow Poisson distribution with parameter λ, and the number of new departures D_1, D_2 follow distributions $Bin(x_1 - f_{1,2}, \mu)$ and $Bin(x_2 + f_{1,2}, \mu)$, respectively. Therefore, we have

$$\begin{aligned} & \mathbb{E}_{s' \sim p(\cdot|s,f)}[\hat{v}_\eta(s')] \\ =& \mathbb{E}_{s' \sim p(\cdot|s,f)}[\hat{\beta}_1(x_1' + x_2') + \hat{\beta}_3((x_1')^2 + (x_2')^2)] \\ =& \hat{\beta}_1\big(x_1 + x_2 + 2\lambda - (x_1 + x_2)\mu\big) + \hat{\beta}_3\mathbb{E}\big[(x_1 - f_{1,2} + A_1 - D_1)^2 + (x_2 + f_{1,2} + A_2 - D_2)^2\big]. \end{aligned} \tag{21}$$

Via some algebra to evaluate the expectation term in (21), we have

$$\mathbb{E}_{s' \sim p(\cdot|s,f)}[\hat{v}_\eta(s')] = \hat{\beta}^3(1-\mu)^2[(x_1 - f_{1,2})^2 + (x_2 + f_{1,2})^2] + (\hat{\beta}_1 + \hat{\beta}_3(2\lambda - \mu))(1-\mu)(x_1 + x_2) + 2\hat{\beta}_1\lambda + 2\hat{\beta}_3$$

Plugging the formulas of $g(s, f)$ and $\mathbb{E}[\hat{v}_\eta(s')]$ back into (19), we get

$$\begin{aligned} \hat{A}_\eta(s, f) =& g(s, f) - \gamma + \mathbb{E}_{s' \sim p(\cdot|s,f)}[\hat{v}_\eta(s')] - \hat{v}_\eta(s) \\ =& (B - C)f_{1,2} + \hat{\beta}_3(1-\mu)^2[2f_{1,2}^2 - 2(x_1 - x_2)f_{1,2}] + Const(s), \end{aligned} \tag{22}$$

where $Const(s)$ is a constant that depends on s but is independent of f. Finally, by plugging (22) into (18), we can rewrite the policy gradient $\nabla_0 \hat{N}_1(\theta, s)$ as

$$\begin{aligned} \nabla_0 \hat{N}_1(\theta, s) =& \sum_{f_{1,2}=1}^{q_1} \pi_\theta(f|s)\,(f_{1,2} - q_1\kappa_\theta(2|s,1))\,\hat{A}_\eta(s, \tilde{f}) \\ =& \sum_{f_{1,2}=1}^{q_1} \pi_\theta(f|s)\,(f_{1,2} - q_1\kappa_\theta(2|s,1))\,\Big((B - C)f_{1,2} + \hat{\beta}_3(1-\mu)^2[2f_{1,2}^2 - 2(x_1 - x_2)f_{1,2}] + Const(s) \\ =& q_1\kappa_\theta(2|s,1)\big(1 - \kappa_\theta(2|s,1)\big)\Big(2\hat{\beta}_3(1-\mu)^2\big(2(q_1 - 1)\kappa_\theta(2|s,1) + x_2 - x_1 + 1\big) + B - C\Big). \end{aligned}$$

Here, we have used the binomial distribution property for $f_{1,2}$ and we are able to elimiate the $Const(s)$ since

$$\sum_{f_{1,2}=1}^{q_1} \pi_\theta(f|s)\,(f_{1,2} - q_1\kappa_\theta(2|s,1))\,Const(s) = (\mathbb{E}[f_{1,2}] - q_1\kappa_\theta(2|s,1)) \cdot Const(s) = 0.$$

□

References

J. G. Dai and M. Gluzman. Queueing network controls via deep reinforcement learning. *Stochastic Systems*, 12(1):30–67, 2022.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability.* Springer Science & Business Media, 2012.