A Near-Optimal Control Policy for Data-driven Assemble-to-order Systems

Lun Yu
CUHK-Shenzhen
yulun@cuhk.edu.cn

Zhixuan Cai
Tsinghua University
caizx21@mails.tsinghua.edu.cn

Zhaoran Wang
Northwestern University
zhaoranwang@gmail.com

Tianhu Deng
Soochow University
thdeng@suda.edu.cn

Abstract

We study a data-driven assemble-to-order (ATO) control problem, aiming to synchronize component ordering and product assembly under unknown demand distributions and non-identical lead times. We address two key questions: the statistical tractability of learning a near-optimal policy from limited data and the computational complexity of obtaining it. To the best of our knowledge, our work is the first to analyze the sample efficiency for a general ATO system as a multidimensional control problem and to propose an algorithm that finds a provably near-optimal solution. Methodologically, we introduce a novel asymmetric Lipschitz continuity (ALC) property to establish regularity conditions for the infinite-horizon problem. Surprisingly, we prove that the data-driven ATO problem avoids the curse of dimensionality; the performance gap of our policy scales as $O(M^{-1/2}\log M)$ with sample size M, only logarithmically worse than approximating the demand mean. We develop a specialized reinforcement learning (RL) algorithm that exploits a convex-preserving property in ATO dynamics, using input convex neural networks and interior point methods to achieve computational feasibility. Numerical studies show our algorithm consistently and significantly outperforms existing heuristics and a general-purpose RL benchmark.

1 Introduction

Assemble-to-order (ATO) is a widespread production strategy where products are assembled from components only after customer orders are received. This allows companies like Dell and BMW [9, 12] to offer high product variety while minimizing finished-goods inventory. However, managing ATO systems is challenging due to the need to coordinate procurement of numerous components with varying lead times to meet uncertain future demand for multiple products, and is particularly challenging when the demand distribution is unknown and must be learned from historical data.

Motivated by our collaboration with a processed-food manufacturer, which delivers a variety of 8 products using 15 components, we consider the ATO control problem in a data-driven setting. We formulate it as an infinite-horizon Markov Decision Process (MDP) and address two fundamental questions:

1. **Statistical Tractability:** How does the performance of the best policy learned from a finite demand sample of size M scale with the problem dimension? Can we learn a near-optimal policy without an impractically large dataset?

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: MLxOR: Mathematical Foundations and Operational Integration of Machine Learning for Uncertainty-Aware Decision-Making.

2. **Computational Complexity:** Can we efficiently compute such a policy, given that the optimal policy structure is unknown and the state-action space is high-dimensional?

Our Contributions. Our work provides affirmative answers to these questions.

First, we show that a Convex Fitted Q-Iteration (CFQI) algorithm efficiently computes a near-optimal policy. The optimality gap of CFQI is $O(p^2 M^{-1/2} \log M)$ relative to the intractable optimal policy, where p is the dimension of the system. This demonstrates that accurately approximating the high-dimensional optimal value function requires only logarithmically more samples than approximating the demand mean, which avoids the curse of dimensionality. Our theoretical analysis introduces a novel *asymmetric Lipschitz continuity (ALC)* property. Standard Lipschitz conditions are often too restrictive for infinite-horizon inventory problems. ALC provides a weaker, more suitable regularity condition that reflects the inherent asymmetry between over-stocking and under-stocking costs, enabling a rigorous sample complexity analysis for the ATO control problem.

Second, we propose a specialized reinforcement learning (RL) algorithm as implementation of CFQI. We leverage a convex-preserving property of the control problem by employing Input Convex Neural Networks (ICNNs) [2] to approximate the convex value functions, transforming the complex policy improvement step into a series of unconstrained convex optimizations. Our extensive numerical results validate our theory, showing that our algorithm significantly outperforms strong baselines, including a general-purpose TD3 algorithm and established heuristics.

2 Data-driven ATO as a Markov Decision Process

We model an ATO system with m products and n components. The assembly structure is defined by a Bill-of-Materials (BOM) matrix $A \in \mathbb{R}^{m \times n}$. Component j has a lead time of $L_j \geq 1$ periods. In each period t, the system state $\boldsymbol{s}(t) = (\boldsymbol{x}(t), \tilde{\boldsymbol{u}}(t), R(t))$ is observed, where $\boldsymbol{x}(t) \in \mathbb{R}_+^n$ is the on-hand component inventory, $\tilde{\boldsymbol{u}}(t) \in \mathbb{R}_+^m$ is the total backlog (including new demand $\boldsymbol{d}(t)$), and $R(t) \in \mathbb{R}^{n \times L_n}$ is the pipeline inventory matrix. The demand vector $\boldsymbol{d}(t)$ is i.i.d. from an unknown distribution F with compact support, but we have access to a historical sample $\widehat{D} = \{\boldsymbol{d}_1, \ldots, \boldsymbol{d}_M\}$.

The actions are $\boldsymbol{a}(t) = (\boldsymbol{y}(t), \boldsymbol{z}(t+1))$, where $\boldsymbol{y}(t) \in \mathbb{R}_+^m$ is the assembly quantity and $\boldsymbol{z}(t+1) \in \mathbb{R}_+^n$ is the component order quantity for the next cycle. Actions are constrained by available inventory: $A^\top \boldsymbol{y}(t) \leq \boldsymbol{x}(t)$ and $\boldsymbol{y}(t) \leq \tilde{\boldsymbol{u}}(t)$. The single-period cost is $C(\boldsymbol{s}(t), \boldsymbol{a}(t)) = \boldsymbol{h}^\top(\boldsymbol{x}(t) - A^\top \boldsymbol{y}(t)) + \boldsymbol{b}^\top(\tilde{\boldsymbol{u}}(t) - \boldsymbol{y}(t))$, where $\boldsymbol{h}$ and $\boldsymbol{b}$ are component holding and product backorder costs, respectively. The system evolves according to the transition function $\boldsymbol{s}(t+1) = f(\boldsymbol{s}(t), \boldsymbol{a}(t), \boldsymbol{d}(t+1))$. The objective is to find a policy π that minimizes the total expected discounted cost $V^\pi(\boldsymbol{s}; F) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t C(\boldsymbol{s}(t), \boldsymbol{a}(t))\right]$, where the expectation is over future demands from the true distribution F. The optimal value function is $V^*(\boldsymbol{s}; F) = \inf_\pi V^\pi(\boldsymbol{s}; F)$. For more details on the ATO model, please refer to Appendix A.

3 Algorithm and Theoretical Guarantees

Directly solving the MDP is intractable due to the unknown distribution F and the continuous, high-dimensional state-action spaces. We propose a solution based on *Sample Average Approximation (SAA)* combined with a specialized RL algorithm. To handle the unbounded state space, we first introduce a *Damped Sample Average Approximation (DSAA)* of the problem. We introduce a damping factor $\alpha \in (0, 1)$ to the state transition, making the state and action spaces provably bounded while preserving the linearity of the dynamics and convexity of the state-action space. We then replace the unknown demand distribution F with the empirical distribution $\widehat{F}$ from the sample $\widehat{D}$. This defines a tractable MDP approximation with a known transition kernel.

The optimal action-value function Q^* for this DSAA problem can be found as the unique fixed point of the Bellman optimality operator, $\mathcal{T}^* Q = Q$, where for a state-action pair $(\boldsymbol{s}, \boldsymbol{a})$:

$$(\mathcal{T}^* w)(\boldsymbol{s}, \boldsymbol{a}) := C(\boldsymbol{s}, \boldsymbol{a}) + \gamma \mathbb{E}_{\hat{\boldsymbol{d}} \sim \widehat{F}} \left[\min_{\boldsymbol{a}' \in \mathcal{A}_\alpha(\boldsymbol{s}')} w(\boldsymbol{s}', \boldsymbol{a}') \right], \quad \text{for } \boldsymbol{s}' = f_\alpha(\boldsymbol{s}, \boldsymbol{a}, \hat{\boldsymbol{d}}). \tag{1}$$

Our first key insight is that this operator preserves convexity.

Theorem 1 (**Convexity Preservation**). *If the function w is convex over the state-action space, then $\mathcal{T}^* w$ is also convex.*

This theorem implies that the optimal action-value function Q^* of the DSAA problem is convex. This property is critical, as it allows us to restrict the search for Q^* to the space of convex functions, which can be efficiently approximated and optimized.

Based on the convexity property, we propose the Convex Fitted Q-Iteration (CFQI) algorithm (Algorithm 1). CFQI approximates the action-value function Q^* by iteratively applying the Bellman operator to a sequence of convex functions $\{Q_k\}$. In practice, we parameterize Q_k using *Input Convex Neural Networks (ICNNs)*, a neural network architecture whose output is guaranteed to be a convex function of its input [2].

The policy improvement step, $\min_{\boldsymbol{a}} Q_k(\boldsymbol{s}, \boldsymbol{a})$, is a constrained convex optimization problem. To handle this efficiently across thousands of states in each iteration, we employ a penalty method, transforming it into an unconstrained convex optimization that can be solved rapidly using gradient-based methods. The final output is a policy $\hat{\pi}^*(\widehat{D})$ derived from the learned convex Q-function, translated back to be applicable to the original, non-damped system.

Algorithm 1: Convex FQI.

Input: Number of iterations K, Class $\mathcal{G}$ with accuracy level ε, regressor `Regress`, initial actor-value function $Q_0 \in \mathcal{C} \cap \mathcal{G}$, sample $\widehat{D}$, initial state $\boldsymbol{s} = \boldsymbol{s}(0)$, and constants $\alpha \in (0, \bar{d}/\|\boldsymbol{s}\|_\infty)$ and $\eta \in (0, 1)$.

Output: Policy $\hat{\pi}^*(\widehat{D})$.

for $k = 0$ **to** $K - 1$ **do**

apply `Regress` to obtain convex function

$$Q_{k+1} \in \texttt{Regress}\Big(\mathcal{T}^*\big(Q_k(\,\cdot\,) + \Lambda(\,\cdot\,; \eta^k)\big)\Big), \tag{2}$$

where the Bellman operator $\mathcal{T}^*$ is characterized by taking $G = \widehat{F}$.

take a map ξ_K on $\mathcal{S}_\alpha$ satisfying

$$\xi_K(\boldsymbol{s}) \in \arg\min_{\boldsymbol{a} \in \mathcal{A}(\boldsymbol{s})} Q_K(\boldsymbol{s}, \boldsymbol{a}).$$

return policy $\hat{\pi}^*(\widehat{D}) := \{\hat{\boldsymbol{a}}(t) : t \geq 0\}$, for recursively defined sequences

$$\hat{\boldsymbol{a}}(t) := (\hat{\boldsymbol{y}}(t), \hat{\boldsymbol{z}}(t+1)) := \xi_K(\widehat{\boldsymbol{s}}(t) - \boldsymbol{r}(t)), \tag{3}$$

$$\widehat{\boldsymbol{s}}(t+1) = f(\widehat{\boldsymbol{s}}(t), \hat{\boldsymbol{a}}(t), \boldsymbol{d}(t+1)), \quad \text{and}$$

$$\boldsymbol{r}(t+1) = (1-\alpha)\boldsymbol{r}(t) + \alpha(\boldsymbol{x}(t) - A^\top \boldsymbol{y}(t), \tilde{\boldsymbol{u}}(t) - \boldsymbol{y}(t), O). \forall t = 0, 1, \cdots, \tag{4}$$

with $\widehat{\boldsymbol{s}}(0) = \boldsymbol{s}$ and $\boldsymbol{r}(0) = \boldsymbol{0}$. Note that $\{\widehat{\boldsymbol{s}}(t)\}$ is the controlled system state.

A major challenge in analyzing the sample complexity is establishing regularity of the value function. Standard Lipschitz continuity conditions [8] do not hold for the ATO problem. To overcome this, we introduce a novel property.

Theorem 2 (**Asymmetric Lipschitz Continuity (ALC)**). *The Bellman operator $\mathcal{T}^*$ preserves a specific form of asymmetric Lipschitz continuity. This implies that the optimal value function $V_\alpha^*(\cdot; G)$ is uniformly Lipschitz continuous for any demand distribution G.*

The ALC property intuitively captures the different cost implications of positive vs. negative changes in inventory levels (i.e., overage versus underage). Detailed definition for the ALC property can be found in Appendix B. Armed with this tool, we establish our main theoretical result on the performance of the policy generated by CFQI.

Theorem 3 (**Algorithmic Error**). *For any $\delta > 0$, $\boldsymbol{s} \in \mathcal{S}$, and distribution $F \in \mathcal{D}$, the output data driven policy $\hat{\pi}^*$ from Algorithm 1 with*

$$\alpha = \min\Big\{\frac{1-\gamma}{M^{1/2}}, \frac{\bar{d}}{\|\boldsymbol{s}\|_\infty}\Big\}, \quad \eta = \max\Big\{\frac{\gamma}{2}, 2\gamma - 1\Big\}, \quad \textit{and}$$

$$K \geq 2 + \frac{2\log\alpha + \log\varepsilon_m - \log(3\bar{a}\bar{c}\bar{d}L_n(m+n))}{\log(1-\gamma)}$$

gives a data-driven policy $\hat{\pi}^*$ satisfying

$$V^{\hat{\pi}^*(\widehat{D})}(\boldsymbol{s}; F) - V^*(\boldsymbol{s}; F) \leq \frac{1}{\sqrt{M}} V^*(\boldsymbol{s}; F) + \frac{8\varepsilon_m}{(1-\gamma)^3} + \mathcal{E}(\delta) \tag{5}$$

with probability at least $1 - \delta$. Here,

$$\mathcal{E}(\delta) = O\Big(\frac{\bar{a}\bar{c}(p_S + p_A)^2}{(1-\gamma)^2} \sqrt{\frac{1}{M} \log \frac{\bar{a}M}{\alpha\delta}}\Big)$$

and $O(\,\cdot\,)$ hides a constant that does not depend on the model primitives.

Theorem 3 shows that the optimality gap shrinks at a rate of nearly $M^{-1/2}$ and grows polynomially in dimension, thus avoiding the curse of dimensionality. This guarantees that a near-optimal policy can be learned from a reasonably sized dataset.

