Technical Appendix: Proofs, Algorithms, and Supplementary Results

EC.1. Summary of Notation

A	consumption matrix, $A \coloneqq [a_{ij}]_{m\times n}$
A^j	j-th column of A, incidence vector for product j
$\mathcal{A}$	action space, collection of all subsets of $\mathcal{J}$
$\mathcal{A}(x)$	collection of all available assortments at the state x, $\mathcal{A}(x) \coloneqq \{S \in \mathcal{A} : x \geq A^j \text{ for all } j \in S\}$
$B([0,T]\times\mathcal{X})$	space of Borel measurable functions defined on $[0,T]\times\mathcal{X}$ such that $\|\psi\| \coloneqq \sup_{(t,x)\in[0,T]\times\mathcal{X}} \|\psi(t,x)\| < \infty$
c	initial inventory of m resources, $c = (c_1,\ldots,c_m)^\top$
$C^{1,0}([0,T]\times\mathcal{X})$	space of real-valued functions defined on $[0,T]\times\mathcal{X}$ that are continuously differentiable in t over $[0,T]$ for all $x \in \mathcal{X}$
d	degree of polynomial involved in parametric families of the value functions or policies
$\mathcal{F}$	rich enough σ-algebra defined on the sample space Ω
$\{\mathcal{F}_t\}_{t\geq 0}$	filtration generated by standard Poisson process N^λ and two sequences of i.i.d. uniform random variables for randomized actions and customer choices, respectively
$\{\mathcal{F}_t^{X^{\boldsymbol{\pi}}}\}_{t\geq 0}$	natural filtration generated by state process $X^{\boldsymbol{\pi}}$ under randomized policy $\boldsymbol{\pi}$
$\{\mathcal{F}_t^{\tilde{X}^{\boldsymbol{\pi}}}\}_{t\geq 0}$	natural filtration generated by exploratory state process $\tilde{X}^{\boldsymbol{\pi}}$
$\{\mathcal{G}_t\}_{t\geq 0}$	filtration generated by standard Poisson process N^λ and a sequences of i.i.d. uniform random variables for customer choices
$H(t,x,S,v(\cdot,\cdot))$	Hamiltonian of the classical intensity control problem
$\mathcal{H}(\boldsymbol{\pi}(\cdot\mid t,x))$	entropy of randomized policy $\boldsymbol{\pi}(\cdot\mid t,x)$
$J(t,x;\boldsymbol{\pi})$	entropy-regularized value function under policy $\boldsymbol{\pi}$ in continuous-time RL formulation
$J_{\Delta t}(t_k,x;\boldsymbol{\pi})$	entropy-regularized value function under policy $\boldsymbol{\pi}$ in discretization-based RL methods
$J^*(t,x)$	optimal value function in the continuous-time RL formulation
$J^\theta(t,x)$	parametric value function in continuous-time RL formulation
$J^\theta_{\Delta t}(t_k,x)$	parametric value function in discretization-based RL methods
$\mathcal{J}$	set of all products, $\mathcal{J} = \{1,\ldots,n\}$
K	total number of intervals in a uniform grid $0 = t_0 < t_1 < \cdots < t_K = T$ with step size $\Delta t = \frac{T}{K}$
$\mathbb{K}$	set of all valid time-state-action triples, $\mathbb{K} \coloneqq \{(t,x,S) : t \in [0,T],\ x \in \mathcal{X},\ S \in \mathcal{A}(x)\}$
L	total number of state jumps over $[0,T]$
$\bar{L}$	number of disjoint consideration sets for customers
m	number of resources
M	batch size in algorithms
$\mathcal{M}_{\Delta t}$	discrete-time model under a uniform grid $0 = t_0 < t_1 < \cdots < t_K = T$ with step size $\Delta t = \frac{T}{K}$, it is assumed that in each period $(t_k, t_{k+1}]$, one arrival occurs with probability $\lambda\Delta t$, and no arrivals occur with probability $1-\lambda\Delta t$
n	number of products
$\bar{n}$	total number of revenue-ordered assortments
N	number of episodes in algorithms
N^λ, N_t^λ	a standard Poisson process with rate λ, to model the arrival process of all potential customers
$N^{\boldsymbol{u}}$, $N_t^{\boldsymbol{u}}$	$N_t^{\boldsymbol{u}} = (N_{1,t}^{\boldsymbol{u}},\ldots,N_{n,t}^{\boldsymbol{u}})^\top$, a vector of Poisson processes with intensities $(\lambda P_1(S_t),\ldots,\lambda P_n(S_t))$ modulated by deterministic control $\boldsymbol{u} = \{S_t : t \in [0,T]\}$

$N^{\boldsymbol{\pi}}$, $N_t^{\boldsymbol{\pi}}$	$N_t^{\boldsymbol{\pi}} = (N_{1,t}^{\boldsymbol{\pi}}, \ldots, N_{n,t}^{\boldsymbol{\pi}})^\top$, with $N_{j,t}^{\boldsymbol{\pi}}$ representing the number of product j sold up to time t under the randomized policy $\boldsymbol{\pi}$
p	prices vector for n products, $p = (p_1, \ldots, p_n)^\top$
$P_0(S)$	probability that an arriving customer makes no purchase when assortment S is offered
$P_j(S)$	probability that an arriving customer purchases product j when assortment S is offered
$\mathbb{P}$	probability measure defined on $\mathcal{F}$
$q(y \mid t, x, S)$	transition rate of state process
$r(S)$	average revenue rate when offered S, $r(S) \coloneqq \lambda \sum_{j=1}^n p_j P_j(S)$
S	assortment offered by the firm, $S \subseteq \mathcal{J}$
$S_t^{\boldsymbol{\pi}}$	assortment offered at time t under policy $\boldsymbol{\pi}$
T	selling time limit
$\boldsymbol{u}$	deterministic control process, $\boldsymbol{u} = \{S_t \in \mathcal{A} : t \in [0,T]\}$
$\mathcal{U}$	set of all non-anticipating control processes
$V(t,x;\boldsymbol{u})$	value function under control $\boldsymbol{u}$ in the classical formulation
$V^*(t,x)$	optimal value function in the classical formulation
$V_{\Delta t}^*(t_k,x)$	optimal value function under the discrete-time model $\mathcal{M}_{\Delta t}$
$V_{\text{ADP-}\Delta t}(t_k,x)$	approximate value function derived from ADP under the discrete-time model $\mathcal{M}_{\Delta t}$, $V_{\text{ADP-}\Delta t}(t_k,x) = \theta_k + \sum_{i=1}^m V_{k,i} x_i$
$V_{\text{I-ADP-}\Delta t}(t,x)$	continuous-time interpolated approximated value function based on $V_{\text{ADP-}\Delta t}(t_k,x)$
$X^{\boldsymbol{u}}$, $X_t^{\boldsymbol{u}}$, $X_{t-}^{\boldsymbol{u}}$	process of the remaining inventory (state) under deterministic control $\boldsymbol{u}$, $X_{t-}^{\boldsymbol{u}}$ refers to the left limit of process $X^{\boldsymbol{u}}$ at time t
$X^{\boldsymbol{\pi}}$, $X_t^{\boldsymbol{\pi}}$, $X_{t-}^{\boldsymbol{\pi}}$	process of the remaining inventory (state) under randomized policy $\boldsymbol{\pi}$, $X_{t-}^{\boldsymbol{\pi}}$ refers to the left limit of process $X^{\boldsymbol{\pi}}$ at time t
$\tilde{X}^{\boldsymbol{\pi}}$, $\tilde{X}_t^{\boldsymbol{\pi}}$, $\tilde{X}_{t-}^{\boldsymbol{\pi}}$	exploratory state process defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P}; \{\mathcal{G}_t\}_{t\geq 0})$, it has the same distribution as the sample state process $\{X_t^{\boldsymbol{\pi}} : t \in [0,T]\}$ defined on $(\Omega, \mathcal{F}, \mathbb{P}; \{\mathcal{F}_t\}_{t\geq 0}))$
$\mathcal{X}$	state space, $\mathcal{X} = \{0, \ldots, c_1\} \times \cdots \times \{0, \ldots, c_m\}$
α_ϕ	learning rate for actor (policy) network
α_θ	learning rate for critic (value function) network
γ	entropy factor (temperature parameter)
λ	customer arrival rate
$\boldsymbol{\pi}(S \mid t,x)$	randomized policy
$\boldsymbol{\pi}^*(S \mid t,x)$	optimal randomized policy for the entropy-regularized optimization problem
$\boldsymbol{\pi}^\phi(S \mid t,x)$	parametric randomized policy
$\boldsymbol{\sigma}^\phi(S^{[j]} \mid t)$	parametric policy with action space that comprises all revenue-ordered assortments $S^{[1]}, \cdots, S^{[\bar{n}]}$
$\varphi(t,x)$	vector of basis functions for the linear value function approximation, $\varphi(t,x) \coloneqq (\varphi_1(t,x), \ldots, \varphi_W(t,x))^\top$
ϕ	parameters in actor (policy) network
θ	parameters in critic (value function) network
τ_l	l-th jump time the state process

EC.2. The Reformulation of (6) in Section 2.2

We first prove that for each $j = 1, \ldots, n$, the process $\{N_{j,t}^{\boldsymbol{\pi}} - \int_0^t \sum_{S \in \mathcal{A}(X_{t-}^{\boldsymbol{\pi}})} \lambda P_j(S) \boldsymbol{\pi}(S \mid s, X_{t-}^{\boldsymbol{\pi}}) \, \mathrm{d}t : t \in [0,T]\}$ is an $\{\mathcal{F}_t\}_{t\geq 0}$-martingale. Let $N_{(t,t+\Delta t]}^{\lambda}$ denote the number of customer arrivals in $(t, t+\Delta t]$; if there is exactly one customer arrival during this time interval, we denote the arrival time by t_a. Given $N_{(t,t+\Delta t]}^{\lambda} = 1$, the random variable t_a is uniformly distributed over $(t, t+\Delta t]$ by the property of the Poisson process N^λ. To

demonstrate the Markov property of $N^{\boldsymbol{\pi}}$ and derive its infinitesimal generator, for any bounded, measurable, real-valued function f, we evaluate $\mathbb{E}[f(N^{\boldsymbol{\pi}}_{t+\Delta t}) \mid N^{\boldsymbol{\pi}}_t = z, \mathcal{F}_t]$ as follows:

$$
\begin{aligned}
&\mathbb{E}[f(N^{\boldsymbol{\pi}}_{t+\Delta t}) \mid N^{\boldsymbol{\pi}}_t = z, \mathcal{F}_t] \\
&= \sum_{y\in\mathcal{X}} f(y)\cdot \mathbb{P}(N^{\boldsymbol{\pi}}_{t+\Delta t} = y \mid N^{\boldsymbol{\pi}}_t = z, \mathcal{F}_t) \\
&= \sum_{y\in\mathcal{X}} f(y)\cdot \Bigg\{ \mathbb{P}(N^{\boldsymbol{\pi}}_{t+\Delta t} = y \mid N^{\lambda}_{(t,t+\Delta t]} = 0,\, N^{\boldsymbol{\pi}}_t = z, \mathcal{F}_t)\cdot \mathbb{P}(N^{\lambda}_{(t,t+\Delta t]} = 0 \mid N^{\boldsymbol{\pi}}_t = z, \mathcal{F}_t) \\
&\quad + \left[\int_t^{t+\Delta t} \mathbb{P}(N^{\boldsymbol{\pi}}_{t+\Delta t} = y \mid t_a = u,\, N^{\lambda}_{(t,t+\Delta t]} = 1,\, N^{\boldsymbol{\pi}}_t = z, \mathcal{F}_t)\cdot \frac{1}{\Delta t}\,\mathrm{d}u\right]\cdot \mathbb{P}(N^{\lambda}_{(t,t+\Delta t]} = 1 \mid N^{\boldsymbol{\pi}}_t = z, \mathcal{F}_t) \\
&\quad + \mathbb{P}(N^{\boldsymbol{\pi}}_{t+\Delta t} = y \mid N^{\lambda}_{(t,t+\Delta t]} \geq 2,\, N^{\boldsymbol{\pi}}_t = z, \mathcal{F}_t)\cdot \mathbb{P}(N^{\lambda}_{(t,t+\Delta t]} \geq 2 \mid N^{\boldsymbol{\pi}}_t = z, \mathcal{F}_t)\Bigg\} \\
&= f(z)\cdot[1-\lambda\Delta t + o(\Delta t)] + \left[f(z)\left(\sum_{S\in\mathcal{A}(c-Az)} P_0(S)\frac{\int_t^{t+\Delta t}\boldsymbol{\pi}(S\mid u,\, c-Az)\,\mathrm{d}u}{\Delta t}\right)\right. \\
&\quad \left. + \sum_{j=1}^n f(z+\boldsymbol{e}^j)\left(\sum_{S\in\mathcal{A}(c-Az)} P_j(S)\frac{\int_t^{t+\Delta t}\boldsymbol{\pi}(S\mid u, c-Az)\,\mathrm{d}u}{\Delta t}\right)\right]\cdot[\lambda\Delta t + o(\Delta t)] + o(\Delta t) \\
&= f(z) + \lambda\Delta t\cdot\sum_{j=1}^n\left(\sum_{S\in\mathcal{A}(c-Az)} P_j(S)\frac{\int_t^{t+\Delta t}\boldsymbol{\pi}(S\mid u, c-Az)\,\mathrm{d}u}{\Delta t}\right)\cdot[f(z+\boldsymbol{e}^j) - f(z)] + o(\Delta t).
\end{aligned}
$$

Since the result depends only on the current state z, the process $N^{\boldsymbol{\pi}}$ satisfies the Markov property. Next, its infinitesimal generator is given by

$$
\lim_{\Delta t\to 0}\frac{\mathbb{E}[f(N^{\boldsymbol{\pi}}_{t+\Delta t}) \mid N^{\boldsymbol{\pi}}_t = z] - f(z)}{\Delta t} = \lambda\sum_{j=1}^n\sum_{S\in\mathcal{A}(c-Az)} P_j(S)\boldsymbol{\pi}(S\mid t, c-Az)\cdot[f(z+\boldsymbol{e}^j) - f(z)]. \quad \text{(EC.1)}
$$

Given that $N^{\boldsymbol{\pi}}$ has generator (EC.1) and $X^{\boldsymbol{\pi}}_t = c - AN^{\boldsymbol{\pi}}_t$ is bounded for all $t\in[0,T]$, it follows from Dynkin's formula that for each $j = 1,\ldots,n$, the process $\{N^{\boldsymbol{\pi}}_{j,t} - \int_0^t \sum_{S\in\mathcal{A}(X^{\boldsymbol{\pi}}_{t-})}\lambda P_j(S)\boldsymbol{\pi}(S\mid s, X^{\boldsymbol{\pi}}_{t-})\,\mathrm{d}t : t\in[0,T]\}$ is an $\{\mathcal{F}_t\}_{t\geq 0}$-martingale.

Next, we introduce the exploratory dynamics of $\{\tilde{X}^{\boldsymbol{\pi}}_t : t\in[0,T]\}$, which can be interpreted as the average of trajectories of the sample state process $\{X^{\boldsymbol{\pi}}_t : t\in[0,T]\}$ over infinitely many randomized actions. In the case of controlled diffusions, Wang et al. (2020) derived heuristically the exploratory state process by applying a law-of-large-numbers argument to the drift and diffusion coefficients of the controlled diffusion process. However, the sample state process $\{X^{\boldsymbol{\pi}}_t : t\in[0,T]\}$ in our setting is a pure jump process, so their approach does not apply. Instead, we derive the infinitesimal generator of the sample state process $\{X^{\boldsymbol{\pi}}_t : t\in[0,T]\}$, from which we will identify the dynamics of the exploratory state process.

From the above derivation for the generator of $N^{\boldsymbol{\pi}}$, it is evident that for $s<t$, given the present state $X^{\boldsymbol{\pi}}_s$, the increment $X^{\boldsymbol{\pi}}_t - X^{\boldsymbol{\pi}}_s = -A(N^{\boldsymbol{\pi}}_t - N^{\boldsymbol{\pi}}_s)$ does not depend on any other past information in $\mathcal{F}_s$. Therefore, the sample state process $\{X^{\boldsymbol{\pi}}_t : t\in[0,T]\}$ is also a Markov process. Using a similar derivation as for (EC.1), we

can obtain the generator of $\{X_t^{\boldsymbol{\pi}} : t \in [0,T]\}$. For each time-state pair (t,x) and any bounded and measurable function $g : \mathcal{X} \mapsto \mathbb{R}$,

$$\lim_{\Delta t \to 0} \frac{\mathbb{E}[g(X_{t+\Delta t}^{\boldsymbol{\pi}}) \mid X_t^{\boldsymbol{\pi}} = x] - g(x)}{\Delta t} = \lambda \sum_{j=1}^{n} \sum_{S \in \mathcal{A}(x)} P_j(S) \boldsymbol{\pi}(S \mid t,x) \cdot [g(x - A^j) - g(x)]. \tag{EC.2}$$

One can observe from (EC.2) that the effect of individually sampled actions has been averaged out. We can construct an ("averaged") exploratory process $\{\tilde{X}_t^{\boldsymbol{\pi}} : t \in [0,T]\}$, defined on the original probability space $(\Omega, \mathcal{F}, \mathbb{P}; \{\mathcal{G}_t\}_{t \geq 0})$, as a continuous-time Markov chain with the generator given by (EC.2) and $\tilde{X}_0^{\boldsymbol{\pi}} = c$. Because the generator and the initial state of $\{X_t^{\boldsymbol{\pi}} : t \in [0,T]\}$ and $\{\tilde{X}_t^{\boldsymbol{\pi}} : t \in [0,T]\}$ are identical, we infer that the sample state process $\{X_t^{\boldsymbol{\pi}} : t \in [0,T]\}$ has the same distribution as the distribution of the exploratory state process $\{\tilde{X}_t^{\boldsymbol{\pi}} : t \in [0,T]\}$. This completes the proof of (6).

EC.3. Actor-Critic Algorithms

EC.4. Proofs of Statements

Proof of Lemma 1 Let $B([0,T] \times \mathcal{X})$ denote the space of all Borel measurable functions $\psi : [0,T] \times \mathcal{X} \to \mathbb{R}$ such that $\|\psi\| \coloneqq \sup_{(t,x) \in [0,T] \times \mathcal{X}} |\psi(t,x)| < \infty$. This space becomes a Banach space when endowed with the supreme norm $\|\cdot\|$. We define an operator $\mathcal{L}$ on $B([0,T] \times \mathcal{X})$ by

$$\mathcal{L}\psi(t,x) \coloneqq e^{\beta t} \int_t^T \left\{ R^{\boldsymbol{\pi}}(s,x) + \sum_{S \in \mathcal{A}(x)} \left(\sum_{y \in \mathcal{X}} e^{-\beta s} \psi(s,y) q(y \mid s,x,S) \right) \boldsymbol{\pi}(S \mid s,x) \right\} \mathrm{d}s, \tag{EC.3}$$

where $R^{\boldsymbol{\pi}}(t,x)$ is as defined in (5), and $\beta \coloneqq 2\lambda + 1$. Given that the integrand on the right-hand side of (EC.3) also belongs to the space $B([0,T] \times \mathcal{X})$, it follows that the operator $\mathcal{L}$ is well-defined. Then, for $\psi_1, \psi_2 \in B([0,T] \times \mathcal{X})$, we obtain

$$\begin{aligned}
|\mathcal{L}\psi_1(t,x) - \mathcal{L}\psi_2(t,x)| &\leq e^{\beta t} \int_t^T e^{-\beta s} \left(\sum_{S \in \mathcal{A}(x)} \sum_{y \in \mathcal{X}} |\psi_1(s,y) - \psi_2(s,y)| \cdot |q(y \mid s,x,S)| \, \boldsymbol{\pi}(S \mid s,x) \right) \mathrm{d}s \\
&\leq e^{\beta t} \int_t^T e^{-\beta s} \left(\|\psi_1 - \psi_2\| \cdot \sum_{S \in \mathcal{A}(x)} \left(\sum_{y \in \mathcal{X}} |q(y \mid s,x,S)| \right) \boldsymbol{\pi}(S \mid s,x) \right) \mathrm{d}s \\
&\leq 2\lambda \, \|\psi_1 - \psi_2\| \, e^{\beta t} \int_t^T e^{-\beta s} \, \mathrm{d}s \\
&\leq \frac{2\lambda}{\beta} (1 - e^{-\beta(T-t)}) \, \|\psi_1 - \psi_2\| \\
&\leq \frac{2\lambda}{\beta} \, \|\psi_1 - \psi_2\| .
\end{aligned}$$