4 Numerical Studies

We validate our proposed algorithm, CTD3 (a practical implementation of CFQI using TD3 architecture with ICNN critics), against two strong baselines: 1) **PTD3**, a penalized TD3 algorithm using standard (non-convex) networks, and 2) **NV-PRP**, a state-of-the-art heuristic based on newsvendor decomposition and priority rules [6]. We test on three systems of increasing complexity: a small N-system (2 components, 2 products) where the optimal policy is known, a mid-sized PC-system (6 components, 4 products), and a large-scale real-world system from a food manufacturer (15 components, 8 products).

Results. Across all settings, CTD3 demonstrates superior performance.

- **N-System:** With independent demands, all methods perform well. However, when demands are negatively correlated (substitutes), CTD3 provides a cost reduction of over 14% compared to the NV-PRP heuristic, which struggles with demand correlation. CTD3 also converges faster and more stably than PTD3.
- **PC-System:** CTD3 consistently achieves lower costs than both PTD3 and NV-PRP across all sample sizes. The performance improves rapidly as M increases from 10 to 200, confirming our theory that a large sample is not required.
- **Large-System:** In the real-world industrial case (Table 6), the advantage of exploiting convexity becomes crucial. CTD3 achieves a cost of 32.0, an 11.6% improvement over NV-PRP. In contrast, the general-purpose PTD3 algorithm fails to find a competitive policy, highlighting the importance of incorporating problem structure into the RL algorithm design for high-dimensional control tasks.

More detailed results can be found in Appendix E.

Insights on Heuristics. Our learned policy reveals insights into common heuristics. In the PC-system, we found that the CTD3 policy frequently violates the well-known "no-holdback" rule, which dictates that available components should always be used to satisfy current demand. By strategically holding back components for more valuable future products, our policy achieves lower costs. This suggests that the optimal no-holdback rule for simpler systems can be significantly suboptimal in more complex, multi-product settings, potentially incurring double-digit performance losses.

5 Conclusion

We developed a specialized RL algorithm to solve the data-driven ATO control problem, providing both strong theoretical guarantees and superior empirical performance. Our analysis, enabled by the novel ALC property, is the first to show that this high-dimensional control problem is statistically tractable, avoiding the curse of dimensionality. Our algorithm, by exploiting the problem's inherent convexity with ICNNs, provides a new and effective benchmark for complex ATO systems. This work demonstrates the power of integrating structural properties from operations research into modern deep reinforcement learning frameworks to solve challenging real-world control problems.

References

[1] Praveen Agarwal, Mohamed Jleli, and Bessem Samet. *Fixed point theory in metric spaces: Recent advances and applications*. Springer Singapore, Singapore, 2018.

[2] Brandon Amos, Lei Xu, and J. Zico Kolter. Input convex neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 146–155. PMLR, 06–11 Aug 2017.

[3] Dimitri Bertsekas and Steven E Shreve. *Stochastic optimal control: The discrete-time case*, volume 5. Athena Scientific, 1996.

[4] Dimitri P Bertsekas et al. Dynamic programming and optimal control 3rd edition, volume ii. *Belmont, MA: Athena Scientific*, 1, 2011.

[5] David Blackwell. Discounted dynamic programming. *The Annals of Mathematical Statistics*, 36(1):226–235, 1965.

[6] Shuyu Chen, Lijian Lu, Jing-Sheng Jeannette Song, and Hanqin Zhang. Optimizing assemble-to-order systems: Decomposition heuristics and scalable algorithms. *SSRN Electronic Journal*, 2021.

[7] Levi DeValve, Saša Pekeč, and Yehua Wei. A primal-dual approach to analyzing ATO systems. *Management Science*, 66(11):5389–5407, November 2020.

[8] Karl Hinderer. Lipschitz continuity of value functions in markovian decision processes. *Mathematical Methods of Operations Research*, 62(1):3–22, 2005.

[9] Roman Kapuscinski, Rachel Q. Zhang, Paul Carbonneau, Robert Moore, and Bill Reeves. Inventory decisions in Dell's supply chain. *INFORMS Journal on Applied Analytics*, 34(3):191–205, June 2004.

[10] Lijian Lu, Jing-Sheng Song, and Hanqin Zhang. Optimal and asymptotically optimal policies for assemble-to-order N- and W-systems. *Naval Research Logistics*, 62(8):617–645, 2015.

[11] Yingdong Lu and Jing-Sheng Song. Order-based cost optimization in assemble-to-order systems. *Operations Research*, 53(1):151–169, February 2005.

[12] Joann Muller. BMW's push for made-to-order cars. *Forbes,*, 2010. URL `https://www.forbes.com/forbes/2010/0927/companies-bmw-general-motors-cars-bespoke-auto.html`.

[13] Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *Journal of Machine Learning Research*, 9:815–857, jun 2008.

[14] Martin J. Wainwright. *High-dimensional statistics: a non-asymptotic viewpoint*. Cambridge University Press, first edition, February 2019.

[15] Yaqi Xie, Will Ma, and Linwei Xin. Vc theory for inventory policies. *arXiv preprint arXiv:2404.11509*, 2024.

Appendix

A MDP Formulation of the ATO Problem

The ATO control problem can be analyzed using an infinite-horizon Markov Decision Process (MDP) formulation. In the formulation, states are updated whenever new information is realized, at which actions are also taken. Therefore, the state of the ATO system takes the form of $\boldsymbol{s}(t) := (\boldsymbol{x}(t), \tilde{\boldsymbol{u}}(t), R(t))$, which updates immediately after *observing new demands*. Here

- $\boldsymbol{x}(t)$ is the inventory level after receiving component replenishment,
- $\tilde{\boldsymbol{u}}(t) := \boldsymbol{u}(t) + \boldsymbol{d}(t)$ is the augmented unsatisfied demands,
- $R(t) \in \mathbb{R}^{n \times L_n}$ is the pipeline inventory matrix after procurement $\boldsymbol{z}(t)$ is executed.

For $j \in [n]$ and $\ell \in [L_j]$, the element on the (j, ℓ)-th entry of $R(t)$ is the quantity of component j ordered in period $t + \ell - L_j$, which is scheduled to arrive in period $t + \ell$. The entry is 0 if $j > L_j$.

We emphasize that the update of state $\boldsymbol{s}(t)$ occurs at a different time from the advancement of $\boldsymbol{x}(t)$ to $\boldsymbol{x}(t+1)$. Consequently, we specify the t-th action by $\boldsymbol{a}(t) := (\boldsymbol{y}(t), \boldsymbol{z}(t+1))$ since the product assembly quantity $\boldsymbol{y}(t)$ and component orders $\boldsymbol{z}(t+1)$ are both decided between the observations of demands $\boldsymbol{d}(t)$ and $\boldsymbol{d}(t+1)$. Let $p_S := m + n + nL_n$ and $p_A := m + n$ be the dimensional of the state and action vector, respectively, the state transition is characterized by a function $f : \mathbb{R}^{p_S + p_A + m} \to \mathbb{R}^{p_S}$,

$$\boldsymbol{s}(t+1) = f\big(\boldsymbol{s}(t), \boldsymbol{a}(t), \boldsymbol{d}(t+1)\big), \forall t \in \mathbb{Z}_+. \tag{6}$$

Given $\boldsymbol{s} = (\boldsymbol{x}, \tilde{\boldsymbol{u}}, R)$ and $\boldsymbol{a} = (\boldsymbol{y}, \boldsymbol{z})$, the function f is characterized by $f(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d}) = (\boldsymbol{x}', \tilde{\boldsymbol{u}}', R')$, where $\boldsymbol{x}' = \boldsymbol{x} - A^\top \boldsymbol{y} + R\boldsymbol{e}_1$, $\tilde{\boldsymbol{u}}' = \tilde{\boldsymbol{u}} + \boldsymbol{d} - \boldsymbol{y}$, and $R' \in \mathbb{R}^{n \times L_n}$ satisfies that

$$R'_{j,\ell} = \begin{cases} R_{j,\ell+1}, & \text{if } \ell < L_j, \\ z_j, & \text{if } \ell = L_j, \\ 0, & \text{if } \ell > L_j. \end{cases}$$

The manufacturer starts to control the system after the first demand is realized. The manufacturer observes the state $\boldsymbol{s}(0)$ and has control over the actions $\boldsymbol{a}(t)$ for $t \in \mathbb{Z}_+$. We represent the control policy of the ATO system by a stochastic process $\pi := \{\boldsymbol{a}(t) : t \in \mathbb{Z}_+\}$, which encapsulates all procurement and assembly decisions to be made. Let $\mathcal{R}$ to be the family of $n \times L_n$ nonnegative matrices R such that $R_{j,\ell} = 0$ for $\ell > L_j$. Let $\mathcal{S} := \mathbb{R}_+^{n+m} \times \mathcal{R} \subseteq \mathbb{R}_+^{p_S}$ be the state space. For state $\boldsymbol{s} = (\boldsymbol{x}, \tilde{\boldsymbol{u}}, R) \in \mathcal{S}$, denote by $\mathcal{A}(\boldsymbol{s})$ the state-dependent action space, i.e., $\mathcal{A}(\boldsymbol{s}) := \big\{(\boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}_+^{m+n} : \ A^\top \boldsymbol{y} \le \boldsymbol{x}, \ \boldsymbol{y} \le \tilde{\boldsymbol{u}}\big\}$. Note that $\mathcal{S}$ is a noncompact subset of $\mathbb{R}^{p_S}$. For each $\boldsymbol{s} \in \mathcal{S}$, $\mathcal{A}(\boldsymbol{s})$ is a noncompact subset $\mathbb{R}^{p_A}$. Policy π is admissible if: (i) the decision $\boldsymbol{a}(t) \in \mathcal{A}(\boldsymbol{s}(t))$ for all $t \in \mathbb{Z}_+$, where the controlled state process $\{\boldsymbol{s}(t) : t \in \mathbb{Z}_+\}$ is recursively defined by (6); and (ii) $\boldsymbol{a}(t)$ is independent of $\{\boldsymbol{d}(t') : t' > t\}$, for all $t \in \mathbb{Z}_+$. Let Π be the family of all admissible policies.

For $\boldsymbol{s} = (\boldsymbol{x}, \tilde{\boldsymbol{u}}, R) \in \mathcal{S}$ and $\boldsymbol{a} = (\boldsymbol{y}, \boldsymbol{z}) \in \mathcal{A}(\boldsymbol{s})$, define

$$C(\boldsymbol{s}, \boldsymbol{a}) := \boldsymbol{h}^\top (\boldsymbol{x} - A^\top \boldsymbol{y}) + \boldsymbol{b}^\top (\tilde{\boldsymbol{u}} - \boldsymbol{y}). \tag{7}$$

For ease of notation, define the domain of function C as $\mathcal{SA} := \{(\boldsymbol{s}, \boldsymbol{a}) : \boldsymbol{s} \in \mathcal{S}, \boldsymbol{a} \in \mathcal{A}(\boldsymbol{s})\}$. The value function, which characterizes the infinite horizon discounted cost associated with policy $\pi = \{\boldsymbol{a}(t) : t \in \mathbb{Z}_+\} \in \Pi$ and the initial state $\boldsymbol{s}$, is specified by

$$V^\pi(\boldsymbol{s}; F) := \mathbb{E}\Big[\sum_{t \in \mathbb{Z}_+} \gamma^t C(\boldsymbol{s}(t), \ \boldsymbol{a}(t)) \Big| \ \boldsymbol{s}(0) = \boldsymbol{s} \Big], \forall \boldsymbol{s} \in \mathcal{S}. \tag{8}$$

Here $\{\boldsymbol{s}(t)\}$ is the corresponding controlled state process and the expectation is taken with respect to $\{\boldsymbol{d}(t) : t \in \mathbb{Z}_+\}$. In the expression (8), we specify the dependency on the demand distribution F, as it is the only system primitive that is unknown to the manufacturer. Therefore, an ATO control policy π is optimal if $V^\pi(\boldsymbol{s}; F)$ is minimized over $\pi \in \Pi$. Define the optimal value function as

$$V^*(\boldsymbol{s}; F) := \inf_{\pi \in \Pi} V^\pi(\boldsymbol{s}; F), \forall \boldsymbol{s} \in \mathcal{S}. \tag{9}$$

Our goal in this paper is to find a *data-driven policy* that approximately solves the data-driven ATO problem. More formally, we aim to find a policy $\hat{\pi}(\widehat{D})$ that can be characterized by i.i.d. sample $\widehat{D}$ from distribution F, such that the optimality gap

$$V^{\hat{\pi}(\widehat{D})}(\boldsymbol{s}; F) - V^*(\boldsymbol{s}; F)$$

is small with high probability. This problem is challenging as the optimal value function $V^*(\ \cdot\ ; F)$ is multidimensional, has noncompact support, and is also unknown to the manufacturer.

B Asymmetric Lipschitz Continuity

Definition 1 (Asymmetric Seminorm)**.** *Let $\rho : \mathbb{R}^p \to \mathbb{R}_+$ be an asymmetric seminorm if*

(i) For $\boldsymbol{x} \in \mathbb{R}^p$ and scalar $a \geq 0$, $\rho(a\boldsymbol{x}) = a\rho(\boldsymbol{x})$.

(ii) For $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$, $\rho(\boldsymbol{x} + \boldsymbol{y}) \leq \rho(\boldsymbol{x}) + \rho(\boldsymbol{y})$.

Asymmetric seminorm is a relaxation of a norm such that $\rho(-\boldsymbol{x})$ may not equal $\rho(-\boldsymbol{x})$ and $\rho(\boldsymbol{x}) = 0$ may hold for some $\boldsymbol{x} \neq \boldsymbol{0}$.

Definition 2 (Asymmetric Lipschitz Continuity (ALC))**.** *A function $g : E \subseteq \mathbb{R}^p \to \mathbb{R}$ is L-ALC with respect to an asymmetric seminorm ρ if for $x_1, x_2 \in E$,*

$$-L\rho(x_1 - x_2) \leq w(x_2) - w(x_1) \leq L\rho(x_2 - x_1).$$

In general, ALC functions are Lipschitz continuous with possibly different Lipschitz constants, while the opposite is not necessarily true.

For the ATO control problem, consider the following asymmetric seminorms. First, for all $p \in \mathbb{N}$, we take asymmetric seminorm ρ_0 on $\mathbb{R}^p$, such that $\rho_0(\boldsymbol{u}) := (\bar{a}+1)u_i^- + u_i^+$ for $\boldsymbol{u} = (u_1, u_2, \cdots, u_p)$. In particular, we can write

$$\rho_0(\boldsymbol{s}) = \rho_0(\boldsymbol{x}) + \rho_0(\tilde{\boldsymbol{u}}) + \rho_0(R) = \sum_{i=1}^{n} \rho_0(x_i) + \sum_{j=1}^{m} \rho_0(\tilde{u}_i) + \sum_{\ell=1}^{L_n}\sum_{i=1}^{n} \rho_0(R_{i\ell}). \tag{10}$$

We also take a different asymmetric seminorm ρ_1 on $\mathbb{R}^{p_S+p_A}$. Let

$$\rho_1(\boldsymbol{s}, \boldsymbol{a}) = \rho_0(\boldsymbol{x} - A\boldsymbol{y}) + \rho_0(\tilde{\boldsymbol{u}} - \boldsymbol{y}) + \rho_0(R) + \rho_0(\boldsymbol{z})$$

for $\boldsymbol{s} = (\boldsymbol{x}, \tilde{\boldsymbol{u}}, R)$ and $\boldsymbol{a} = (\boldsymbol{y}, \boldsymbol{z})$. It is clear that the functions ρ_0 and ρ_1 are asymmetric seminorms on $\mathbb{R}^{p_S}$ and $\mathbb{R}^{p_S+p_A}$, respectively.

Theorem 4 (THE ALC PROPERTY)**.** *If w is an $\bar{c}/(1-\gamma)$-ALC function on $\mathcal{SA}_\alpha$, then so is $\mathcal{T}^* w$.*

The goal of the ALC property is to imply Lipschitz continuity of the optimal value function.

Corollary 1. *For any distribution G on $\mathcal{D}$, $V_\alpha^*(\,\cdot\,; G)$ is L_v-Lipschitz continuous on $\mathcal{S}_\alpha$ for*

$$L_v := \frac{\bar{c}(\bar{a}+1)p_S}{1-\gamma}.$$

The use of ALC in Theorem 4 is unusual but necessary, as illustrated by the following example.

Example 1. *Consider an ATO system, where a single product is assembled from two components with unit lead times. Assembly takes one unit of each component. There are no arriving demand. Holding and backlog costs have rate 1. We consider discount factor $\gamma = 0.9$ and damp rate $\alpha = 0.1$.*

First, note that Theorem 4 fails under common notions of Lipschitz continuity. Take

$$\boldsymbol{s}^1 := (\boldsymbol{x}^1, \tilde{u}^1, R^1) = ((1,1), 1, (0,0)), \quad \boldsymbol{s}^2 := (\boldsymbol{x}^2, \tilde{u}^2, R^2) = ((2,2), 2, (0,0)),$$
$$\boldsymbol{a}^1 := (y, \boldsymbol{z}) = (1, (0,0)), \quad \boldsymbol{a}^2 := (y^2, \boldsymbol{z}^2) = (0, (0,0)), \quad \text{and} \quad w(\boldsymbol{s}, \boldsymbol{a}) := L(x_1 + x_2 + \tilde{u} + y).$$

Note that the Lipschitz constant depends on the choice of norm. For the commonly used L^1 norm and L^∞ norm, it is easy to check that w is $(L, \|\cdot\|_1)$-Lipschitz and $(4L, \|\cdot\|_\infty)$-Lipschitz. However,

$$\|(\boldsymbol{s}^1, \boldsymbol{a}^1) - (\boldsymbol{s}^2, \boldsymbol{a}^2)\|_\infty = 1, \quad \|(\boldsymbol{s}^1, \boldsymbol{a}^1) - (\boldsymbol{s}^2, \boldsymbol{a}^2)\|_1 = 4,$$
$$\mathcal{T}^* w(\boldsymbol{s}^1, \boldsymbol{a}^1) = 0, \quad \text{and} \quad T^* w(\boldsymbol{s}^2, \boldsymbol{a}^2) = 6 + 4.86L.$$

There exists no L such that $\mathcal{T}^* w$ is $(L, \|\cdot\|_1)$-Lipschitz or $(4L, \|\cdot\|_\infty)$-Lipschitz continuous.

It is therefore significant that the Lipschitz constant does not necessarily exhibit any contraction property under the Bellman operator $\mathcal{T}^*$, despite the discount of cost and the damping of state. In contrast, the function w can be excluded from Lipschitz continuous classes by considering seminorms. In Theorem 4, w is not continuous with the proposed asymmetric seminorm ρ_1.

Next, a seminorm also needs to be asymmetric in order that the Lipschitz constant be preserved by Bellman operator. Take

$$\boldsymbol{s}^3 = (\boldsymbol{x}^3, \tilde{\boldsymbol{u}}^3, R^3) = ((1,2), 2, (0,0)), \quad w_2(\boldsymbol{s}, \boldsymbol{a}) = L(x_2 - x_1 + \tilde{u} - y), \quad \text{and}$$
$$\rho_1'(\boldsymbol{s}, \boldsymbol{a}) = \|\boldsymbol{x} - A\boldsymbol{y}\|_1 + \|\tilde{\boldsymbol{u}} - \boldsymbol{y}\|_1 + \sum_{i=1}^{n} \|R_i\|_1 = |x_1 - y| + |x_2 - y| + |\tilde{u} - y| + |R_1| + |R_2|.$$

Here seminorm ρ_1' is the symmetric version of ρ_1. We can check that

$$|w_2(\boldsymbol{s}, \boldsymbol{a})| \leq L\rho_1'(\boldsymbol{s}, \boldsymbol{a}), \forall (\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{SA}_\alpha,$$

indicating that w_2 is (L, ρ_1')-Lipschitz. However,

$$\rho_1'\big((\boldsymbol{s}^2, \boldsymbol{a}^2) - (\boldsymbol{s}^3, \boldsymbol{a}^2)\big) = 1, \quad \mathcal{T}^* w_2(\boldsymbol{s}_2, \boldsymbol{a}_2) = 2, \quad \text{and} \quad \mathcal{T}^* w_2(\boldsymbol{s}_3, \boldsymbol{a}_2) = 3 + 1.62L.$$

Therefore, there is again no L such that $\mathcal{T}^* w_2$ is (L, ρ_1')-Lipschitz continuous. In contrast, an L-ALC function with respect to asymmetric seminorm ρ_1 can decrease by $2L$ when x_1 increases by 1 (but can increase by at most L). We can check that w_2 is L-ALC. By Theorem 4 that $\mathcal{T}^* w_2$ is also L-ALC for sufficiently large L. Note that the choice of ρ_1 aligns well with the asymmetry between overage and underage in the assembly. An overage of component 1 results in a unit cost from holding the component in the inventory, while an underage of component 1 costs twice as much, i.e., by holding component 2 in the inventory and backlogging one unit of demand. It turns out that, with such asymmetry captured, a sufficiently large ALC constant is preserved by the Bellman's operator.

C Theoretical Results

Proposition 1 (BOUNDEDNESS). *For $\alpha \in (0,1)$, there exist proper $x_\alpha, u_\alpha, R_\alpha, y_\alpha, z_\alpha \in \mathbb{R}_+$ such that $f_\alpha(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d}) \in \mathcal{S}_\alpha$ holds whenever $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{SA}_\alpha$ and $\boldsymbol{d} \in \mathcal{D}$, where*

$$\mathcal{S}_\alpha := [0, x_\alpha]^n \times [0, u_\alpha]^m \times [0, R_\alpha]^{n \times L_n},$$
$$\mathcal{SA}_\alpha := \Big\{(\boldsymbol{s}, \boldsymbol{a}) : \boldsymbol{s} \in \mathcal{S}_\alpha, \;\; \boldsymbol{a} \in [0, y_\alpha]^m \times [0, z_\alpha]^n, \; \boldsymbol{a} \in \mathcal{A}(\boldsymbol{s})\Big\}.$$

Moreover, $(x_\alpha, u_\alpha, R_\alpha, y_\alpha, z_\alpha)$ increases as α decreases, becomes boundless as $\alpha \to 0^+$.

Lemma 1. *For any $\alpha \in (0,1)$, f_α is a linear function and $\mathcal{SA}_\alpha$ is a bounded convex set.*

Lemma 2 (THE BELLMAN OPERATOR). *For any G on $\mathcal{D}$ and $\alpha \in (0,1)$,*

(i) the operator $\mathcal{T}^*$ has a unique fixed point Q^*, namely, $\mathcal{T}^* Q^* = Q^*$.

(ii) the function $V_\alpha^*(\boldsymbol{s}; G)$ satisfies $V_\alpha^*(\boldsymbol{s}; G) = \min_{\boldsymbol{a} \in \mathcal{A}_\alpha(\boldsymbol{s})} Q^*(\boldsymbol{s}, \boldsymbol{a})$, for $\boldsymbol{s} \in \mathcal{S}_\alpha$.

Theorem 5 (PRESERVATION PROPERTY). *$\mathcal{T}^* w$ is convex whenever $w : \mathcal{SA}_\alpha \to \mathbb{R}$ is convex. Moreover, $\mathcal{T}^*(w) \in \mathcal{C}$ if $w \in \mathcal{C}$.*

Corollary 2. *$Q^* \in \mathcal{C}$ and $V_\alpha^*(\,\cdot\,; \widehat{F})$ is convex on $\mathcal{S}_\alpha$.*

Proposition 2. *For any sample $\widehat{D}$, Algorithm 1 is well-posed in the sense that*

(i) there exists a selection ξ_K that is measurable;

(ii) the output policy $\hat{\pi}^*(\widehat{D}) \in \Pi_\alpha \subseteq \Pi$ is admissible.

Lemma 3 (REPRESENTATION POWER OF ICNN). *There exists an ICNN family $\{Q_\phi : \phi \in \Phi\}$ such that*

$$\inf_{\phi \in \Phi} \|Q - Q_\phi\|_\infty \le \epsilon$$

holds for any Lipschitz convex function $Q \in \mathcal{C}$.

D Proofs

D.1 Preliminary

In this section, we prove Proposition 1 and Lemma 1.

Proof. Proof of Proposition 1. We take

$$u_\alpha = y_\alpha := 2\alpha^{-1}\bar{d}, \quad R_\alpha = z_\alpha := 3\alpha^{-1}\bar{a}L_n\bar{d}, \quad \text{and} \quad x_\alpha := 3\alpha^{-2}\bar{a}L_n\bar{d}.$$

Take an arbitrary $\boldsymbol{d} \in [0, \bar{d}]^m$ and $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{SA}$ with $\boldsymbol{s} = (\boldsymbol{x}, \tilde{\boldsymbol{u}}, R)$, $\boldsymbol{a} = (\boldsymbol{y}, \boldsymbol{z})$. The choice $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{SA}$ implies that

$$\begin{aligned} &\boldsymbol{x} \in [0, x_\alpha]^n, \quad \tilde{\boldsymbol{u}} \in [0, u_\alpha]^m, \quad R \in [0, R_\alpha]^{n \times L_n}, \\ &\boldsymbol{y} \in [0, u_\alpha]^n, \quad \boldsymbol{z} \in [0, z_\alpha]^m \quad \boldsymbol{y} \leq \tilde{\boldsymbol{u}}, \quad \text{and} \quad A^\top \boldsymbol{y} \leq \boldsymbol{x}. \end{aligned}$$

It holds that

$$\begin{aligned} &\boldsymbol{s}' = (\boldsymbol{x}', \tilde{\boldsymbol{u}}', R') = f_\alpha(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d}) \quad \text{for} \\ &\boldsymbol{x}' = (\boldsymbol{x} - A^\top \boldsymbol{y})(1-\alpha) + R\boldsymbol{e}_1, \quad \tilde{\boldsymbol{u}}' = (\tilde{\boldsymbol{u}} - \boldsymbol{y})(1-\alpha) + \boldsymbol{d}, \quad \text{and} \quad R' = \Gamma(R, z). \end{aligned} \tag{11}$$

Our goal is to show that

$$\boldsymbol{x}' \in [0, x_\alpha]^n, \quad \tilde{\boldsymbol{u}}' \in [0, u_\alpha]^m, \quad \text{and} \quad R' \in [0, R_\alpha]^{n \times L_n}.$$

Plugging the minimal and maximal values of $(\boldsymbol{x}, \tilde{\boldsymbol{u}}, \boldsymbol{y}, R)$, it is straightforward to check that

$$\begin{aligned} &\boldsymbol{x}' \leq (x_\alpha(1-\alpha) + R_\alpha)\mathbf{1} = x_\alpha \mathbf{1}, \quad \tilde{\boldsymbol{u}}' \leq (u_\alpha(1-\alpha) + \bar{d})\mathbf{1} \leq u_\alpha \mathbf{1}, \\ &\boldsymbol{x}' = (\boldsymbol{x} - A^\top \boldsymbol{y})(1-\alpha) \geq \mathbf{0}, \quad \text{and} \quad \tilde{\boldsymbol{u}}' = (\tilde{\boldsymbol{u}} - \boldsymbol{y})\alpha \geq \mathbf{0}, \end{aligned}$$

where the last two inequalities utilize $A^\top \boldsymbol{y} \leq \boldsymbol{x}$ and $\boldsymbol{y} \leq \tilde{\boldsymbol{u}}$, respectively. As $\boldsymbol{z}$, R, U_1, and U_2 have nonnegative entries, it is clear that

$$R' = RU_1 + diag(\boldsymbol{z})U_2 \geq O.$$

Finally, the choices of U_1, U_2, and $R \in \mathcal{R}$ imply that RU_1 and $diag(\boldsymbol{z})U_2$ has no common nonzero entry, so that $R \in [0, R_\alpha]^{n \times (L_n - 1)}$ and $z_\alpha = R_\alpha$ imply

$$(R')_{j,\ell} = (RU_1 + diag(\boldsymbol{z})U_2)_{j,\ell} \leq R_\alpha, \quad \text{for all } j \in [n] \quad \text{and} \quad \ell \in [L_n].$$

Therefore, $R' \in [0, R_\alpha]^{n \times L_n}$, finishing the proof. □

Proof of Lemma 1. Let U_1 be an $L_n \times L_n$ 0-1 matrix such that $(U_1)_{ij} = 1$ if $j = i - 1$. Let U_2 be an $n \times L_n$ 0-1 matrix such that $(U_2)_{ij} = 1$ when $j = L_i - 1$. Define linear mapping

$$\Gamma(R, \boldsymbol{z}) := RU_1 + diag(\boldsymbol{z})U_2.$$

Here, $diag(\boldsymbol{z})$ returns an n by n diagonal matrix generated from $\boldsymbol{z}$. We can write

$$\begin{aligned} &f(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d}) := \Big((\boldsymbol{x} - A^\top \boldsymbol{y}) + R\boldsymbol{e}_1,\ (\tilde{\boldsymbol{u}} - \boldsymbol{y}) + \boldsymbol{d},\ \Gamma(R, \boldsymbol{z}) \Big), \quad \text{for} \\ &\boldsymbol{s} = (\boldsymbol{x}, \tilde{\boldsymbol{u}}, R) \in \mathbb{R}^{p_S}, \quad \boldsymbol{a} = (\boldsymbol{y}, \boldsymbol{z}) \in \mathbb{R}^{p_A}, \quad \text{and} \quad \boldsymbol{d} \in \mathbb{R}^m. \end{aligned} \tag{12}$$

In particular, mapping f is linear.