Observing that $\frac{2\lambda}{\beta} = \frac{2\lambda}{2\lambda+1} < 1$, we identify $\mathcal{L}$ as a contraction operator on the Banach space $B([0,T] \times \mathcal{X})$. Let $\psi^* \in B([0,T] \times \mathcal{X})$ be the fixed point of $\mathcal{L}$, i.e.,

$$\psi^*(t,x) \coloneqq e^{\beta t} \int_t^T \left\{ R^{\boldsymbol{\pi}}(s,x) + \sum_{S \in \mathcal{A}(x)} \left(\sum_{y \in \mathcal{X}} e^{-\beta s} \psi^*(s,y) q(y \mid s,x,S) \right) \boldsymbol{\pi}(S \mid s,x) \right\} \mathrm{d}s.$$

Algorithm EC.1 Actor-Critic Algorithm with Linear Value Function Approximation (PE via TD(0))

Inputs: initial state c, time horizon T, number of episodes N, batch size M; functional forms of basis functions $\varphi_1, \ldots, \varphi_W$, functional form of the policy $\boldsymbol{\pi}^{\phi}(\cdot \mid \cdot, \cdot)$ and an initial value ϕ_0; entropy factor γ, learning rate α_{ϕ}

Required program: an environment simulator $(t', x', S', r') = Environment(t, x, \boldsymbol{\pi}^{\phi}(\cdot \mid \cdot, \cdot))$

Initialize $\phi \leftarrow \phi_0$

for $i = 1$ to N **do**

  Store $(\tau_0^{(i)}, x_0^{(i)}) \leftarrow (0, c)$

  Initialize $l = 0$, $(t, x) = (0, c)$ ▷ Initialize l to count jumps in each state trajectory, and (t, x) to record the time and state right after each jump

  while True **do**

    Apply (t, x) to the environment simulator to get $(t', x', S', r') = Environment(t, x, \boldsymbol{\pi}^{\phi}(\cdot \mid \cdot, \cdot))$

    if $t' \geq T$ **then**

      Store $L^{(i)} \leftarrow l$, $\tau_{L^{(i)}+1}^{(i)} \leftarrow T$

      break

    end if

    Update $l \leftarrow l + 1$

    Store observation at jump time: $(\tau_l^{(i)}, x_l^{(i)}, S_l^{(i)}, r_l^{(i)}) \leftarrow (t', x', S', r')$

    Update $(t, x) \leftarrow (t', x')$

  end while

  Denote $G_l^{\varphi}(i) \coloneqq \varphi(\tau_l^{(i)}, x_{l-1}^{(i)})[\varphi(\tau_l^{(i)}, x_l^{(i)}) - \varphi(\tau_l^{(i)}, x_{l-1}^{(i)})]$, $F_l^{\varphi}(i) \coloneqq F^{\varphi}(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)})$

    $E_l^{\varphi}(\boldsymbol{\pi}^{\phi}, i) \coloneqq E(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)}; \varphi(\cdot, x_l^{(i)}), \boldsymbol{\pi}^{\phi})$, $E_l^{\mathbf{1}}(\boldsymbol{\pi}^{\phi}, i) = E(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)}; \mathbf{1}, \boldsymbol{\pi}^{\phi})$

    $U_l^{\theta^*}(\boldsymbol{\pi}^{\phi}, i) \coloneqq \log \boldsymbol{\pi}^{\phi}(S_l^{(i)} \mid \tau_l^{(i)}, x_{l-1}^{(i)})[J^{\theta^*}(\tau_l^{(i)}, x_l^{(i)}) - J^{\theta^*}(\tau_l^{(i)}, x_{l-1}^{(i)}) + r_l^{(i)}]$

  if $i = 0 \pmod{M}$ **then** ▷ Perform an update using M episodes generated under the policy

    Evaluate policy $\boldsymbol{\pi}^{\phi}$: [using formula (19), incorporating techniques (20) and (21)]

$$\theta^* = \left[\frac{1}{M}\sum_{k=i-M+1}^{i}\left(\sum_{l=1}^{L^{(k)}} G_l^{\varphi}(k) + \sum_{l=0}^{L^{(k)}} F_l^{\varphi}(k)\right)\right]^{(-1)} \times \left[\frac{1}{M}\sum_{k=i-M+1}^{i}\left(\sum_{l=1}^{L^{(k)}} \varphi(\tau_l^{(k)}, x_{l-1}^{(k)}) r_{\tau_l}^{(k)} + \gamma \sum_{l=0}^{L^{(k)}} E_l^{\varphi}(\boldsymbol{\pi}^{\phi}, k)\right)\right]$$

    Compute policy gradient at ϕ: [using formula (25)]

$$\Delta\phi = \nabla_{\phi}\left\{\frac{1}{M}\sum_{k=i-M+1}^{i}\left(\sum_{l=1}^{L^{(k)}} U_l^{\theta^*}(\boldsymbol{\pi}^{\phi}, k) + \gamma \sum_{l=0}^{L^{(k)}} E_l^{\mathbf{1}}(\boldsymbol{\pi}^{\phi}, k)\right)\right\}$$

    Update ϕ by

$$\phi \leftarrow \phi + \alpha_{\phi}\Delta\phi$$

  end if

end for

Algorithm EC.2 Actor-Critic Algorithm with Neural Networks (PE via Monte Carlo)

Inputs: initial state c, time horizon T, number of episodes N, batch size M; critic network $J^\theta(t,x)$ and an initial value θ_0; actor network $\boldsymbol{\pi}^\phi(S \mid t,x)$ and an initial value ϕ_0; entropy factor γ, learning rates α_θ, α_ϕ

Required program: an environment simulator $(t',x',S',r') = Environment(t,x,\boldsymbol{\pi}^\phi(\cdot \mid \cdot,\cdot))$

Initialize $\theta \leftarrow \theta_0$, $\phi \leftarrow \phi_0$

for $i = 1$ to N **do**

  Store $(\tau_0^{(i)}, x_0^{(i)}) \leftarrow (0,c)$

  Initialize $l = 0$, $(t,x) = (0,c)$ ▷ Initialize l to count jumps in each state trajectory, and (t,x) to record the time and state right after each jump

  while True **do**

    Apply (t,x) to the environment simulator to get $(t',x',S',r') = Environment(t,x,\boldsymbol{\pi}^\phi(\cdot \mid \cdot,\cdot))$

    if $t' \geq T$ **then**

      Store $L^{(i)} \leftarrow l$, $\tau_{L^{(i)}+1}^{(i)} \leftarrow T$

      break

    end if

    Update $l \leftarrow l+1$

    Store observation at jump time: $(\tau_l^{(i)}, x_l^{(i)}, S_l^{(i)}, r_l^{(i)}) \leftarrow (t',x',S',r')$

    Update $(t,x) \leftarrow (t',x')$

  end while

  Denote $D_l(\theta,i) \coloneqq D(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)}; \theta)$, $b_l(\theta,i) \coloneqq b(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)}; \theta)$

    $C_l(\boldsymbol{\pi}^\phi, i) \coloneqq \sum_{l'=l+1}^{L^{(i)}} [r_{l'}^{(i)} + \gamma E(\tau_{l'}^{(i)}, \tau_{l'+1}^{(i)}, x_{l'}^{(i)}; \mathbf{1}, \boldsymbol{\pi}^\phi)]$

    $E_l(\theta, \boldsymbol{\pi}^\phi, i) \coloneqq E(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)}; b(\tau_l^{(i)}, \cdot, x_l^{(i)}; \theta), \boldsymbol{\pi}^\phi)$, $E_l^{\mathbf{1}}(\boldsymbol{\pi}^\phi, i) = E(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)}; \mathbf{1}, \boldsymbol{\pi}^\phi)$

    $U_l^\theta(\boldsymbol{\pi}^\phi, i) \coloneqq \log \boldsymbol{\pi}^\phi(S_l^{(i)} \mid \tau_l^{(i)}, x_{l-1}^{(i)})[J^\theta(\tau_l^{(i)}, x_l^{(i)}) - J^\theta(\tau_l^{(i)}, x_{l-1}^{(i)}) + r_l^{(i)}]$

  if $i = 0 \pmod M$ **then** ▷ Perform an update using M episodes generated under the policy

    Update critic network: [using formula (14)]

$$\theta \leftarrow \theta - \alpha_\theta \nabla_\theta \left\{ \frac{1}{M} \sum_{k=i-M+1}^{i} \sum_{l=0}^{L^{(k)}} \left(\frac{1}{2} D_l(\theta,k) - b_l(\theta,k) C_l(\boldsymbol{\pi}^\phi, k) - \gamma E_l(\theta, \boldsymbol{\pi}^\phi, k) \right) \right\}$$

    Update actor network: [using formula (25)]

$$\phi \leftarrow \phi + \nabla_\phi \left\{ \frac{1}{M} \sum_{k=i-M+1}^{i} \left(\sum_{l=1}^{L^{(k)}} U_l^\theta(\boldsymbol{\pi}^\phi, k) + \gamma \sum_{l=0}^{L^{(k)}} E_l^{\mathbf{1}}(\boldsymbol{\pi}^\phi, k) \right) \right\}$$

  end if

end for

Let $v(t,x) = e^{-\beta t}\psi^*(t,x)$ for all $(t,x) \in [0,T]\times\mathcal{X}$, then v is continuously differentiable in t, i.e. $v \in C^{1,0}([0,T]\times\mathcal{X})$, and satisfies (7). Thus, we establish the existence of the solution to (7).