Next, the boundedness of $\mathcal{SA}_\alpha$ clearly follows from Proposition 1. It follows from

$$\mathcal{SA}_\alpha = \{(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{S}_\alpha \times \mathcal{A}_\alpha : A^\top \boldsymbol{y} \leq \boldsymbol{x} \quad \text{and} \quad \boldsymbol{y} \leq \tilde{\boldsymbol{u}}\}$$

that $\mathcal{SA}_\alpha$ is a convex polyhedron. □

D.2 The Shape Preservation Properties

In this section, we present our original proofs to Theorems 5 and 4, which establish that Bellman operator $\mathcal{T}^*$ preserves boundedness, convexity, and asymmetric Lipschitz continuity.

Proof. Proof of Theorem 5. We first show the boundedness of the operator $\mathcal{T}^*$. Take

$$B_\alpha = \bar{c}(m+n)x_\alpha/(1-\gamma).$$

For $(\boldsymbol{s},\boldsymbol{a})\in\mathcal{SA}_\alpha$, (7) gives

$$0\le C(\boldsymbol{s},\boldsymbol{a})=\boldsymbol{h}^\top(\boldsymbol{x}-A^\top\boldsymbol{y})+\boldsymbol{b}^\top(\tilde{\boldsymbol{u}}-\boldsymbol{y}_1)\le mx_\alpha\|\boldsymbol{h}\|_\infty+nu_\alpha\|\boldsymbol{b}\|_\infty,\quad \text{for all } (\boldsymbol{s},\boldsymbol{a})\in\mathcal{SA}_\alpha.$$

$$(\mathcal{T}^*w)(\boldsymbol{s},\boldsymbol{a}):=C(\boldsymbol{s},\boldsymbol{a})+\gamma\mathbb{E}_{\hat{\boldsymbol{d}}\sim G}\left[\min_{\boldsymbol{a}'\in\mathcal{A}_\alpha(\boldsymbol{s}')}w(\boldsymbol{s}',\boldsymbol{a}')\right],\quad \text{for}\quad \boldsymbol{s}'=f_\alpha(\boldsymbol{s},\boldsymbol{a},\hat{\boldsymbol{d}}),\quad \boldsymbol{s}\in\mathcal{S}_\alpha. \tag{13}$$

Together with (13) and $u_\alpha\le x_\alpha$,

$$\mathcal{T}^*[w(\boldsymbol{s},\boldsymbol{a})]=C(\boldsymbol{s},\boldsymbol{a})+\gamma\mathbb{E}_{\boldsymbol{d}}\big[\min_{\boldsymbol{a}'\in\mathcal{A}_\alpha(\boldsymbol{s}')}w(\boldsymbol{s}',\boldsymbol{a}')\big]\le\bar{c}(m+n)x_\alpha+\gamma B_\alpha\le B_\alpha,\quad \text{for all } (\boldsymbol{s},\boldsymbol{a})\in\mathcal{SA}_\alpha.$$

Therefore, $\mathcal{T}^*w$ is bounded by B_α if w is bounded by B_α.

For the convex property of $\mathcal{T}^*$, we take an arbitrary convex function w on $\mathcal{SA}_\alpha$, take arbitrary state and action pairs $(\boldsymbol{s}_i,\boldsymbol{a}_i)\in\mathcal{SA}_\alpha$, $i=1,2$ and let

$$\boldsymbol{s}_3=\frac{\boldsymbol{s}_1+\boldsymbol{s}_2}{2}\quad\text{and}\quad\boldsymbol{a}_3=\frac{\boldsymbol{a}_1+\boldsymbol{a}_2}{2}.$$

Temporarily write

$$\boldsymbol{s}_i':=f(\boldsymbol{s}_i,\boldsymbol{a}_i,\boldsymbol{d}),\quad c_i:=C(\boldsymbol{s}_i,\boldsymbol{a}_i),\quad\text{for}\quad i=1,2,3,\quad \boldsymbol{a}_i'\in\arg\min_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s}_i')}w(\boldsymbol{s}_i',\boldsymbol{a})\quad\text{for}\quad i=1,2.$$

Note that the minimal is obtained as $w(\boldsymbol{s}_i',\ \cdot\ )$ is convex on the compact set $\mathcal{A}_\alpha(\boldsymbol{s}_i')$. It follows from the linearity of f and C that

$$\boldsymbol{s}_3'=(\boldsymbol{s}_1'+\boldsymbol{s}_2')/2\quad\text{and}\quad c_3=(c_1+c_2)/2.$$

Clearly $(\boldsymbol{s}_i',\boldsymbol{a}_i')\in\mathcal{SA}_\alpha$ for $i=1,2$. It follows from Lemma 1 that $\mathcal{SA}_\alpha$ is convex, implying that $(\boldsymbol{s}_3',\boldsymbol{a}_3')\in\mathcal{SA}_\alpha$ and, in turn, $\boldsymbol{a}_3'\in\mathcal{A}_\alpha(\boldsymbol{s}_3')$ for $\boldsymbol{a}_3':=(\boldsymbol{a}_1'+\boldsymbol{a}_2')/2$. Therefore,

$$\begin{aligned}&\mathcal{T}^*w(\boldsymbol{s}_1,\boldsymbol{a}_1)+\mathcal{T}^*w(\boldsymbol{s}_2,\boldsymbol{a}_2)=c_1+c_2+\gamma\mathbb{E}_{\boldsymbol{d}}\left[w(\boldsymbol{s}_1',\boldsymbol{a}_1')+w(\boldsymbol{s}_2',\boldsymbol{a}_2')\right]\\ \ge&c_1+c_2+2\gamma\mathbb{E}_{\boldsymbol{d}}\big[w(\boldsymbol{s}_3',\boldsymbol{a}_3')\big]\ge 2c_3+2\gamma\mathbb{E}_{\boldsymbol{d}}\big[\min_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s}_3')}\big(\boldsymbol{s}_3',\boldsymbol{a}\big)\big]=2\mathcal{T}^*w(\boldsymbol{s}_3,\boldsymbol{a}_3),\end{aligned}$$

in which first inequality results from the convexity of w. Because the choice of $(\boldsymbol{s}_i,\boldsymbol{a}_i)$ is arbitrary, $\mathcal{T}^*w$ is convex on $\mathcal{SA}_\alpha$ whenever w is convex on $\mathcal{SA}_\alpha$, finishing the proof of Theorem 5. □

The proof of Theorem 4 is based on Lemma 4, a supporting result that relates the asymmetric norms ρ_1 to ρ_0.

Lemma 4. *For any $\alpha\in[0,1)$, $\boldsymbol{s}_1\in\mathcal{S}_\alpha$, $\boldsymbol{s}_2\in\mathcal{S}_\alpha$, it holds that*

$$\max_{\boldsymbol{a}_1\in\mathcal{A}_\alpha(\boldsymbol{s}_1)}\min_{\boldsymbol{a}_2\in\mathcal{A}_\alpha(\boldsymbol{s}_2)}\rho_1\big((\boldsymbol{s}_2-\boldsymbol{s}_1,\boldsymbol{a}_2-\boldsymbol{a}_1)\big)\le\rho_0(\boldsymbol{s}_2-\boldsymbol{s}_1). \tag{14}$$

Note that the maximum and minimum is attained as ρ_1 is a continuous function and $\mathcal{A}_\alpha(\boldsymbol{s})$ is compact for $\boldsymbol{s}\in\mathcal{S}_\alpha$.

Proof. Proof of Theorem 4. Take an arbitrary L-ALC function w on $\mathcal{SA}_\alpha$, $(\boldsymbol{s}_1,\boldsymbol{a}_1)\in\mathcal{SA}_\alpha$, and $(\boldsymbol{s}_2,\boldsymbol{a}_2)\in\mathcal{SA}_\alpha$.

First, for the single period cost function C in (7), we can compute that.

$$\begin{aligned}C(\boldsymbol{s}_2,\boldsymbol{a}_2)-C(\boldsymbol{s}_1,\boldsymbol{a}_1)=&\boldsymbol{h}^\top(\boldsymbol{x}_2-\boldsymbol{x}_1)-\boldsymbol{h}^\top A^\top(\boldsymbol{y}_2-\boldsymbol{y}_1)+\boldsymbol{b}^\top(\tilde{\boldsymbol{u}}_2-\tilde{\boldsymbol{u}}_1-\boldsymbol{y}_2+\boldsymbol{y}_1)\\ \le&\bar{c}\rho_0(\boldsymbol{x}_2-\boldsymbol{x}_1-A^\top(\boldsymbol{y}_2-\boldsymbol{y}_1))+\bar{c}\rho_0(\tilde{\boldsymbol{u}}_2-\tilde{\boldsymbol{u}}_1-\boldsymbol{y}_2+\boldsymbol{y}_1)\\ \le&\bar{c}\rho_1(\boldsymbol{s}_2-\boldsymbol{s}_1,\boldsymbol{a}_2-\boldsymbol{a}_1).\end{aligned}$$

Second, we temporarily denote $\boldsymbol{s}_i':=f_\alpha(\boldsymbol{s}_i,\boldsymbol{a}_i,\boldsymbol{d})$ for $i=1,2$. By (13),

$$\begin{aligned}&\mathcal{T}^*w(\boldsymbol{s}_2,\boldsymbol{a}_2)-\mathcal{T}^*w(\boldsymbol{s}_1,\boldsymbol{a}_1)\\ =&C(\boldsymbol{s}_2,\boldsymbol{a}_2)-C(\boldsymbol{s}_1,\boldsymbol{a}_1)+\gamma\mathbb{E}\Big[\min_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s}_2')}w(\boldsymbol{s}_2',\boldsymbol{a})-\min_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s}_1')}w(\boldsymbol{s}_1',\boldsymbol{a})\Big]\\ =&C(\boldsymbol{s}_2,\boldsymbol{a}_2)-C(\boldsymbol{s}_1,\boldsymbol{a}_1)+\gamma\mathbb{E}\Big[\max_{\boldsymbol{a}_1'\in\mathcal{A}_\alpha(\boldsymbol{s}_1')}\min_{\boldsymbol{a}_2'\in\mathcal{A}_\alpha(\boldsymbol{s}_2')}w(\boldsymbol{s}_2',\boldsymbol{a}_2')-w(\boldsymbol{s}_1',\boldsymbol{a}_1')\Big]\\ \le&\bar{c}\rho_1(\boldsymbol{s}_2-\boldsymbol{s}_1,\boldsymbol{a}_2-\boldsymbol{a}_1)+\gamma L\mathbb{E}\Big[\max_{\boldsymbol{a}_1'\in\mathcal{A}_\alpha(\boldsymbol{s}_1')}\min_{\boldsymbol{a}_2'\in\mathcal{A}_\alpha(\boldsymbol{s}_2')}\rho_1(\boldsymbol{s}_2'-\boldsymbol{s}_1',\boldsymbol{a}_2'-\boldsymbol{a}_1')\Big]\\ \le&\bar{c}\rho_1(\boldsymbol{s}_2-\boldsymbol{s}_1,\boldsymbol{a}_2-\boldsymbol{a}_1)+\gamma L\mathbb{E}[\rho_0(\boldsymbol{s}_2'-\boldsymbol{s}_1')]\\ =&\bar{c}\rho_1(\boldsymbol{s}_2-\boldsymbol{s}_1,\boldsymbol{a}_2-\boldsymbol{a}_1)+\gamma L\rho_0(f_\alpha(\boldsymbol{s}_2-\boldsymbol{s}_1,\boldsymbol{a}_2-\boldsymbol{a}_1,\boldsymbol{0})).\end{aligned} \tag{15}$$

Note that the first inequality follows from the fact that C is $\bar{c}$-ALC and w is L, the second inequality results from Lemma 4, and the last equality in (15) uses the linearity of f_α.

Next, we expand the expression of f_α in the right-hand side of (15). For

$$(\boldsymbol{x}', \tilde{\boldsymbol{u}}', R') := f_\alpha(\boldsymbol{s}_2 - \boldsymbol{s}_1, \boldsymbol{a}_2 - \boldsymbol{a}_1, \boldsymbol{0}),$$

the expressions in (11) gives

$$\begin{aligned}
\boldsymbol{x}' &= (1-\alpha)(\boldsymbol{x}_2 - \boldsymbol{x}_1 - A^\top(\boldsymbol{y}_2 - \boldsymbol{y}_1)) + (R_2 - R_1)\boldsymbol{e}_1, \\
\tilde{\boldsymbol{u}}' &= (1-\alpha)(\tilde{\boldsymbol{u}}_2 - \tilde{\boldsymbol{u}}_1 - (\boldsymbol{y}_2 - \boldsymbol{y}_1)), \quad \text{and} \\
R' &= (R_2 - R_1)U_1 + U_2 diag(\boldsymbol{z}_2 - \boldsymbol{z}_1).
\end{aligned}$$

Using the triangular inequality for asymmetric seminorm ρ_0 and dropping the scalar $(1-\alpha)$,

$$\begin{aligned}
\rho_0(\boldsymbol{x}') &\le \rho_0(\boldsymbol{x}_2 - \boldsymbol{x}_1 - A^\top(\boldsymbol{y}_2 - \boldsymbol{y}_1)) + \rho_0((R_2 - R_1)\boldsymbol{e}_1), \\
\rho_0(\tilde{\boldsymbol{u}}') &\le \rho_0(\tilde{\boldsymbol{u}}_2 - \tilde{\boldsymbol{u}}_1 - (\boldsymbol{y}_2 - \boldsymbol{y}_1)), \quad \text{and} \\
\rho_0(R') &\le \rho_0((R_2 - R_1)U_1) + \rho_0(U_2 diag(\boldsymbol{z}_2 - \boldsymbol{z}_1)).
\end{aligned}$$

Expanding the expressions of matrices U_1 and U_2,

$$\rho_0((R_2 - R_1)U_1) = \sum_{\ell=2}^{L_n} \rho_0((R_2 - R_1)\boldsymbol{e}_\ell) = \rho_0(R_2 - R_1) \quad \text{and} \quad \rho_0(U_2 diag(\boldsymbol{z}_2 - \boldsymbol{z}_1)) = \rho_0(\boldsymbol{z}_2 - \boldsymbol{z}_1).$$

To summarize, we have shown that

$$\begin{aligned}
&\rho_0(f_\alpha(\boldsymbol{s}_2 - \boldsymbol{s}_1, \boldsymbol{a}_2 - \boldsymbol{a}_1, \boldsymbol{0})) \\
=&\rho_0(\boldsymbol{x}') + \rho_0(\tilde{\boldsymbol{u}}') + \rho_0(R') \\
\le&\rho_0(\boldsymbol{x}_2 - \boldsymbol{x}_1 - A^\top(\boldsymbol{y}_2 - \boldsymbol{y}_1)) + \rho_0(\tilde{\boldsymbol{u}}_2 - \tilde{\boldsymbol{u}}_1 - (\boldsymbol{y}_2 - \boldsymbol{y}_1)) + \rho_0(\boldsymbol{z}_2 - \boldsymbol{z}_1) + \rho_0(R_2 - R_1) \\
=&\rho_1(\boldsymbol{s}_2 - \boldsymbol{s}_1, \boldsymbol{a}_2 - \boldsymbol{a}_1).
\end{aligned}$$

Together with (15),

$$\mathcal{T}^* w(\boldsymbol{s}_2, \boldsymbol{a}_2) - \mathcal{T}^* w(\boldsymbol{s}_1, \boldsymbol{a}_1) \le (\bar{c} + \gamma L)\rho_1(\boldsymbol{s}_1 - \boldsymbol{s}_2, \boldsymbol{a}_1 - \boldsymbol{a}_2), \quad \text{for all } (\boldsymbol{s}_i, \boldsymbol{a}_i) \in \mathcal{SA}_\alpha, \quad i = 1, 2.$$

implying that $\mathcal{T}w$ is a $(\bar{c} + \gamma L)$-ALC function. □

The proof of Lemma 4 is based on the following linear algebra result.

Lemma 5. *Let p by q matrix $H = (h_{ij})$ satisfy $h_{ij} \in \{0\} \cup [1, \infty)$ for all $i \in [p]$ and $j \in [q]$. Then for all $\boldsymbol{w} \in \mathbb{R}^q_+$ and $\boldsymbol{b} \in \mathbb{R}^p_+$ such that $\boldsymbol{b} \le H\boldsymbol{w}$, there exists $0 \le \boldsymbol{v} \le \boldsymbol{w}$ such that*

$$H\boldsymbol{v} \le \boldsymbol{b} \quad \textit{and} \quad \|\boldsymbol{b} - H\boldsymbol{v}\|_1 \le \|\boldsymbol{1}^\top H\|_\infty \, \|H\boldsymbol{w} - \boldsymbol{b}\|_1. \tag{16}$$

Proof of Lemma 4. Fix states $\boldsymbol{s}_1 := (\boldsymbol{x}_1, \tilde{\boldsymbol{u}}_1, R_1) \in \mathcal{S}_\alpha$, $\boldsymbol{s}_2 := (\boldsymbol{x}_2, \tilde{\boldsymbol{u}}_2, R_2) \in \mathcal{S}_\alpha$, an arbitrary action $\boldsymbol{a}_1 := (\boldsymbol{y}_1, \boldsymbol{z}_1) \in \mathcal{A}_\alpha(\boldsymbol{s}_1)$. Our goal is to find an action $\boldsymbol{a}_2 = (\boldsymbol{y}_2, \boldsymbol{z}_2) \in \mathcal{A}_\alpha(\boldsymbol{s}_2)$ such that

$$\rho_1(\boldsymbol{s}_2 - \boldsymbol{s}_1, \boldsymbol{a}_2 - \boldsymbol{a}_1) \le \rho_0(\boldsymbol{s}_2 - \boldsymbol{s}_1). \tag{17}$$

We take $\boldsymbol{z}_2 = \boldsymbol{z}_1$ and find $\boldsymbol{y}_2$ by Lemma 5. Let $\boldsymbol{b}_i := (\boldsymbol{x}_i, \tilde{\boldsymbol{u}}_i)$ for $i = 1, 2$, $H = [A, I]^\top$ and $\boldsymbol{b}'_2 = \min\{H\boldsymbol{y}_1, \boldsymbol{b}_2\}$. It is straightforward to check that H satisfies the condition for Lemma 5. It is also clear that $\boldsymbol{b}'_2 \le H\boldsymbol{y}_1 \in \mathbb{R}^{m+n}_+$ and $\boldsymbol{y}_2 \in \mathbb{R}^m_+$. In Lemma 5 we take $\boldsymbol{b} = \boldsymbol{b}'_2$ and $\boldsymbol{w} = \boldsymbol{y}_1$, then there exists $0 \le \boldsymbol{y}_2 \le \boldsymbol{y}_1 \in \mathbb{R}^m_+$ such that

$$H\boldsymbol{y}_2 \le \boldsymbol{b}'_2, \quad \text{and} \quad \boldsymbol{1}^\top(\boldsymbol{b}'_2 - H\boldsymbol{y}_2) \le \|\boldsymbol{1}^\top H\|_\infty (H\boldsymbol{y}_1 - \boldsymbol{b}'_2). \tag{18}$$

Because $H\boldsymbol{y}_2 \le \boldsymbol{b}'_2 \le \boldsymbol{b}_2$, it is easy to check that $A^\top \boldsymbol{y}_2 \le \boldsymbol{x}_2$ and $\boldsymbol{y}_2 \le \tilde{\boldsymbol{u}}_2$, implying that $\boldsymbol{a}_2 \in \mathcal{A}_\alpha(\boldsymbol{s}_2)$. To show (17), the choice of ρ_1 gives

$$\rho_1(\boldsymbol{s}_2 - \boldsymbol{s}_1, \boldsymbol{a}_2 - \boldsymbol{a}_1) = \rho_0(\boldsymbol{b}_2 - \boldsymbol{b}_1 - H(\boldsymbol{y}_2 - \boldsymbol{y}_1)) + \rho_0(R_2 - R_1). \tag{19}$$

For the first term on the right-hand side of (19), the triangular inequality of ρ_0 gives

$$\rho_0(\boldsymbol{b}_2 - \boldsymbol{b}_1 - H(\boldsymbol{y}_2 - \boldsymbol{y}_1)) \le \rho_0(\boldsymbol{b}_2 - \boldsymbol{b}'_2 - \boldsymbol{b}_1 + H\boldsymbol{y}_1) + \rho_0(\boldsymbol{b}'_2 - H\boldsymbol{y}_2).$$

To obtain an upper bound of $\rho_0(\boldsymbol{b}'_2 - H\boldsymbol{y}_2)$, the second inequality in (18) gives $\boldsymbol{b}'_2 - H\boldsymbol{y}_2 \ge \boldsymbol{0}$. It follows from the choice of ρ_0 and the second inequality in (18) that

$$\rho_0(\boldsymbol{b}'_2 - H\boldsymbol{y}_2) = \|\boldsymbol{b}'_2 - H\boldsymbol{y}_2\|_1 \le \|\boldsymbol{1}^\top H\|_\infty \, \|H\boldsymbol{y}_1 - \boldsymbol{b}'_2\|_1,$$

Because $\boldsymbol{b}_2' - H\boldsymbol{y}_1 \leq \mathbf{0}$, the choice of ρ_0 gives

$$\rho_0(\boldsymbol{b}_2' - H\boldsymbol{y}_1) = \|\mathbf{1}^\top H\|_\infty (H\boldsymbol{y}_1 - \boldsymbol{b}_2'),$$

and thus

$$\rho_0(\boldsymbol{b}_2' - H\boldsymbol{y}_2) \leq \rho_0(\boldsymbol{b}_2' - H\boldsymbol{y}_1).$$

Therefore, we have shown that

$$\rho_0(\boldsymbol{b}_2 - \boldsymbol{b}_1 - H(\boldsymbol{y}_2 - \boldsymbol{y}_1)) \leq \rho_0(\boldsymbol{b}_2' - H\boldsymbol{y}_1) + \rho_0(\boldsymbol{b}_2 - \boldsymbol{b}_2' - \boldsymbol{b}_1 + H\boldsymbol{y}_1).$$

We next show that the two vectors on the right-hand side have the same sign in every entry, i.e.,

$$(\boldsymbol{b}_2' - H\boldsymbol{y}_1)_\ell (\boldsymbol{b}_2 - \boldsymbol{b}_2' - \boldsymbol{b}_1 + H\boldsymbol{y}_1)_\ell \geq 0, \quad \text{for all } \ell \in [m+n], \tag{20}$$

It is clear that $\boldsymbol{b}_2' - H\boldsymbol{y}_1 = \min\{0, \boldsymbol{b}_2 - H\boldsymbol{y}_1\} \leq 0$. For any index $\ell \in [m+n]$ such that $\boldsymbol{b}_2' - H\boldsymbol{y}_1 = \min\{0, \boldsymbol{b}_2 - H\boldsymbol{y}_1\}$ has a strictly negative ℓth entry, it must hold that

$$(\boldsymbol{b}_2 - H\boldsymbol{y}_1)_\ell = (\boldsymbol{b}_2' - H\boldsymbol{y}_1)_\ell < 0.$$

As $H\boldsymbol{y}_1 \leq \boldsymbol{b}_1$ by $\boldsymbol{a}_1 \in \mathcal{A}_\alpha(\boldsymbol{s}_1)$, we have shown that $(\boldsymbol{b}_2 - \boldsymbol{b}_2' - \boldsymbol{b}_1 + H\boldsymbol{y}_1)_\ell = (H\boldsymbol{y}_1 - \boldsymbol{b}_1)_\ell \leq 0$, giving (20).

Now the choice of ρ_0 and (20) imply that

$$\rho_0(\boldsymbol{b}_2' - H\boldsymbol{y}_1) + \rho_0(\boldsymbol{b}_2 - \boldsymbol{b}_2' - \boldsymbol{b}_1 + H\boldsymbol{y}_1) = \rho_0(\boldsymbol{b}_2 - \boldsymbol{b}_1),$$

and we have thus computed that

$$\rho_0(\boldsymbol{b}_2 - \boldsymbol{b}_1 - H(\boldsymbol{y}_2 - \boldsymbol{y}_1)) \leq \rho_0(\boldsymbol{b}_2 - \boldsymbol{b}_1) = \rho_0(\boldsymbol{x}_2 - \boldsymbol{x}_1) + \rho_0(\tilde{\boldsymbol{u}}_2 - \tilde{\boldsymbol{u}}_1).$$

Together with (19) and (10), we obtain (17), finishing the proof. □

Proof of Lemma 5. Take $\boldsymbol{w}_0 = \boldsymbol{w}$ and recursively define the sequence $\{\boldsymbol{w}_i : i \in [p]\}$ such that $\boldsymbol{w}_i$ maximizes $\mathbf{1}^\top \boldsymbol{w}'$ over $0 \leq \boldsymbol{w}' \leq \boldsymbol{w}_{i-1}$ subject to $(H\boldsymbol{w}')_i \leq b_i$. Here b_i is the ith entry of vector $\boldsymbol{b}$. We claim that $\boldsymbol{v} = \boldsymbol{w}_p$ satisfies (16). To show this, it follows from $\boldsymbol{w}_p \leq \boldsymbol{w}_i$ and that H is a nonnegative matrix that $(H\boldsymbol{v})_i = (H\boldsymbol{w}_p)_i \leq (H\boldsymbol{w}_i)_i \leq b_i$ for all $i \in [p]$, indicating that $H\boldsymbol{v} \leq \boldsymbol{b}$.