Suppose $v^* \in C^{1,0}([0,T]\times\mathcal{X})$ is a solution to (7). One can readily see that the transition rates of the Markov process $\tilde{X}^{\boldsymbol{\pi}}$, introduced in Section EC.2, at time s is given by

$$\sum_{S\in\mathcal{A}(\tilde{X}^{\boldsymbol{\pi}}_{s-})} q(y \mid s, \tilde{X}^{\boldsymbol{\pi}}_{s-}, S)\boldsymbol{\pi}(S \mid s, \tilde{X}^{\boldsymbol{\pi}}_{s-}), \quad y \neq \tilde{X}^{\boldsymbol{\pi}}_{s-}.$$

It follows from Theorem 3.1 in Guo et al. (2015) that we have the Dynkin's formula:

$$\begin{aligned}
&\mathbb{E}\left[v^*(T,\tilde{X}^{\boldsymbol{\pi}}_T) \mid \tilde{X}^{\boldsymbol{\pi}}_t = x\right] - v^*(t,x)\\
&= \mathbb{E}\left[\int_t^T \left(\frac{\partial v^*}{\partial s}(s,\tilde{X}^{\boldsymbol{\pi}}_{s-}) + \sum_{y\in\mathcal{X}} v^*(s,y)\Big[\sum_{S\in\mathcal{A}(\tilde{X}^{\boldsymbol{\pi}}_{s-})} q(y \mid s, \tilde{X}^{\boldsymbol{\pi}}_{s-}, S)\boldsymbol{\pi}(S \mid s, \tilde{X}^{\boldsymbol{\pi}}_{s-})\Big]\right)\mathrm{d}s \mid \tilde{X}^{\boldsymbol{\pi}}_t = x\right].
\end{aligned}$$

Since v^* satisfies Equation (7) with $v^*(T,x) = 0$, we then infer that

$$\begin{aligned}
v^*(t,x) &= \mathbb{E}\left[\int_t^T \left\{\sum_{S\in\mathcal{A}(\tilde{X}^{\boldsymbol{\pi}}_{s-})} r(S)\boldsymbol{\pi}(S \mid s, \tilde{X}^{\boldsymbol{\pi}}_{s-}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s, \tilde{X}^{\boldsymbol{\pi}}_{s-}))\right\}\mathrm{d}s \mid \tilde{X}^{\boldsymbol{\pi}}_t = x\right]\\
&= J(t,x;\boldsymbol{\pi}), \quad (t,x) \in [0,T]\times\mathcal{X}.
\end{aligned}$$

The proof is therefore complete.

Proof of the optimal stochastic policy (4) *in Section 2.2* Let the space $B([0,T]\times\mathcal{X})$ and the endowed norm $\|\cdot\|$ be as specified in the proof of Lemma 1. We then define an operator $\overline{\mathcal{L}}$ on $B([0,T]\times\mathcal{X})$ by

$$\overline{\mathcal{L}}\psi(t,x) := e^{\beta t}\int_t^T \sup_{\boldsymbol{\pi}\in\Pi}\left\{R^{\boldsymbol{\pi}}(s,x) + \sum_{S\in\mathcal{A}(x)}\left(\sum_{y\in\mathcal{X}} e^{-\beta s}\psi(s,y)q(y \mid s,x,S)\right)\boldsymbol{\pi}(S \mid s,x)\right\}\mathrm{d}s,$$

where $\beta := 2\lambda + 1$. Using a similar argument as in the proof of Lemma 1, we can identify $\overline{\mathcal{L}}$ as a contraction operator on the Banach space $B([0,T]\times\mathcal{X})$ and then establish the existence of the solution in the space $C^{1,0}([0,T]\times\mathcal{X})$ to the following exploratory HJB equation:

$$\begin{cases}
\dfrac{\partial v}{\partial t}(t,x) + \sup\limits_{\boldsymbol{\pi}\in\Pi}\left\{\sum\limits_{S\in\mathcal{A}(x)}[H(t,x,S,v(\cdot,\cdot)) - \gamma\log\boldsymbol{\pi}(S \mid t,x)]\boldsymbol{\pi}(S \mid t,x)\right\} = 0, \quad (t,x) \in [0,T)\times\mathcal{X}\\
v(T,x) = 0, \quad x \in \mathcal{X}.
\end{cases} \tag{EC.4}$$

Assume $v^* \in C^{1,0}([0,T]\times\mathcal{X})$ is a solution to (EC.4), then $v^*(T,x) = 0$ for all $x \in \mathcal{X}$ and the following inequality holds for all $\boldsymbol{\pi} \in \Pi$ and $(t,x) \in [0,T]\times\mathcal{X}$:

$$\frac{\partial v^*}{\partial t}(t,x) + \sum_{S\in\mathcal{A}(x)}\{H(t,x,S,v^*(\cdot,\cdot)) - \gamma\log\boldsymbol{\pi}(S \mid t,x)\}\boldsymbol{\pi}(S \mid t,x) \leq 0. \tag{EC.5}$$

Then, it follows from Theorem 3.1 in Guo et al. (2015) that

$$
\begin{aligned}
& v^*(t,x) \\
&= \mathbb{E}\left[v^*(T,\tilde{X}_T^{\boldsymbol{\pi}}) \mid \tilde{X}_t^{\boldsymbol{\pi}} = x\right] \\
&\quad - \mathbb{E}\left[\int_t^T \left(\frac{\partial v^*}{\partial s}(s,\tilde{X}_{s-}^{\boldsymbol{\pi}}) + \sum_{y\in\mathcal{X}} v^*(s,y)\Big[\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^{\boldsymbol{\pi}})} q(y \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}},S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}})\Big]\right)\mathrm{d}s \mid \tilde{X}_t^{\boldsymbol{\pi}} = x\right] \\
&\geq \mathbb{E}\left[\int_t^T \sum_{S\in\mathcal{A}(\tilde{X}_{s-}^{\boldsymbol{\pi}})} \{r(S) - \gamma\log\boldsymbol{\pi}(S \mid t,\tilde{X}_{s-}^{\boldsymbol{\pi}})\}\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}})\,\mathrm{d}s \mid \tilde{X}_t^{\boldsymbol{\pi}} = x\right] \\
&= J(t,x;\boldsymbol{\pi}), \qquad \text{for all } \boldsymbol{\pi}\in\Pi,\ (t,x)\in[0,T]\times\mathcal{X}.
\end{aligned}
$$

Moreover, the equality in (EC.5) holds for policy $\boldsymbol{\pi}^*$ defined as follows:

$$
\boldsymbol{\pi}^*(S \mid t,x) = \frac{\exp\{\frac{1}{\gamma}H(t,x,S,v^*(\cdot,\cdot))\}}{\sum_{S\in\mathcal{A}(x)}\exp\{\frac{1}{\gamma}H(t,x,S,v^*(\cdot,\cdot))\}}, \qquad \text{for } (t,x,S)\in\mathbb{K}.
$$

Thus, we conclude that $v^*(t,x) = J(t,x;\boldsymbol{\pi}^*) = J^*(t,x)$ for all $(t,x)\in[0,T]\times\mathcal{X}$. This establishes the characterization of the optimal policy in (4).

Proof of Proposition 1 Denote

$$
\tilde{M}_t := J(t,\tilde{X}_t^{\boldsymbol{\pi}};\boldsymbol{\pi}) + \int_0^t \left\{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^{\boldsymbol{\pi}})} r(S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}}))\right\}\mathrm{d}s, \quad t\in[0,T].
$$

We then show that $\{\tilde{M}_t : t\in[0,T]\}$ is an $\{\mathcal{F}_t^{\tilde{X}^{\boldsymbol{\pi}}}\}_{t\geq 0}$-martingale. We integrate the expression for $J(t,x;\boldsymbol{\pi})$ in (6), yielding

$$
\begin{aligned}
\tilde{M}_t = \mathbb{E}&\left[\int_t^T \left\{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^{\boldsymbol{\pi}})} r(S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}}))\right\}\mathrm{d}s \mid \tilde{X}_t^{\boldsymbol{\pi}}\right] \\
&+ \left[\int_0^t \left\{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^{\boldsymbol{\pi}})} r(S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}}))\right\}\mathrm{d}s\right].
\end{aligned} \tag{EC.6}
$$

Due to the Markov property of $\{\tilde{X}_t^{\boldsymbol{\pi}} : t\in[0,T]\}$, and also note that the second term on the right side of equation (EC.6) is $\mathcal{F}_t^{\tilde{X}^{\boldsymbol{\pi}}}$-measurable, we obtain

$$
\tilde{M}_t = \mathbb{E}\left[\int_0^T \left\{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^{\boldsymbol{\pi}})} r(S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s,\tilde{X}_{s-}^{\boldsymbol{\pi}}))\right\}\mathrm{d}s \mid \mathcal{F}_t^{\tilde{X}^{\boldsymbol{\pi}}}\right].
$$

Since $J(T,x;\boldsymbol{\pi}) = 0$ for all $x\in\mathcal{X}$, we conclude that $\tilde{M}_t = \mathbb{E}[\tilde{M}_T \mid \mathcal{F}_t^{\tilde{X}^{\boldsymbol{\pi}}}]$, which implies $\{\tilde{M}_t : t\in[0,T]\}$ is an $\{\mathcal{F}_t^{\tilde{X}^{\boldsymbol{\pi}}}\}_{t\geq 0}$-martingale. Moreover, since $\tilde{M}_t$ is bounded for any fixed time $t\in[0,T]$, it follows that $\{\tilde{M}_t : t\in[0,T]\}$ is square-integrable.

Conversely, assume that $\{\tilde{M}_t^v : t \in [0,T]\}$ is an $\{\mathcal{F}_t^{\tilde{X}^\pi}\}_{t\geq 0}$-martingale and $v(T,x) = 0$ for all $x \in \mathcal{X}$. From this, it follows directly that $\tilde{M}_t^v = \mathbb{E}[\tilde{M}_T^v \mid \mathcal{F}_t^{\tilde{X}^\pi}]$ for all $t \in [0,T]$, which is equivalent to

$$\begin{aligned} v(t,\tilde{X}_t^\pi) =& \mathbb{E}\left[\int_0^T \left\{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s,\tilde{X}_{s-}^\pi))\right\} \mathrm{d}s \mid \mathcal{F}_t^{\tilde{X}^\pi}\right] \\ &- \left[\int_0^t \left\{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s,\tilde{X}_{s-}^\pi))\right\} \mathrm{d}s\right], \quad t \in [0,T]. \end{aligned} \tag{EC.7}$$

Since the second term on the right side of Equation (EC.7) is $\mathcal{F}_t^{\tilde{X}^\pi}$-measurable, then it follows from the Markov property of $\{\tilde{X}_t^\pi : t \in [0,T]\}$ that

$$\begin{aligned} v(t,\tilde{X}_t^\pi) &= \mathbb{E}\left[\int_t^T \left\{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s,\tilde{X}_{s-}^\pi))\right\} \mathrm{d}s \mid \mathcal{F}_t^{\tilde{X}^\pi}\right] \\ &= \mathbb{E}\left[\int_t^T \left\{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s,\tilde{X}_{s-}^\pi))\right\} \mathrm{d}s \mid \tilde{X}_t^\pi\right] \\ &= J(t,\tilde{X}_t^\pi;\boldsymbol{\pi}), \quad t \in [0,T]. \end{aligned}$$

Therefore, we conclude that $v(t,x) = J(t,x;\boldsymbol{\pi})$ for all $(t,x) \in [0,T]\times\mathcal{X}$.

Proof of Theorem 1 Denote $U_t \coloneqq \int_{(t,T]} p^\top \mathrm{d}N_s^\pi - \int_t^T \sum_{S\in\mathcal{A}(X_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s, X_{s-}^\pi)\,\mathrm{d}s$, then we have

$$\mathbb{E}\left[U_t \mid X_t^\pi\right] = 0. \tag{EC.8}$$

We further denote $M_t \coloneqq J(t,X_t^\pi;\boldsymbol{\pi}) + \int_0^t \{\sum_{S\in\mathcal{A}(X_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s, X_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s, X_{s-}^\pi))\}\,\mathrm{d}s$ and $M_t^\theta \coloneqq J^\theta(t,X_t^\pi) + \int_0^t \{\sum_{S\in\mathcal{A}(X_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s, X_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s, X_{s-}^\pi))\}\,\mathrm{d}s$. It follows from $J(T,x;\boldsymbol{\pi}) \equiv 0$ and (EC.8) that

$$\begin{aligned} 2L(\theta) &= \mathbb{E}\left[\int_0^T (U_t + M_T - M_t^\theta)^2\,\mathrm{d}t\right] \\ &= \mathbb{E}\left[\int_0^T [U_t^2 + 2U_t(M_T - M_t^\theta) + (M_T - M_t^\theta)^2]\,\mathrm{d}t\right] \\ &= \mathbb{E}\left[\int_0^T \left[U_t^2 + 2U_t\left(\int_t^T \left\{\sum_{S\in\mathcal{A}(X_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s, X_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s, X_{s-}^\pi))\right\}\mathrm{d}s\right)\right]\mathrm{d}t\right] \\ &\quad - \mathbb{E}\left[\int_0^T 2J^\theta(t,X_t^\pi)\mathbb{E}[U_t \mid X_t^\pi]\,\mathrm{d}t\right] + \mathbb{E}\left[\int_0^T (M_T - M_t^\theta)^2\,\mathrm{d}t\right] \\ &= \mathbb{E}\left[\int_0^T \left[U_t^2 + 2U_t\left(\int_t^T \left\{\sum_{S\in\mathcal{A}(X_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s, X_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s, X_{s-}^\pi))\right\}\mathrm{d}s\right)\right]\mathrm{d}t\right] \\ &\quad + \mathbb{E}\left[\int_0^T (M_T - M_t^\theta)^2\,\mathrm{d}t\right]. \end{aligned}$$

Since the first term does not rely on θ, we have

$$\arg\min_\theta L(\theta) = \arg\min_\theta \mathbb{E}\left[\int_0^T (M_T - M_t^\theta)^2 \,\mathrm{d}t\right]. \tag{EC.9}$$

Next, we denote $\tilde{M}_t := J(t, \tilde{X}_t^\pi; \boldsymbol{\pi}) + \int_0^t \{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s, \tilde{X}_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s, \tilde{X}_{s-}^\pi))\}\,\mathrm{d}s$ and $\tilde{M}_t^\theta := J^\theta(t, \tilde{X}_t^\pi) + \int_0^t \{\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^\pi)} r(S)\boldsymbol{\pi}(S \mid s, \tilde{X}_{s-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s, \tilde{X}_{s-}^\pi))\}\,\mathrm{d}s$. From Proposition 1 we know that $\{\tilde{M}_t : t \in [0,T]\}$ is an $\{\mathcal{F}_t^{\tilde{X}^\pi}\}_{t\geq 0}$-martingale. Therefore,

$$\begin{aligned}
\mathbb{E}\left[\int_0^T (\tilde{M}_T - \tilde{M}_t^\theta)^2 \,\mathrm{d}t\right] &= \mathbb{E}\left[\int_0^T (\tilde{M}_T - \tilde{M}_t + \tilde{M}_t - \tilde{M}_t^\theta)^2 \,\mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_0^T [(\tilde{M}_T - \tilde{M}_t)^2 + (\tilde{M}_t - \tilde{M}_t^\theta)^2 + 2(\tilde{M}_T - \tilde{M}_t)(\tilde{M}_t - \tilde{M}_t^\theta)] \,\mathrm{d}t\right] \\
&= \mathbb{E}\left[\int_0^T (\tilde{M}_T - \tilde{M}_t)^2 \,\mathrm{d}t\right] + \mathbb{E}\left[\int_0^T (\tilde{M}_t - \tilde{M}_t^\theta)^2 \,\mathrm{d}t\right] \\
&\quad + 2\int_0^T \mathbb{E}\left[(\tilde{M}_t - \tilde{M}_t^\theta)\cdot \mathbb{E}\big[\tilde{M}_T - \tilde{M}_t \mid \mathcal{F}_t^{\tilde{X}^\pi}\big]\right] \mathrm{d}t \\
&= \mathbb{E}\left[\int_0^T (\tilde{M}_T - \tilde{M}_t)^2 \,\mathrm{d}t\right] + \mathbb{E}\left[\int_0^T |J(t, \tilde{X}_t^\pi; \boldsymbol{\pi}) - J^\theta(t, \tilde{X}_t^\pi)|^2 \,\mathrm{d}t\right].
\end{aligned}$$

Noting that the first term does not rely on θ, we obtain

$$\arg\min_\theta \mathbb{E}\left[\int_0^T |\tilde{M}_T - \tilde{M}_t^\theta|^2 \,\mathrm{d}t\right] = \arg\min_\theta \mathbb{E}\left[\int_0^T |J(t, \tilde{X}_t^\pi; \boldsymbol{\pi}) - J^\theta(t, \tilde{X}_t^\pi)|^2 \,\mathrm{d}t\right].$$

Since the distribution of $\{\tilde{X}_t^\pi : t \in [0,T]\}$ is the same as the distribution of $\{X_t^\pi : t \in [0,T]\}$ together with (EC.9), we conclude that $\arg\min_\theta L(\theta) = \arg\min_\theta \mathrm{MSVE}(\theta)$.

Proof of Theorem 2 Note that a bounded process ξ with $\xi_t \in \mathcal{F}_{t-}^{X^\pi}$ is also $\{\mathcal{F}_t\}_{t\geq 0}$-predictable. On the other hand, for each $j = 1, \cdots, n$, the process $\{N_{j,t}^\pi - \int_0^t \sum_{S\in\mathcal{A}(X_{t-}^\pi)} \lambda P_j(S)\boldsymbol{\pi}(S \mid s, X_{t-}^\pi)\,\mathrm{d}t : t \in [0,T]\}$ is an $\{\mathcal{F}_t\}_{t\geq 0}$-martingale. Thus we have

$$\mathbb{E}\left[\int_0^T \xi_t (p^\top \mathrm{d}N_t^\pi)\right] = \mathbb{E}\left[\int_0^T \xi_t \Big[\sum_{S\in\mathcal{A}(X_{t-}^\pi)} r(S)\boldsymbol{\pi}(S \mid t, X_{t-}^\pi)\Big] \mathrm{d}t\right].$$

It suffices to prove that $v(t,x) = J(t,x;\boldsymbol{\pi})$ for all $(t,x) \in [0,T]\times\mathcal{X}$, if and only if v satisfies $v(T,x) = 0$ for all $x \in \mathcal{X}$, and

$$\mathbb{E}\left[\int_0^T \xi_t \left\{\mathrm{d}v(t, X_t^\pi) + \Big[\sum_{S\in\mathcal{A}(X_{t-}^\pi)} r(S)\boldsymbol{\pi}(S \mid t, X_{t-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid t, X_{t-}^\pi))\Big] \mathrm{d}t\right\}\right] \tag{EC.10}$$

for any bounded process ξ with $\xi_t \in \mathcal{F}_{t-}^{X^\pi}$.

We first establish that $v(t,x) = J(t,x;\boldsymbol{\pi})$ for all $(t,x) \in [0,T]\times\mathcal{X}$, if and only if v satisfies $v(T,x) = 0$ for all $x \in \mathcal{X}$, and

$$\mathbb{E}\left[\int_0^T \tilde{\xi}_t \left\{\mathrm{d}v(t, \tilde{X}_t^\pi) + \Big[\sum_{S\in\mathcal{A}(\tilde{X}_{t-}^\pi)} r(S)\boldsymbol{\pi}(S \mid t, \tilde{X}_{t-}^\pi) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid t, \tilde{X}_{t-}^\pi))\Big] \mathrm{d}t\right\}\right] = 0, \tag{EC.11}$$

for any bounded process $\tilde{\xi}$ with $\tilde{\xi}_t \in \mathcal{F}_{t-}^{\tilde{X}^{\pi}}$. The "only if" part follows immediately from Proposition 1. To establish the "if" part, assume that $v(T,x)=0$ for all $x\in\mathcal{X}$, and (EC.11) holds for any bounded process $\tilde{\xi}$ with $\tilde{\xi}_t \in \mathcal{F}_{t-}^{\tilde{X}^{\pi}}$. It follows from Theorem 3.1 of Guo et al. (2015) that the process

$$v(t,\tilde{X}_t^{\pi}) - \int_0^t \left(\frac{\partial v}{\partial s}(s,\tilde{X}_{s-}^{\pi}) + \sum_{y\in\mathcal{X}} v(s,y)\Big[\sum_{S\in\mathcal{A}(\tilde{X}_{s-}^{\pi})} q(y \mid s,\tilde{X}_{s-}^{\pi},S)\boldsymbol{\pi}(S \mid s,\tilde{X}_{s-}^{\pi})\Big]\right)\mathrm{d}s$$

defines an $\{\mathcal{F}_t^{\tilde{X}^{\pi}}\}_{t\geq 0}$-martingale, and its boundedness at any fixed time $t\in[0,T]$ ensures square-integrability. Then, it follows that

$$\mathbb{E}\left[\int_0^T \tilde{\xi}_t\left\{dv(t,\tilde{X}_t^{\pi}) - \left(\frac{\partial v}{\partial t}(t,\tilde{X}_{t-}^{\pi}) + \sum_{y\in\mathcal{X}} v(t,y)\Big[\sum_{S\in\mathcal{A}(\tilde{X}_{t-}^{\pi})} q(y \mid t,\tilde{X}_{t-}^{\pi},S)\boldsymbol{\pi}(S \mid t,\tilde{X}_{t-}^{\pi})\Big]\right)\mathrm{d}t\right\}\right] = 0, \tag{EC.12}$$