To prove the second inequality in (16), it follows from $H\boldsymbol{v} \leq \boldsymbol{b} \leq H\boldsymbol{w}$ that

$$\begin{aligned}\|\boldsymbol{b} - H\boldsymbol{v}\|_1 \leq & \|H\boldsymbol{w} - H\boldsymbol{v}\|_1 = \mathbf{1}^\top H(\boldsymbol{w} - \boldsymbol{v}) \leq \|\mathbf{1}^\top H\|_\infty \|\boldsymbol{w} - \boldsymbol{v}\|_1 \\ = & \sum_{i=1}^{p} \|\mathbf{1}^T H\|_\infty \|\boldsymbol{w}_{i-1} - \boldsymbol{w}_i\|_1. \end{aligned} \tag{21}$$

Here, the last equality follows from $\boldsymbol{w} = \boldsymbol{w}_0 \geq \boldsymbol{w}_1 \geq \cdots \geq \boldsymbol{w}_p = \boldsymbol{v}$. Because $\boldsymbol{w}_i$ maximizes $\mathbf{1}^\top \boldsymbol{w}'$ subject to $0 \leq \boldsymbol{w}' \leq \boldsymbol{w}_{i-1}$ and $(H\boldsymbol{w}')_i = \sum_{j=1}^q h_{ij} w_j' \leq b_i$, the optimality condition gives

$$1\{h_{ij} = 0\}(\boldsymbol{w}_i - \boldsymbol{w}_{i-1})_j = 0 \quad \text{for all } j \in [q] \quad \text{and} \quad (H\boldsymbol{w}_i)_i = \min\{b_i, (H\boldsymbol{w}_{i-1})_i\}.$$

Together with the fact that $h_{ij} \in \{0\} \cup [1, \infty)$ gives

$$\|\boldsymbol{w}_{i-1} - \boldsymbol{w}_i\|_1 = \sum_{j \in [q]} 1\{h_{ij} \neq 0\}(\boldsymbol{w}_{i-1} - \boldsymbol{w}_i)_j \leq \sum_{j \in [q]} h_{ij}(\boldsymbol{w}_{i-1} - \boldsymbol{w}_i) = (H\boldsymbol{w}_{i-1} - H\boldsymbol{w}_i)_i.$$

We can compute the right-hand side by

$$(H\boldsymbol{w}_{i-1} - H\boldsymbol{w}_i)_i = (H\boldsymbol{w}_{i-1})_i - (H\boldsymbol{w}_i)_i = (H\boldsymbol{w}_{i-1})_i - \min\{(H\boldsymbol{w}_{i-1})_i, b_i) = \max\{0, (H\boldsymbol{w}_{i-1} - \boldsymbol{b})_i\}.$$

Together with $\boldsymbol{w}_{i-1} \leq \boldsymbol{w}$ and $H\boldsymbol{w} \geq \boldsymbol{b}$, we have shown that

$$\|\boldsymbol{w}_{i-1} - \boldsymbol{w}_i\|_1 \leq (H\boldsymbol{w}_{i-1} - H\boldsymbol{w}_i)_i = \max\{0, (H\boldsymbol{w}_{i-1} - \boldsymbol{b})_i\} \leq (H\boldsymbol{w} - \boldsymbol{b})_i, \quad \text{for all } i \in [p].$$

Plugging in (21), we obtain the second inequality in (16), finishing the proof. □

D.3 Contraction of Bellman Operators

In this section, we state and prove Proposition 3, an extended version of Lemma 2. The statements in Proposition 3 are mostly standard. The proofs of Corollaries 2 and 1 also appear at this section.

We introduce the following notations.

$$\mathcal{A}_\alpha(\boldsymbol{s}) := \{\boldsymbol{a} \in \mathbb{R}^{p_A} : (\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{SA}_\alpha\} = \mathcal{A}(\boldsymbol{s}) \cap \left([0, y_\alpha]^m \times [0, z_\alpha]^n\right) \quad \text{for} \quad \boldsymbol{s} \in \mathcal{S}_\alpha. \tag{22}$$

First, define Markovian control ξ by a Borel measurable map from $\mathcal{S}_\alpha$ to $\mathbb{R}_+^{p_A}$, such that $\xi(\boldsymbol{s}) \in \mathcal{A}_\alpha(\boldsymbol{s})$ for $\boldsymbol{s} \in \mathcal{S}_\alpha$, where we recall $\mathcal{A}_\alpha$ from (22).

$$\boldsymbol{s}_\alpha(t+1) = f_\alpha(\boldsymbol{s}_\alpha(t), \boldsymbol{a}(t), \boldsymbol{d}(t+1)), \quad \text{for all } t \geq 0. \tag{23}$$

We say that policy $\pi \in \Pi_\alpha$ is induced by ξ if $\pi = \{\boldsymbol{a}(t)\}$ is recursively characterized by $\boldsymbol{a}(t) = \xi(\boldsymbol{s}_\alpha(t))$ and $\{\boldsymbol{s}_\alpha(t)\}$ from (23). Second, define Bellman operator $\mathcal{T}^\xi$ for Markovian policy $\xi \in \Xi$ by

$$\mathcal{T}^\xi[w(\boldsymbol{s}, \boldsymbol{a})] := C(\boldsymbol{s}, \boldsymbol{a}) + \gamma \mathbb{E}_{\boldsymbol{d} \sim G}[w(\boldsymbol{s}', \xi(\boldsymbol{s}'))], \tag{24}$$

for $(\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{SA}_\alpha$ and $\boldsymbol{s}' = f(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d})$. The following result summarizes all the properties that we obtain for the Bellman operators $\mathcal{T}^*$ and $\mathcal{T}^\xi$ for later reference.

Proposition 3. *For any $\alpha \in (0,1)$ and distribution G on $\mathcal{D}$,*

(i) operators $\mathcal{T}^\xi$ and $\mathcal{T}^*$ are contraction mappings, i.e., for any w_1 and w_2 that are measurable on $\mathcal{SA}_\alpha$,

$$\|\mathcal{T}^\xi(w_1) - \mathcal{T}^\xi(w_2)\|_\infty \le \gamma\|w_1 - w_2\|_\infty \quad \textit{and} \quad \|\mathcal{T}^*(w_1) - \mathcal{T}^*(w_2)\|_\infty \le \gamma\|w_1 - w_2\|_\infty.$$

Here, the infinite norm $\|\cdot\|_\infty$ is defined by

$$\|w\|_\infty = \sup_{(\boldsymbol{s},\boldsymbol{a})\in\mathcal{SA}_\alpha} |w(\boldsymbol{s},\boldsymbol{a})|;$$

(ii) operator $\mathcal{T}^\xi$ has a unique fixed point $Q^\xi : \mathcal{SA}_\alpha \to \mathbb{R}$, i.e., $\mathcal{T}^\xi(Q^\xi) = Q^\xi$. Moreover, Q^ξ satisfies $V_\alpha^\pi(\boldsymbol{s}; G) = Q^\xi(\boldsymbol{s}, \xi(\boldsymbol{s}))$ for all $\boldsymbol{s} \in \mathcal{S}_\alpha$, where π is the induced policy by ξ.

(iii) operator $\mathcal{T}^*$ has a unique fixed point $Q^* \in \mathcal{C}$. For $w_0 \in \mathcal{C}$, the recursively defined sequence $\{w_k : k \ge 0\}$ by $w_{k+1} := \mathcal{T}^* w_k$ satisfies $w_k \in \mathcal{C}$ and

$$\|w_k - Q^*\|_\infty \le \gamma^k \|w_0 - Q^*\|_\infty, \quad \textit{for} \quad k \in \mathbb{Z}_+. \tag{25}$$

(iv) there exists an optimal Markovian control $\xi^* \in \Xi$ satisfying $Q^{\xi^*} = Q^*$ and

$$V_\alpha^{\pi^*}(\boldsymbol{s}; G) = V_\alpha^*(\boldsymbol{s}; G) = \min_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s})} Q^*(\boldsymbol{s},\boldsymbol{a}), \quad \textit{for all } \boldsymbol{s} \in \mathcal{S}_\alpha. \tag{26}$$

Here $\pi^* \in \Pi_\alpha$ is the induced policy by ξ^*.

Note that all results in Corollary 2 and Lemma 2 immediately follow from Proposition 3.

Proof. Proof of Proposition 3.

Proof of Assertion (i): It follows from Blackwell's sufficient conditions (see [5]) that assertion (i) is implied by the following monotonicity and discounting properties of the operator $\mathcal{T} \in \{\mathcal{T}^\xi, \mathcal{T}^*\}$:

- *Monotonicity*: If $w(\boldsymbol{s},\boldsymbol{a}) \ge w'(\boldsymbol{s},\boldsymbol{a})$ for all $(\boldsymbol{s},\boldsymbol{a}) \in \mathcal{SA}_\alpha$ where w, w' are measurable functions on $\mathcal{SA}_\alpha$, then $\mathcal{T}(w)(\boldsymbol{s},\boldsymbol{a}) \ge \mathcal{T}(w')(\boldsymbol{s},\boldsymbol{a})$.
- *Discounting*: For all measurable function w on $\mathcal{SA}_\alpha$, $\mathcal{T}(w+b)(\boldsymbol{s},\boldsymbol{a}) \le \mathcal{T}(w)(\boldsymbol{s},\boldsymbol{a}) + \gamma b$ for all $b \ge 0$ and $(\boldsymbol{s},\boldsymbol{a}) \in \mathcal{SA}_\alpha$. Here $(w+b)(\boldsymbol{s},\boldsymbol{a}) = w(\boldsymbol{s},\boldsymbol{a}) + b$.

For the operator $\mathcal{T}^\xi$, the two properties follows from

$$T^\xi w(\boldsymbol{s},\boldsymbol{a}) = C(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}_{\boldsymbol{d}}[w(\boldsymbol{s}', \xi(\boldsymbol{s}'))] \ge C(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}_{\boldsymbol{d}}[w'(\boldsymbol{s}', \xi(\boldsymbol{s}'))] = \mathcal{T}^\xi w'(\boldsymbol{s},\boldsymbol{a}), \quad \text{and}$$
$$\mathcal{T}^\xi(w+b)(\boldsymbol{s},\boldsymbol{a}) = C(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}_{\boldsymbol{d}}[(w(\boldsymbol{s}', \xi(\boldsymbol{s}')) + b)] = \mathcal{T}^\xi w(\boldsymbol{s},\boldsymbol{a}) + \gamma b,$$

It also follows from similar arguments that the operator $\mathcal{T}^*$ also has the two properties, finishing the proof of assertion (i).

Proof of Assertion (ii): It follows from the contraction property of $\mathcal{T}^\xi$ in Assertion (i) and the Banach Fixed Point Theorem (see Theorem 1.1 in [1]) that $\mathcal{T}^\xi$ has a unique fixed point Q^ξ. To show that $Q^\xi(\boldsymbol{s}, \xi(\boldsymbol{s})) = V_\alpha^\pi(\boldsymbol{s}; F)$ for induced policy $\pi = \{\boldsymbol{a}(t)\}$ and all $\boldsymbol{s}$, temporarily define $Q_V : \mathcal{SA}_\alpha \to \mathbb{R}$ by

$$Q_V(\boldsymbol{s},\boldsymbol{a}) := C(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}_{\boldsymbol{d}}[V_\alpha^{\pi'}(f_\alpha(\boldsymbol{s},\boldsymbol{a},\boldsymbol{d}); G)],$$

for $\pi' := \{\boldsymbol{a}(t+1) : t \ge 0\}$. As a brief explanation, $Q_V(\boldsymbol{s},\boldsymbol{a})$ is the actor-value function if the system takes action $\boldsymbol{a}$ in the first decision period and then follows policy π. The fact that ξ induces $\pi = \{\boldsymbol{a}(t) : t \ge 0\}$ and the choice of π' imply $\xi(\boldsymbol{s}) = \boldsymbol{a}(0)$ and

$$Q_V(\boldsymbol{s}, \xi(\boldsymbol{s})) := C(\boldsymbol{s}, \boldsymbol{a}(0)) + \gamma\mathbb{E}_{\boldsymbol{d}}[V_\alpha^{\pi'}(f_\alpha(\boldsymbol{s}, \boldsymbol{a}(0), \boldsymbol{d}); G)] = V_\alpha^\pi(\boldsymbol{s}; G), \quad \text{for all } \boldsymbol{s} \in \mathcal{S}_\alpha. \tag{27}$$

Together with (24),

$$\mathcal{T}^\xi Q_V(\boldsymbol{s},\boldsymbol{a}) = C(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}_{\boldsymbol{d}}[Q_V(\boldsymbol{s}', \xi(\boldsymbol{s}')] = C(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}_{\boldsymbol{d}}[V_\alpha^\pi(\boldsymbol{s}'; G)] = Q_V(\boldsymbol{s},\boldsymbol{a}),$$

for all $(\boldsymbol{s},\boldsymbol{a}) \in \mathcal{SA}_\alpha$ and $\boldsymbol{s}' = f(\boldsymbol{s},\boldsymbol{a},\boldsymbol{d})$. As a result, Q_V is a fixed point of $\mathcal{T}^\pi$ and the uniqueness of Q^ξ implies $Q_V = Q^\xi$. Using (27) again,

$$Q^\xi(\boldsymbol{s}, \xi(\boldsymbol{s})) = Q_V(\boldsymbol{s}, \xi(\boldsymbol{s})) = V_\alpha^\pi(\boldsymbol{s}; G) \quad \text{for all } \boldsymbol{s} \in \mathcal{S}_\alpha.$$

Proof of Assertion (iii): It follows similar from the Banach fixed point theorem to the proof of Assertion (ii) that $\mathcal{T}^*$ has a unique fixed point Q^*, where (25) immediately follows from Assertion (i). The fact that $w_k \in \mathcal{C}$

for all $\mathcal{C}$ follows immediately from Theorem 5. Because $\mathcal{C}$ is a closed set with respect to norm $\|\cdot\|_\infty$, we obtain $Q^* \in \mathcal{C}$ as the limit point of $\{w_k\}$.

Proof of Assertion (iv): It follows from Assertion (iii) that Q^* is convex, which is thus lower semi-continuous. Also, the set $\mathcal{A}_\alpha(\boldsymbol{s})$ is closed for any $\boldsymbol{s} \in \mathcal{S}_\alpha$. By Proposition 7.33 in [4] there exists a Markovian control $\xi^* \in \Xi$ such that

$$\xi^*(\boldsymbol{s}) \in \arg\min_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s})} Q^*(\boldsymbol{s},\boldsymbol{a}) \quad \text{for all } \boldsymbol{s} \in \mathcal{S}_\alpha.$$

To show that $Q^* = Q^{\xi^*}$, for any $(\boldsymbol{s},\boldsymbol{a}) \in \mathcal{SA}_\alpha$ and $\boldsymbol{s}' = f(\boldsymbol{s},\boldsymbol{a},\boldsymbol{d})$, we have

$$\begin{aligned}\mathcal{T}^{\xi^*}[Q^*(\boldsymbol{s},\boldsymbol{a})] &= C(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}_{\boldsymbol{d}}\big[Q^*(\boldsymbol{s}',\xi^*(\boldsymbol{s}'))\big] \\ &= C(\boldsymbol{s},\boldsymbol{a}) + \gamma\mathbb{E}_{\boldsymbol{d}}\big[\inf_{\boldsymbol{a}'\in\mathcal{A}(\boldsymbol{s}')} Q^*(\boldsymbol{s}',\boldsymbol{a}')\big] = \mathcal{T}^*Q^*(\boldsymbol{s},\boldsymbol{a}) = Q^*(\boldsymbol{s},\boldsymbol{a}),\end{aligned}$$

As a result, Q^* is a fixed point of $\mathcal{T}^{\xi^*}$, implying that $Q^* = Q^{\xi^*}$.

We next adopt existing results in [4], Proposition 9.8 to characterize $V_\alpha^*(\boldsymbol{s};G)$ for all $\boldsymbol{s} \in \mathcal{S}_\alpha$. First, the DSAA control problem is an infinite horizon stochastic optimal control problem as in Definition 9.1 in the reference under the discounted case. Apply Proposition 9.8, we obtain

$$V_\alpha^*(\boldsymbol{s};G) = \inf_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s})} C(\boldsymbol{s},\boldsymbol{a}) + \gamma E_{\boldsymbol{d}\sim G}[V_\alpha^*(f_\alpha(\boldsymbol{s},\boldsymbol{a},\boldsymbol{d}))], \quad \text{for all } \boldsymbol{s} \in \mathcal{S}_\alpha.$$

Temporarily define

$$Q_V^*(\boldsymbol{s},\boldsymbol{a}) := C(\boldsymbol{s},\boldsymbol{a}) + \gamma E_{\boldsymbol{d}\sim G}[V_\alpha^*(f_\alpha(\boldsymbol{s},\boldsymbol{a},\boldsymbol{d}))], \quad \text{for all } (\boldsymbol{s},\boldsymbol{a}) \in \mathcal{SA}_\alpha,$$

we can rewrite the equality by

$$\inf_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s})} Q_V^*(\boldsymbol{s},\boldsymbol{a}) = V_\alpha^*(\boldsymbol{s};G), \quad \text{for all } \boldsymbol{s} \in \mathcal{S}_\alpha.$$

Together with (24), it holds for all $(\boldsymbol{s},\boldsymbol{a}) \in \mathcal{SA}_\alpha$ that

$$\mathcal{T}^*Q_V^*(\boldsymbol{s},\boldsymbol{a}) = C(\boldsymbol{s},\boldsymbol{a}) + \gamma E_{\boldsymbol{d}\in\mathcal{G}}[V_\alpha^*(\boldsymbol{s}';G)] = Q_V^*(\boldsymbol{s},\boldsymbol{a}), \quad \text{for} \quad \boldsymbol{s}' = f_\alpha(\boldsymbol{s},\boldsymbol{a},\boldsymbol{d}).$$

Therefore, $Q_V^* = Q^*$, giving the second equality of (26). The first equality of (26) follows from

$$V_\alpha^{\pi^*}(\boldsymbol{s};G) = Q^{\xi^*}(\boldsymbol{s},\xi^*(\boldsymbol{s})) = Q^*(\boldsymbol{s},\xi^*(\boldsymbol{s})) = \min_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s})} Q^*(\boldsymbol{s},\boldsymbol{a}) = V_\alpha^*(\boldsymbol{s};G), \quad \text{for all } \boldsymbol{s} \in \mathcal{S}.$$

Here, the first two equalities follow from Assertion (ii) and $Q^* = Q^{\xi^*}$, respectively. □

Proof. Proof of Lemma 6 and Corollary 2.

The statements in Lemma 6 immediately follow from Proposition 3(iii) and (iv). To prove Corollary 2, it follows from Proposition 3(iii) that $Q^* \in C$ is convex. By (iv), $V_\alpha^*(\,\cdot\,;G)$ is convex. □

Proof. Proof of Corollary 1. Temporarily take $L_0 = \bar{c}/(1-\gamma)$. We take w_0 an L_0-ALC function on $\mathcal{SA}_\alpha$ (for example, the zeroth function) and recursively generate the sequence $\{w_k : k \geq 1\}$ as in Proposition 3(ii). By Theorem 4, w_k is L_0-ALC for all $k \geq 1$. Together with (25), we have shown that Q^* is L_0-ALC.

Next, Proposition 3(v) gives

$$V_\alpha^*(\boldsymbol{s};G) = \min_{\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s})} Q^*(\boldsymbol{s},\boldsymbol{a}), \quad \text{for all } \boldsymbol{s} \in \mathcal{S}_\alpha.$$

For any $\boldsymbol{s}_1 \in \mathcal{S}$ and $\boldsymbol{s}_2 \in \mathcal{S}$, Lemma 4 gives

$$\begin{aligned}V_\alpha^*(\boldsymbol{s}_2;G) - V_\alpha^*(\boldsymbol{s}_1;G) &= \min_{\boldsymbol{a}_2\in\mathcal{A}_\alpha(\boldsymbol{s}_2)} \max_{\boldsymbol{a}_1\in\mathcal{A}_\alpha(\boldsymbol{s}_1)} Q^*(\boldsymbol{s}_2,\boldsymbol{a}_2) - Q^*(\boldsymbol{s}_1,\boldsymbol{a}_1) \\ &\leq \min_{\boldsymbol{a}_2\in\mathcal{A}(\boldsymbol{s}_2)} \max_{\boldsymbol{a}_1\in\mathcal{A}(\boldsymbol{s}_1)} L_0\rho_1(\boldsymbol{s}_2-\boldsymbol{s}_1,\boldsymbol{a}_2-\boldsymbol{a}_1) \leq L_0\rho_0(\boldsymbol{s}_2-\boldsymbol{s}_1).\end{aligned}$$

In particular, the first inequality follows from the definition of L_0-ALC and the second follows from Lemma 4. Finally, it is clear from the choice of ρ_0 that, for $\boldsymbol{s} = (\boldsymbol{x},\tilde{\boldsymbol{u}},R)$,

$$\rho_0(\boldsymbol{s}) \leq (\bar{a}+1)\Big(\sum_{i=1}^n |x_i| + \sum_{j=1}^m |\tilde{u}_i| + \sum_{\ell=1}^{L_n}\sum_{i=1}^n |R_{i\ell}|\Big) \leq (\bar{a}+1)p_S\|\boldsymbol{s}\|_\infty.$$

As $L_v := (\bar{a}+1)p_SL_0 = \bar{c}(\bar{a}+1)p_S/(1-\gamma)$, function $V_\alpha^*(\,\cdot\,;G)$ is thus L_v-Lipschitz continuous. □

D.4 Proofs of Theorem 3.

In this section, we provide our proof of Theorem 3. Given Theorem 5, Theorem 4, and Proposition 3, most of the remaining arguments are standard, despite the proof is still long and technical. We hereby provide a roadmap and a list of supporting lemmas before we present the proof.

Proof Roadmap. The proof of Theorem 3 relies on an auxiliary Markovian policy and multiple versions of the value functions. Recall from Algorithm 1 that ξ_K is a measurable function from $\mathcal{S}_\alpha$ to $\mathbb{R}^{p_A}$. The map ξ_K induces a Markovian policy $\pi_K = \{\boldsymbol{a}(t)\}$, for $\boldsymbol{a}(t) = \xi_K(\boldsymbol{s}_\alpha(t))$ for all $t \geq 0$ and $\boldsymbol{s}_\alpha(t)$ from (23). Note that π_K is different from the output policy $\hat{\pi}^*(\widehat{D})$ of Algorithm 1 because policy $\hat{\pi}^*(\widehat{D})$ depends on ξ_K through the damped relation (3). For short, we write $\hat{\pi}^*$ for $\hat{\pi}^*(\widehat{D})$.

There are six different value functions with respect to three different transition dynamics. Take $V^{\hat{\pi}^*}(\,\cdot\,; F)$ and $V^*(\,\cdot\,; F)$ the value function of policy $\hat{\pi}^*$ and the optimal value function under the true demand distribution, as defined by (8) and (9), respectively. Take $V_\alpha^{\pi_K}(\,\cdot\,; F)$ and $V_\alpha^*(\,\cdot\,; F)$ the value function of policy π_K and the optimal value function of damped control problem under the true demand distribution, respectively. Here, the controlled dynamics evolves by (23). Finally, take $V_\alpha^{\pi_K}(\,\cdot\,; \widehat{F})$ and $V_\alpha^*(\,\cdot\,; \widehat{F})$ the value function of policy π_K and the optimal value function of the DSAA control problem, respectively, where the demand follows from the sampling distribution and controlled dynamics evolve by (23). It holds that

$$\begin{aligned} 0 \leq V^{\hat{\pi}^*}(\boldsymbol{s}; F) - V^*(\boldsymbol{s}; F) &= \big(V^{\hat{\pi}^*}(\boldsymbol{s}; F) - V_\alpha^{\pi_K}(\boldsymbol{s}; F)\big) \\ &\quad + \big(V_\alpha^{\pi_K}(\boldsymbol{s}; F) - V_\alpha^{\pi_K}(\boldsymbol{s}; \widehat{F})\big) + \big(V_\alpha^{\pi_K}(\boldsymbol{s}; \widehat{F}) - V_\alpha^*(\boldsymbol{s}; \widehat{F})\big) \\ &\quad + \big(V_\alpha^*(\boldsymbol{s}; \widehat{F}) - V_\alpha^*(\boldsymbol{s}; F)\big) + \big(V_\alpha^*(\boldsymbol{s}; F) - V^*(\boldsymbol{s}; F)\big). \end{aligned} \tag{28}$$

We need separate estimates of the five terms on the right-hand side of (28). Intuitively, the first and last terms on the right-hand side of (28) capture the error of using a damped system to approximate the nond-damped system. To bound the two terms, we employ coupling arguments in Section D.6 that align the damped system and a non-damped system in the same probability space and bound the difference of the cost function on each sample path. The second and fourth terms on the right-hand side of (28) capture the error of using a sampling distribution to approximate the unknown true distribution. To bound this source of error, we employ covering arguments in Section D.7 on the compact domain $\mathcal{S}_\alpha$ and then make use of an existing concentration result. Finally, the third term on the right-hand side of (28) gives the error of having a finite number of iterations in the CFQI algorithm. This source of error can be bounded by the contraction of the Bellman operator $\mathcal{T}^*$, see Section D.5. The following four supporting lemmas provide an upper bound for each of the five terms.

Lemma 6. *For any $\alpha \in (0,1)$ and $\boldsymbol{s} \in \mathcal{S}_\alpha$,*

$$0 \leq \sup_{\boldsymbol{s} \in \mathcal{S}_\alpha} \big(V_\alpha^{\pi_K}(\boldsymbol{s}; \widehat{F}) - V_\alpha^*(\boldsymbol{s}; \widehat{F})\big) = \frac{4\varepsilon_m}{(1-\gamma)^2}. \tag{29}$$

Lemma 7. *Take $\alpha \in (0,1)$ and*

$$\mathcal{E}_1(\delta) := \frac{3(\bar{a}+1)\bar{c}p_S}{(1-\gamma)^2}\sqrt{\frac{p_S+p_A}{2M}\log\frac{3M(\bar{a}+1)x_\alpha}{\delta\bar{d}}}.$$

For any $\pi \in \Pi_\alpha$, event

$$\|V_\alpha^\pi(\,\cdot\,; F) - V_\alpha^\pi(\,\cdot\,; \widehat{F})\|_\infty \leq \frac{\gamma\|V_\alpha^\pi(\,\cdot\,; \widehat{F}) - V_\alpha^*(\,\cdot\,; \widehat{F})\|_\infty}{1-\gamma} + \mathcal{E}_1(\delta) \tag{30}$$

has probability at least $1-\delta$. Also, event

$$\|V_\alpha^\pi(\,\cdot\,; F) - V_\alpha^\pi(\,\cdot\,; \widehat{F})\|_\infty \leq \frac{\gamma\|V_\alpha^\pi(\,\cdot\,; F) - V_\alpha^*(\,\cdot\,; F)\|_\infty}{1-\gamma} + \mathcal{E}_1(\delta) \tag{31}$$

has probability at least $1-\delta$. Here, the probability is taken with respect to the sample $\widehat{D}$.