for any bounded process $\tilde{\xi}$ with $\tilde{\xi}_t \in \mathcal{F}_{t-}^{\tilde{X}^{\pi}}$. By taking the difference between equations (EC.11) and (EC.12), we obtain that

$$\mathbb{E}\left[\int_0^T \tilde{\xi}_t\left(\frac{\partial v}{\partial t}(t,\tilde{X}_{t-}^{\pi}) + \sum_{S\in\mathcal{A}(\tilde{X}_{t-}^{\pi})} H(t,\tilde{X}_{t-}^{\pi},S,v(\cdot,\cdot))\boldsymbol{\pi}(S \mid t,\tilde{X}_{t-}^{\pi}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid t,\tilde{X}_{t-}^{\pi}))\right)\mathrm{d}t\right] = 0 \tag{EC.13}$$

holds for any bounded process $\tilde{\xi}$ with $\tilde{\xi}_t \in \mathcal{F}_{t-}^{\tilde{X}^{\pi}}$. Define the test function $\tilde{\xi}_t$ as follows:

$$\tilde{\xi}_t = \operatorname{sgn}\left(\frac{\partial v}{\partial t}(t,\tilde{X}_{t-}^{\pi}) + \sum_{S\in\mathcal{A}(\tilde{X}_{t-}^{\pi})} H(t,\tilde{X}_{t-}^{\pi},S,v(\cdot,\cdot))\boldsymbol{\pi}(S \mid t,\tilde{X}_{t-}^{\pi}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid t,\tilde{X}_{t-}^{\pi}))\right),$$

where $\operatorname{sgn}(\cdot)$ is the sign function. Then, (EC.13) implies that

$$\frac{\partial v}{\partial t}(t,\tilde{X}_{t-}^{\pi}) + \sum_{S\in\mathcal{A}(\tilde{X}_{t-}^{\pi})} H(t,\tilde{X}_{t-}^{\pi},S,v(\cdot,\cdot))\boldsymbol{\pi}(S \mid t,\tilde{X}_{t-}^{\pi}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid t,\tilde{X}_{t-}^{\pi})) = 0, \quad t\in[0,T]. \tag{EC.14}$$

It follows from Lemma 1 that $v(t,x)=J(t,x;\boldsymbol{\pi})$ for all $(t,x)\in[0,T]\times\mathcal{X}$. Hence, we complete the proof of the equivalent condition (EC.11).

Let $D[0,T]$ denote the space of all mappings $f:[0,T]\mapsto\mathcal{X}$. For any process $Y\in D[0,T]$ and fixed $t\in[0,T]$, define the stopped process $Y_{t-\wedge\cdot}^{\pi}\in D[0,T]$ such that $Y_{t-\wedge\cdot}^{\pi}(s)=Y_s^{\pi}$ for $s\in[0,t)$, and $Y_{t-\wedge\cdot}^{\pi}(s)=Y_{t-}^{\pi}$ for $s\in[t,T]$. Note that any process ξ with $\xi_t\in\mathcal{F}_{t-}^{X^{\pi}}$ corresponds to a measurable function $\boldsymbol{\xi}:[0,T]\times D([0,T])\mapsto\mathbb{R}$ such that $\xi_t=\boldsymbol{\xi}(t,X_{t-\wedge\cdot}^{\pi})$. Then, $\boldsymbol{\xi}(t,\tilde{X}_{t-\wedge\cdot}^{\pi})$ is $\mathcal{F}_{t-}^{\tilde{X}^{\pi}}$-measurable for each $t\in[0,T]$. Since the distribution of $\{X_t^{\pi}:t\in[0,T]\}$ is the same as the distribution of $\{\tilde{X}_t^{\pi}:t\in[0,T]\}$, it follows that

$$\begin{aligned}
&\mathbb{E}\left[\int_0^T \boldsymbol{\xi}(t,X_{t-\wedge\cdot}^{\pi})\left\{\mathrm{d}v(t,X_t^{\pi}) + \Big[\sum_{S\in\mathcal{A}(X_{t-}^{\pi})} r(S)\boldsymbol{\pi}(S \mid t,X_{t-}^{\pi}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid t,X_{t-}^{\pi}))\Big]\mathrm{d}t\right\}\right] \\
=&\,\mathbb{E}\left[\int_0^T \boldsymbol{\xi}(t,\tilde{X}_{t-\wedge\cdot}^{\pi})\left\{\mathrm{d}v(t,\tilde{X}_t^{\pi}) + \Big[\sum_{S\in\mathcal{A}(\tilde{X}_{t-}^{\pi})} r(S)\boldsymbol{\pi}(S \mid t,\tilde{X}_{t-}^{\pi}) + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot \mid t,\tilde{X}_{t-}^{\pi}))\Big]\mathrm{d}t\right\}\right].
\end{aligned} \tag{EC.15}$$

Conversely, any process $\tilde{\xi}$ with $\tilde{\xi}_t \in \mathcal{F}_{t-}^{\tilde{X}^{\pi}}$ corresponds to a measurable function $\tilde{\boldsymbol{\xi}} : [0,T] \times D([0,T]) \mapsto \mathbb{R}$ such that $\tilde{\xi}_t = \tilde{\boldsymbol{\xi}}(t, \tilde{X}^{\pi}_{t-\wedge\cdot})$. Then, $\tilde{\boldsymbol{\xi}}(t, X^{\pi}_{t-\wedge\cdot})$ is $\mathcal{F}_{t-}^{X^{\pi}}$-measurable for each $t \in [0,T]$, and equation (EC.15) holds for $\tilde{\boldsymbol{\xi}}(\cdot,\cdot)$. This establishes the equivalence between condition (EC.11) and condition (EC.10). The proof is therefore complete.

Proof of Theorem 3 By the representation (23) of $g(t,x;\phi) = \nabla_\phi J(t,x;\boldsymbol{\pi}^\phi)$, we have

$$\begin{aligned}\nabla_\phi J(0,c;\boldsymbol{\pi}^\phi) = \mathbb{E}\Bigg[\int_0^T \Bigg(\sum_{S \in \mathcal{A}(X^{\boldsymbol{\pi}^\phi}_{t-})} \nabla_\phi \log \boldsymbol{\pi}^\phi(S \mid t, X^{\boldsymbol{\pi}^\phi}_{t-}) H(t, X^{\boldsymbol{\pi}^\phi}_{t-}, S, J(\cdot,\cdot;\boldsymbol{\pi}^\phi)) \boldsymbol{\pi}^\phi(S \mid t, X^{\boldsymbol{\pi}^\phi}_{t-}) \Bigg) \mathrm{d}t \\ + \gamma \int_0^T \nabla_\phi \mathcal{H}(\boldsymbol{\pi}^\phi(\cdot \mid t, \tilde{X}^{\boldsymbol{\pi}^\phi}_{t-})) \, \mathrm{d}t \mid X^{\boldsymbol{\pi}^\phi}_0 = c \Bigg].\end{aligned} \tag{EC.16}$$

It suffices to show the first term of (EC.16) also equals to the first term of (24). To simplify the notation, we use $\mathbb{E}_c$ to denote the expectation conditional on the initial state $X^{\boldsymbol{\pi}^\phi}_0 = c$ in the subsequent derivations.

Denote the arrival time of the k-th customer as T_k and the product chosen by the customer as ζ_{T_k}, where $\zeta_{T_k} \in \{0,1,\ldots,n\}$. For $j = 1,\ldots,n$, denote

$$I_j(t,x,S) \coloneqq \nabla_\phi \log \boldsymbol{\pi}^\phi(S \mid t,x)[J(t,x-A^j;\boldsymbol{\pi}^\phi) - J(t,x;\boldsymbol{\pi}^\phi) + p_j].$$

Then, the first term of (24) can be rewritten as

$$\sum_{k=1}^{\infty} \mathbb{E}_c \Bigg[\sum_{j=1}^{n} I_j(T_k, X^{\boldsymbol{\pi}^\phi}_{T_k-}, S^{\boldsymbol{\pi}^\phi}_{T_k}) \mathbf{1}_{\{\zeta_{T_k}=j\}} \mathbf{1}_{\{T_k \le T\}} \Bigg]. \tag{EC.17}$$

For each $k \ge 1$, by taking conditional expectation with respect to $\sigma(T_k, X^{\boldsymbol{\pi}^\phi}_{T_k-})$ inside the expectation in (EC.17), we obtain

$$\begin{aligned}&\sum_{k=1}^{\infty} \mathbb{E}_c \Bigg[\sum_{S \in \mathcal{A}(X^{\boldsymbol{\pi}^\phi}_{T_k-})} \Bigg(\sum_{j=1}^{n} I_j(T_k, X^{\boldsymbol{\pi}^\phi}_{T_k-}, S) P_j(S) \boldsymbol{\pi}^\phi(S \mid T_k, X^{\boldsymbol{\pi}^\phi}_{T_k-}) \Bigg) \mathbf{1}_{\{T_k \le T\}} \Bigg] \\ &= \mathbb{E}_c \Bigg[\int_0^T \sum_{S \in \mathcal{A}(X^{\boldsymbol{\pi}^\phi}_{t-})} \Bigg(\sum_{j=1}^{n} I_j(t, X^{\boldsymbol{\pi}^\phi}_{t-}, S) P_j(S) \boldsymbol{\pi}^\phi(S \mid t, X^{\boldsymbol{\pi}^\phi}_{t-}) \Bigg) \mathrm{d}N^{\lambda}_t \Bigg] \\ &= \mathbb{E}_c \Bigg[\int_0^T \sum_{S \in \mathcal{A}(X^{\boldsymbol{\pi}^\phi}_{t-})} \Bigg(\sum_{j=1}^{n} I_j(t, X^{\boldsymbol{\pi}^\phi}_{t-}, S) \lambda P_j(S) \boldsymbol{\pi}^\phi(S \mid t, X^{\boldsymbol{\pi}^\phi}_{t-}) \Bigg) \mathrm{d}t \Bigg],\end{aligned} \tag{EC.18}$$

where (EC.18) is derived from the fact that $\{N^{\lambda}_t - \lambda t : t \in [0,T]\}$ is an $\{\mathcal{F}_t\}_{t\ge0}$-martingale, and the integrand is $\{\mathcal{F}_t\}_{t\ge0}$-predictable and bounded. The boundedness of the integrand is ensured by Assumption 1 and the inherent boundedness of the state process $\{X^{\boldsymbol{\pi}^\phi}_t : t \in [0,T]\}$. On the other hand, given the definition of H in (2), we have

$$H(t,x,S,J(\cdot,\cdot;\boldsymbol{\pi}^\phi)) = \lambda \sum_{j=1}^{n} p_j P_j(S) + \sum_{y \ne x} \Bigg(\sum_{\{j \in \mathcal{J} : A^j = x-y\}} \lambda P_j(S) \Bigg) J(t,y;\boldsymbol{\pi}^\phi) - \lambda[1 - P_0(S)] J(t,x;\boldsymbol{\pi}^\phi)$$

$$= \sum_{j=1}^{n} [J(t, x - A^j; \boldsymbol{\pi}^\phi) - J(t, x; \boldsymbol{\pi}^\phi) + p_j] \lambda P_j(S).$$

Thus, (EC.18) is exactly the first term of (EC.16). This concludes the proof.

EC.5. The instability of ADP time discretization

Figure EC.1 illustrates the variation of the coefficients $V_{0,i}$ for $i = 1, \ldots, 6$, selected from the ADP solution in Section 6.3, across five closely spaced levels of time discretization $\Delta t = 0.48, 0.49, 0.50, 0.51$ and 0.52.

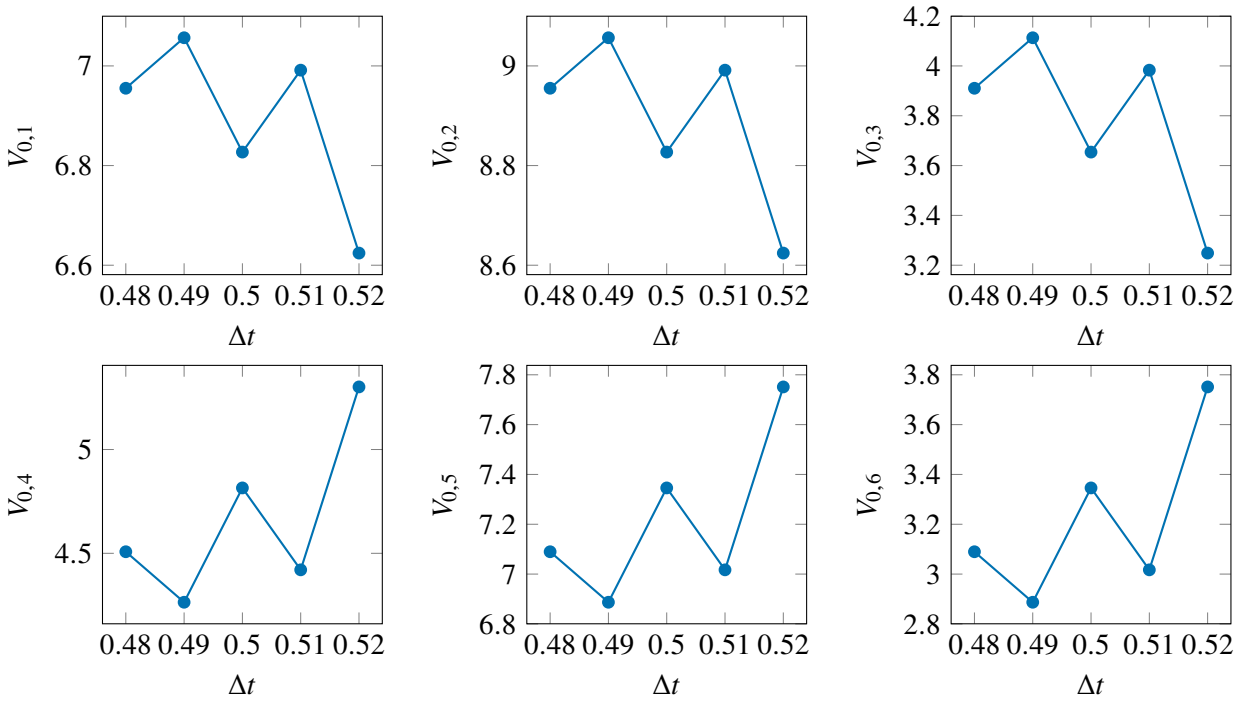

Figure EC.1 Variations of selected coefficients from the ADP solution across different time discretization levels

EC.6. Extension to Admission Control in a Single-Server Queue

To demonstrate the generality of our framework, we apply it to another classic problem in Operations Research: admission control in a single-server queue. We assume the customer arrival process and service process are non-homogeneous Poisson process with rate function $\lambda(t)$ and $\mu(t)$, respectively. If $\mu(t) \equiv \mu$, that means the service times of customers are i.i.d. exponentially distributed with rate μ. An agent is responsible for determining whether to admit arrivals in order to maximize the expected total revenue over a finite horizon $[0, T]$. The system has a maximum capacity of C customers, and it is intially empty. Let A_t denote the admitted arrival process, and D_t denote the customer departure process. The queue length (which includes the customer in service), represented by $X_t = A_t - D_t$, is the system state. The action space is a binary set $\{0, 1\}$, where 0 indicates rejecting an arriving customer, and 1 indicates admitting the customer. Then, a

(randomized) policy $\boldsymbol{\pi}$ is defined as a mapping from each time-state pair (t,x) to a Bernoulli distribution with success probability p. The entropy-regularized value function is given by

$$J(t,x;\boldsymbol{\pi}) \coloneqq \mathbb{E}\left[K_1\int_{(t,T]} \mathrm{d}A_s^{\boldsymbol{\pi}} - K_2\int_t^T X_{s-}^{\boldsymbol{\pi}}\,\mathrm{d}s - K_3 X_T^{\boldsymbol{\pi}} + \gamma\int_t^T \mathcal{H}(\boldsymbol{\pi}(\cdot\mid s, X_{s-}^{\boldsymbol{\pi}}))\,\mathrm{d}s \mid X_t^{\boldsymbol{\pi}} = x\right].$$

The positive coefficients K_1, K_2, K_3 are interpreted as follows: K_1 is the jump reward for admitting a customer, K_2 is the holding cost per unit of time for each admitted customer, and K_3 denotes the penalty for an admitted customer not served by the terminal time T. The earlier analysis can be readily extended to this queueing formulation, with almost identical proof methodologies. We present the results as follows.

For an approximate value function parametrized by θ, we define the loss function $L(\theta)$ as

$$L(\theta) = \frac{1}{2}\mathbb{E}\left[\int_0^T \left(K_1\int_{(t,T]} \mathrm{d}A_s^{\boldsymbol{\pi}} - K_2\int_t^T X_{s-}^{\boldsymbol{\pi}}\,\mathrm{d}s - K_3 X_T^{\boldsymbol{\pi}} + \gamma\int_t^T \mathcal{H}(\boldsymbol{\pi}(\cdot\mid s, X_{s-}^{\boldsymbol{\pi}}))\,\mathrm{d}s - J^\theta(t, X_t^{\boldsymbol{\pi}})\right)^2 \mathrm{d}t\right].$$

It allows us to establish the equivalence theorem, which confirms that the Monte Carlo method for PE remains effective within the queueing framework.

THEOREM EC.1 **(Parallel to Theorem 1)**. *It holds that* $\arg\min_\theta L(\theta) = \arg\min_\theta \mathrm{MSVE}(\theta)$.

Moreover, we have developed the martingale orthogonality conditions for the queueing model, which form the basis for the TD method in PE.

THEOREM EC.2 **(Parallel to Theorem 2)**. *A function $v \in C^{1,0}([0,T]\times\mathcal{X})$ is the value function associated with the policy $\boldsymbol{\pi}$, i.e. $v(t,x) = J(t,x;\boldsymbol{\pi})$ for all $(t,x)\in[0,T]\times\mathcal{X}$, if and only if it satisfies $v(T,x) = -K_3 x$ for all $x\in\mathcal{X}$, and the following martingale orthogonality condition holds for any bounded process ξ with $\xi_t \in \mathcal{F}_{t-}^{X^{\boldsymbol{\pi}}}$ for all $t\in[0,T]$:*

$$\mathbb{E}\left[\int_0^T \xi_t\left\{\mathrm{d}v(t, X_t^{\boldsymbol{\pi}}) + K_1 dA_t^{\boldsymbol{\pi}} - K_2 X_{t-}^{\boldsymbol{\pi}}\,\mathrm{d}t + \gamma\mathcal{H}(\boldsymbol{\pi}(\cdot\mid t, X_{t-}^{\boldsymbol{\pi}}))\,\mathrm{d}t\right\}\right] = 0.$$

Finally, we derive the formula for PG.

THEOREM EC.3 **(Parallel to Theorem 3)**. *Given an admissible parameterized policy $\boldsymbol{\pi}^\phi$ satisfying Assumption 1, the policy gradient $\nabla_\phi J(0,c;\boldsymbol{\pi}^\phi)$ admits the following representation:*

$$\nabla_\phi J(0,c;\boldsymbol{\pi}^\phi) = \mathbb{E}\left[\int_{(0,T]} \nabla_\phi \log\boldsymbol{\pi}^\phi(1\mid t, X_{t-}^{\boldsymbol{\pi}^\phi})[J(t, X_{t-}^{\boldsymbol{\pi}^\phi}+1;\boldsymbol{\pi}^\phi) - J(t, X_{t-}^{\boldsymbol{\pi}^\phi};\boldsymbol{\pi}^\phi) + K_1]dA_t^{\boldsymbol{\pi}^\phi} + \gamma\int_0^T \nabla_\phi\mathcal{H}(\boldsymbol{\pi}^\phi(\cdot\mid t, X_{t-}^{\boldsymbol{\pi}^\phi}))\,\mathrm{d}t\right].$$

On combining PE with PG, we can now design actor-critic algorithms. We present Algorithm EC.3 as an example. This algorithm employs the neural-network-based approximation for both value and policy functions, and utilizes the Monte Carlo method for PE.

To demonstrate the performance of Algorithm EC.3, we consider a simple example: we set the maximum capacity $C = 10$, the time horizon $T = 20$, the coefficients $K_1 = 10$, $K_2 = 1$, $K_3 = 0.1$. The rate functions for customer arrival process and service process are defined as $\lambda(t) = 0.5 + 0.3 \cdot \sin(\frac{2\pi t}{T})$ and $\mu(t) = 0.1 + 0.1 \cdot \frac{t}{T}$, respectively. However, the RL agent is not aware of these underlying rate functions.

We employ the neural-network-based approximation for value functions and policies, following the 2-NNs approach described in Section 6. We continue to use fully connected networks with each hidden layer followed by a ReLU activation function. The only minor variation in this example is that the outputs of the actor network are one-dimensional, and we directly apply the sigmoid function to convert these outputs into admission probabilities. Algorithm EC.3 is implemented with both the actor and critic networks configured with two hidden layers, each having a width of 8. The hyperparameters are set as follows: batch size $M = 100$, entropy factor $\gamma = 1 \times 10^{-3}$, learning rate $\alpha_\theta = 3 \times 10^{-2}$, and learning rate $\alpha_\phi = 1 \times 10^{-5}$. Figure EC.2 illustrates how the average revenues of the updated policies varies throughout the learning process. Upon completing the final update, the simulated average revenues of the resulting policy, achieves a value of 23.930.

We also compare the performance of our RL algorithm with three benchmark policies:

• Optimal Policy from Discretized Dynamic Programming: Similar as in Section 6.1.1, for small Δt, solving the DP problem (without entropy regularization) for the discrete-time model can be expected to yield a reliable approximation $V^*_{\Delta t}(0,0)$ of the true optimal expected revenue $V^*(0,0)$. The associated DP equation is given by

$$\begin{cases} V^*_{\Delta t}(t_k,x) = \lambda(t_{k+1})\Delta t \cdot \mathbf{1}_{\{x \le C-1\}} \cdot \max\limits_{a\in\{0,1\}} \left\{a \cdot [K_1 + V^*_{\Delta t}(t_{k+1},x+1) - V^*_{\Delta t}(t_{k+1},x)]\right\} + K_2 \Delta t \cdot x \\ \qquad\qquad + \mu(t_{k+1})\Delta t \cdot \mathbf{1}_{\{x \ge 1\}} \cdot [V^*_{\Delta t}(t_{k+1},x-1) - V^*_{\Delta t}(t_{k+1},x)] + V^*_{\Delta t}(t_{k+1},x), \quad \forall\, k \\ V^*_{\Delta t}(t_K,x) = K_3 \cdot x \end{cases} \tag{EC.19}$$

• Uniform-Random: At each time and state, the decision to admit or reject an arriving customer is made with an equal probability of 0.5.

• Optimal-Threshold: For each state $\bar{x} \in \{1,\dots,C\}$, a (deterministic) threshold policy $\boldsymbol{\pi}^{\bar{x}}$ admits an arriving customer only if the current state is below the threshold $\bar{x}$. We execute all these threshold policies and refer to the maximum average revenues obtained as the performance of the Optimal-Threshold policy.

We set $\Delta t = 0.001$ and solve the DP problem (EC.19), yielding an optimal expected revenue of $V^*_{0.001}(0,0) = 23.997$, as indicated in Figure EC.2. Table EC.2 reports the simulated average revenues for our 2-NNs policy and the benchmark policies Uniform-Random and Optimal-Threshold. The numerical

Algorithm EC.3 Actor-Critic Algorithm with Neural Networks for Queueing Control

Inputs: maximum capacity C, time horizon T, the values of K_1, K_2, K_3; number of episodes N, batch size M; critic network $J^\theta(t,x)$ and an initial value θ_0; actor network $\boldsymbol{\pi}^\phi(a \mid t,x)$ and an initial value ϕ_0; entropy factor γ, learning rates α_θ, α_ϕ

Required program: an environment simulator $(t',x',r') = Environment(t,x,\boldsymbol{\pi}^\phi(\cdot \mid \cdot,\cdot);C)$

Initialize $\theta \leftarrow \theta_0$, $\phi \leftarrow \phi_0$

for $i = 1$ to N **do**

  Store $(\tau_0^{(i)}, x_0^{(i)}) \leftarrow (0,0)$

  Initialize $l = 0$, $(t,x) = (0,0)$

  while True **do**

    Apply (t,x) and policy $\boldsymbol{\pi}^\phi$ to the environment simulator to get the next jump time t', post-jump state x' and the jump reward r' ▷ $r' = K_1$ for $x' = x+1$ and $r' = 0$ for $x' = x-1$

    if $t' \geq T$ **then**

      Store $L^{(i)} \leftarrow l$, $\tau_{L^{(i)}+1}^{(i)} \leftarrow T$

      break

    end if

    Update $l \leftarrow l+1$. Store observation at jump time: $(\tau_l^{(i)}, x_l^{(i)}, r_l^{(i)}) \leftarrow (t',x',r')$

    Update $(t,x) \leftarrow (t',x')$

  end while

  Denote $D(t_1,t_2,x;\theta) \coloneqq \int_{t_1}^{t_2} [J^\theta(s,x)]^2 \,\mathrm{d}s$, $b(t_1,t_2,x;\theta) \coloneqq \int_{t_1}^{t_2} J^\theta(s,x)\,\mathrm{d}s$

  $E(t_1,t_2,x;v,\boldsymbol{\pi}) \coloneqq \int_{t_1}^{t_2} v(s)\mathcal{H}(\boldsymbol{\pi}(\cdot \mid s,x))\,\mathrm{d}s$, $w(t_1,t_2,x;\theta) \coloneqq \int_{t_1}^{t_2} s \cdot J^\theta(s,x)\,\mathrm{d}s$

  Denote $D_l(\theta,i) \coloneqq D(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)};\theta)$, $b_l(\theta,i) \coloneqq b(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)};\theta)$, $w_l(\theta,i) \coloneqq w(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)};\theta)$

  $\overline{C}_l(\boldsymbol{\pi}^\phi,i) \coloneqq \sum_{l'=l+1}^{L^{(i)}} [r_{l'}^{(i)} - K_2 x_{l'}^{(i)}(\tau_{l'+1}^{(i)} - \tau_{l'}^{(i)}) + \gamma E(\tau_{l'}^{(i)}, \tau_{l'+1}^{(i)}, x_{l'}^{(i)}; \mathbf{1}, \boldsymbol{\pi}^\phi)]$

  $E_l(\theta,\boldsymbol{\pi}^\phi,i) \coloneqq E(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)}; b(\tau_l^{(i)},\cdot,x_l^{(i)};\theta), \boldsymbol{\pi}^\phi)$, $E_l^{\mathbf{1}}(\boldsymbol{\pi}^\phi,i) = E(\tau_l^{(i)}, \tau_{l+1}^{(i)}, x_l^{(i)}; \mathbf{1}, \boldsymbol{\pi}^\phi)$

  $\overline{U}_l^\theta(\boldsymbol{\pi}^\phi,i) \coloneqq \log \boldsymbol{\pi}^\phi(1 \mid \tau_l^{(i)}, x_{l-1}^{(i)})[J^\theta(\tau_l^{(i)}, x_{l-1}^{(i)}+1) - J^\theta(\tau_l^{(i)}, x_{l-1}^{(i)}) + K_1]$

  Store $\mathcal{I}^{(i)} = \{1 \leq l \leq L^{(i)} : x_l^{(i)} = x_{l-1}^{(i)} + 1\}$ ▷ Store indices of customer admission transitions

  if $i = 0 \pmod M$ **then** ▷ Perform an update using M episodes generated under the policy

    Update critic network:

$$\theta \leftarrow \theta - \alpha_\theta \nabla_\theta \left\{ \frac{1}{M} \sum_{k=i-M+1}^{i} \sum_{l=0}^{L^{(k)}} \left(\frac{1}{2} D_l(\theta,k) + b_l(\theta,k)[K_2 \cdot x_l^{(k)} \tau_{l+1}^{(k)} + K_3 \cdot x_{L^{(k)}}^{(k)}] \right.\right.$$
$$\left.\left. - b_l(\theta,k)\overline{C}_l(\boldsymbol{\pi}^\phi,k) - K_2 x_l^{(k)} w_l(\theta,k) - \gamma E_l(\theta,\boldsymbol{\pi}^\phi,k) \right) \right\}$$

    Update actor network:

$$\phi \leftarrow \phi + \alpha_\phi \nabla_\phi \left\{ \frac{1}{M} \sum_{k=i-M+1}^{i} \left(\sum_{l \in \mathcal{I}^{(k)}} \overline{U}_l^\theta(\boldsymbol{\pi}^\phi,k) + \gamma \sum_{l=0}^{L^{(k)}} E_l^{\mathbf{1}}(\boldsymbol{\pi}^\phi,k) \right) \right\}$$

  end if

end for

results demonstrate that our RL algorithm significantly outperforms the UNIFORM-RANDOM and OPTIMAL-THRESHOLD benchmarks, and its performance approaches that of the optimal policy from DP with a negligible gap of only 0.3%.

Figure EC.2 Average revenues of Algorithm EC.3 over episodes for the example in Section EC.6

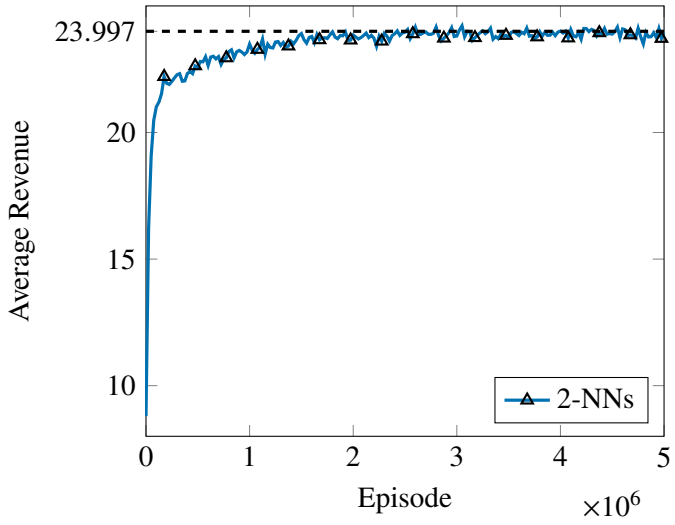

Table EC.2 Simulation results for selected policies in Section EC.6

Policy	Average revenue	99%-CI (±)	$\frac{\text{Average revenue}}{V^*_{0.001}(0,0)}$ (%)	Time cost (second)
2-NNs	23.930	0.390	99.72	1,526
UNIFORM-RANDOM	8.603	0.415	35.85	*
OPTIMAL-THRESHOLD	13.358	0.325	55.67	*

* indicates that the method involves neither solving LP problems nor learning; the time cost is virtually zero.