Lemma 8. *For any $\alpha \in (0,1)$ and $\boldsymbol{s} \in \mathcal{S}_\alpha$,*

$$V^{\hat{\pi}^*}(\boldsymbol{s}; F) - V_\alpha^{\pi_K}(\boldsymbol{s}; F) \leq \frac{\alpha}{(1-\gamma)} V_\alpha^{\pi_K}(\boldsymbol{s}; F). \tag{32}$$

Lemma 9. *For any $\alpha \in (0,1)$ and $\boldsymbol{s} \in \mathcal{S}_\alpha$,*

$$V_\alpha^*(\boldsymbol{s}; F) - V^*(\boldsymbol{s}; F) \leq \frac{\alpha\bar{c}m(L_n+1)\bar{d}}{(1-\gamma)^2}. \tag{33}$$

Proof. Proof of Theorem 3. First, the choice of α in (5) indicates that $\boldsymbol{s} \in \mathcal{S}_\alpha$, so that Lemmas 6 – 9 are all applicable. Similar to the treatment in (28), we can expand the right-hand side of Lemma 8 by

$$\begin{aligned}\frac{\alpha}{1-\gamma}V_\alpha^{\pi_K}(\boldsymbol{s};F) = \frac{\alpha}{1-\gamma}V^*(\boldsymbol{s};F) + \frac{\alpha}{1-\gamma}\big(V_\alpha^{\pi_K}(\boldsymbol{s};F) - V_\alpha^{\pi_K}(\boldsymbol{s};\widehat{F})\big) + \frac{\alpha}{1-\gamma}\big(V_\alpha^{\pi_K}(\boldsymbol{s};\widehat{F}) - V_\alpha^*(\boldsymbol{s};\widehat{F})\big)\\ + \frac{\alpha}{1-\gamma}\big(V_\alpha^*(\boldsymbol{s};\widehat{F}) - V_\alpha^*(\boldsymbol{s};F)\big) + \frac{\alpha}{1-\gamma}\big(V_\alpha^*(\boldsymbol{s};F) - V^*(\boldsymbol{s};F)\big).\end{aligned}$$

Combine (28) with (32) and using $\alpha \le (1-\gamma)/\sqrt{M} \le 1-\gamma$, we obtain

$$\begin{aligned}0 \le& V^{\hat{\pi}^*}(\boldsymbol{s};F) - V^*(\boldsymbol{s};F)\\ \le& \frac{1}{\sqrt{M}}V^*(\boldsymbol{s};F) + 2|V_\alpha^{\pi_K}(\boldsymbol{s};F) - V_\alpha^{\pi_K}(\boldsymbol{s};\widehat{F})| + 2|V_\alpha^{\pi_K}(\boldsymbol{s};\widehat{F}) - V_\alpha^*(\boldsymbol{s};\widehat{F})|\\ &+ 2\big(V_\alpha^*(\boldsymbol{s};\widehat{F}) - V_\alpha^*(\boldsymbol{s};F)\big) + 2\big(V_\alpha^*(\boldsymbol{s};F) - V^*(\boldsymbol{s};F)\big),\end{aligned} \tag{34}$$

In Lemma 7, we take $\pi = \pi_K \in \Pi_\alpha$,

$$|V_\alpha^{\pi_K}(\boldsymbol{s};F) - V_\alpha^{\pi_K}(\boldsymbol{s};\widehat{F})| \le \frac{\gamma\|V_\alpha^{\pi_K}(\,\cdot\,;\widehat{F}) - V_\alpha^*(\,\cdot\,;\widehat{F})\|_\infty}{1-\gamma} + \mathcal{E}_1(\delta/2)$$

holds with probability at least $1-\delta/2$. Together with (29) in Lemma 6,

$$|V_\alpha^{\pi_K}(\boldsymbol{s};F) - V_\alpha^{\pi_K}(\boldsymbol{s};\widehat{F})| \le \frac{4\gamma\varepsilon_m}{(1-\gamma)^3} + \mathcal{E}_1(\delta/2) \tag{35}$$

holds with probability at least $1-\delta/2$.

In Lemma 7, we next take π such that $V_\alpha^\pi(\boldsymbol{s};F) = V_\alpha^*(\boldsymbol{s};F)$, the existence of such π follows from Proposition 3(v). We have

$$V_\alpha^*(\boldsymbol{s};\widehat{F}) - V_\alpha^*(\boldsymbol{s};F) \le V_\alpha^\pi(\boldsymbol{s};\widehat{F}) - V_\alpha^\pi(\boldsymbol{s};F).$$

Together with (31),

$$V_\alpha^*(\boldsymbol{s};\widehat{F}) - V_\alpha^*(\boldsymbol{s};F) \le \mathcal{E}_1(\delta/2) \tag{36}$$

holds with probability at least $1-\delta/2$.

Plug (35), (29), (36), and (33) in Lemma 9 to (34), we can compute that

$$\begin{aligned}0 \le& V^{\hat{\pi}^*}(\boldsymbol{s};F) - V^*(\boldsymbol{s};F)\\ \le& \frac{1}{\sqrt{M}}V^*(\boldsymbol{s};F) + \Big(2\mathcal{E}_1(\delta/2) + \frac{8\gamma\varepsilon_m}{(1-\gamma)^3}\Big) + \frac{8\varepsilon_m}{(1-\gamma)^2} + 2\mathcal{E}_1(\delta/2) + \frac{2\alpha\bar{c}m(L_n+1)\bar{d}}{(1-\gamma)^2}\\ \le& \frac{1}{\sqrt{M}}V^*(\boldsymbol{s};F) + 4\mathcal{E}_1(\delta/2) + \frac{8\varepsilon_m}{(1-\gamma)^3} + \frac{2\alpha\bar{c}m(L_n+1)\bar{d}}{(1-\gamma)^2}\end{aligned}$$

with probability at least $1-\delta$, where the last inequality utilizes the choice that $\alpha \le M^{-1}(1-\gamma)$. We take

$$\mathcal{E}(\delta) := 4\mathcal{E}_1(\delta/2) + \frac{2\alpha\bar{c}m(L_n+1)\bar{d}}{(1-\gamma)^2},$$

giving (5). Finally, using the expression of $\mathcal{E}_1(\delta)$ and x_α, it is clear to check that

$$\mathcal{E}(\delta) = O\Big(\frac{\bar{a}\bar{c}(p_S+p_A)^2}{(1-\gamma)^2}\sqrt{\frac{1}{M}\log\frac{\bar{a}M}{\alpha\delta}}\Big), \quad \text{for} \quad \alpha = \min\Big\{\frac{1-\gamma}{\sqrt{M}}, \frac{\bar{d}}{\|\boldsymbol{s}\|_\infty}\Big\},$$

where $O(\,\cdot\,)$ hides an "absolute" constant that does not depend on any model primitives. □

D.5 The Contraction Arguments

The proof of Lemma 6 mainly utilizes the contraction property of the Bellman operator $\mathcal{T}^*$. The main line of the proof is similar to [13], while the presence of the convex optimization penalty Λ gives an additional source of difficulty.

Proof of Lemma 6. In this proof, we fix the damping factor α and $G = \widehat{F}$ in Proposition 3. Note that the inequality in (29) is trivial and our goal is to prove the equality. Recall ξ_K and Q_K from Algorithm 1. It is clear that

$$Q_K(\boldsymbol{s},\xi_K(\boldsymbol{s})) \le Q_K(\boldsymbol{s},\xi(\boldsymbol{s})) \quad \text{and} \quad \mathcal{T}^{\xi_K}Q_K \le \mathcal{T}^\xi Q_K \quad \text{for all } \boldsymbol{s}\in\mathcal{S} \quad \text{and} \quad \xi\in\Xi. \tag{37}$$

Because π_K is the induced policy by ξ_K, it follows from Proposition 3(ii) that

$$V_\alpha^{\pi_K}(\boldsymbol{s};\widehat{F}) = Q^{\xi_K}(\boldsymbol{s},\xi_K(\boldsymbol{s})) \quad \text{for all } \boldsymbol{s}\in\mathcal{S}_\alpha.$$

Next, recall from Proposition 3(v) the optimal Markovian policy ξ^*. We similarly obtain from Proposition 3(ii) that

$$V_\alpha^*(\boldsymbol{s};\widehat{F}) = Q^{\xi^*}(\boldsymbol{s},\xi^*(\boldsymbol{s})) \quad \text{for all } \boldsymbol{s}\in\mathcal{S}_\alpha.$$

We can now compute that

$$\begin{aligned}
&V_\alpha^{\pi_K}(\boldsymbol{s};\widehat{F}) - V_\alpha^*(\boldsymbol{s};\widehat{F})\\
=&\big(Q^{\xi_K}(\boldsymbol{s},\xi_K(\boldsymbol{s})) - Q_K(\boldsymbol{s},\xi_K(\boldsymbol{s}))\big) + \big(Q_K(\boldsymbol{s},\xi_K(\boldsymbol{s})) - Q_K(\boldsymbol{s},\xi^*(\boldsymbol{s}))\big)\\
&+\big(Q_K(\boldsymbol{s},\xi^*(\boldsymbol{s})) - Q^{\xi^*}(\boldsymbol{s},\xi^*(\boldsymbol{s}))\big)\\
\le&\big(Q^{\xi_K}(\boldsymbol{s},\xi_K(\boldsymbol{s})) - Q_K(\boldsymbol{s},\xi_K(\boldsymbol{s}))\big) + \big(Q_K(\boldsymbol{s},\xi^*(\boldsymbol{s})) - Q^{\xi^*}(\boldsymbol{s},\xi^*(\boldsymbol{s}))\big)\\
\le&\|Q^{\xi_K} - Q_K\|_\infty + \|Q_K - Q^{\xi^*}\|_\infty\\
\le&\|Q^{\xi_K} - Q^{\xi^*}\|_\infty + 2\|Q_K - Q^{\xi^*}\|_\infty, \quad \text{for all } \boldsymbol{s}\in\mathcal{S}.
\end{aligned} \tag{38}$$

Here, the first inequality follows from taking $\xi = \xi^*$ in the first inequality of (37). The second and third inequalities are obtained by taking supreme over $\mathcal{SA}_\alpha$ and then adopting the triangular inequality. Next, it follows from Proposition 3(ii) and (iii) that $Q^{\xi_K} = \mathcal{T}^{\xi_K}Q^{\xi_K}$ and $Q^{\xi^*} = \mathcal{T}^{\xi^*}Q^{\xi^*}$. We can compute that

$$\begin{aligned}
0 &\le Q^{\xi_K} - Q^{\xi^*}\\
&= (\mathcal{T}^{\xi_K}Q^{\xi_K} - \mathcal{T}^{\xi_K}Q^{\xi^*}) + (\mathcal{T}^{\xi_K}Q^{\xi^*} - \mathcal{T}^{\xi_K}Q_K) + (\mathcal{T}^{\xi_K}Q_K - \mathcal{T}^{\xi^*}Q_K) + (\mathcal{T}^{\xi^*}Q_K - \mathcal{T}^{\xi^*}Q^{\xi^*})\\
&\le \|\mathcal{T}^{\xi_K}Q^{\xi_K} - \mathcal{T}^{\xi_K}Q^{\xi^*}\|_\infty + \|\mathcal{T}^{\xi_K}Q^{\xi^*} - \mathcal{T}^{\xi_K}Q_K\|_\infty + \|\mathcal{T}^{\xi^*}Q_K - \mathcal{T}^{\xi^*}Q^{\xi^*}\|_\infty\\
&\le \gamma\|Q^{\xi_K} - Q^{\xi^*}\|_\infty + 2\gamma\|Q^{\xi^*} - Q_K\|_\infty,
\end{aligned}$$

where the first inequality follows from taking $\xi = \xi^*$ in (37) and the second inequality follows from Proposition 3(i). As a result,

$$\|Q^{\xi_K} - Q^{\xi^*}\|_\infty \le \frac{2\gamma}{1-\gamma}\|Q^{\xi^*} - Q_K\|_\infty.$$

Together with (38),

$$\sup_{\boldsymbol{s}\in\mathcal{S}}\big(V_\alpha^{\pi_K}(\boldsymbol{s};\widehat{F}) - V_\alpha^*(\boldsymbol{s};\widehat{F})\big) \le \frac{2}{1-\gamma}\|Q^{\xi^*} - Q_K\|_\infty. \tag{39}$$

To estimate the right-hand side of (39), recall Q_k from Algorithm 1. For $k = [K]\cup\{0\}$, $\boldsymbol{s}\in\mathcal{S}_\alpha$ and $\boldsymbol{a}\in\mathcal{A}_\alpha(\boldsymbol{s})$, let

$$Q_k^\Lambda(\boldsymbol{s},\boldsymbol{a}) := Q_k(\boldsymbol{s},\boldsymbol{a}) + \Lambda(\boldsymbol{s},\boldsymbol{a};\eta^k) \quad \text{and} \quad \xi_k(\boldsymbol{s}) \in \arg\min_{\boldsymbol{a}\in\mathcal{A}(\boldsymbol{s})} Q_k^\Lambda(\boldsymbol{s},\boldsymbol{a}). \tag{40}$$

By (13) and (24), it holds that

$$\mathcal{T}^{\xi^*}Q_k^\Lambda \ge \mathcal{T}^*Q_k^\Lambda = \mathcal{T}^{\xi_k}Q_k^\Lambda, \quad \text{for all } k\in[K]\cup\{0\}. \tag{41}$$

Together with $Q^{\xi^*} = \mathcal{T}^{\xi^*}Q^{\xi^*}$, we can compute that

$$\begin{aligned}
Q^{\xi^*} - Q_{k+1}^\Lambda =&\big(\mathcal{T}^{\xi^*}Q^{\xi^*} - \mathcal{T}^{\xi^*}Q_k^\Lambda\big) + \big(\mathcal{T}^{\xi^*}Q_k^\Lambda - \mathcal{T}^{\xi_k}Q_k^\Lambda\big) + \big(\mathcal{T}^{\xi_k}Q_k^\Lambda - Q_{k+1}^\Lambda\big)\\
\ge&\mathcal{T}^{\xi^*}Q^{\xi^*} - \mathcal{T}^{\xi^*}Q_k^\Lambda + \xi_k,
\end{aligned}$$

where we take $\epsilon_k := \mathcal{T}^{\xi_k}Q_k^\Lambda - Q_{k+1}^\Lambda$, for $k\in[K-1]\cup\{0\}$. Together with Proposition 3(i),

$$Q^{\xi^*} - Q_{k+1}^\Lambda \ge -\gamma\|Q^{\xi^*} - Q_k^\Lambda\|_\infty - \|\epsilon_k\|_\infty.$$

It similarly follows from (24) that $\mathcal{T}^{\xi^*}Q^{\xi^*} \le \mathcal{T}^{\xi_k}Q^{\xi^*}$ for all $k\in[K]\cup\{0\}$, implying that

$$Q^{\xi^*} - Q_{k+1}^\Lambda = \big(\mathcal{T}^{\xi^*}Q^{\xi^*} - \mathcal{T}^{\xi_k}Q^{\xi^*}\big) + \big(\mathcal{T}^{\xi_k}Q^{\xi^*} - \mathcal{T}^{\xi_k}Q_k^\Lambda\big) + \epsilon_k \le \mathcal{T}^{\xi_k}Q^{\xi^*} - \mathcal{T}^{\xi_k}Q_k^\Lambda + \epsilon_k,$$

for $k\le K-1$. Together with Proposition 3(i),

$$Q^{\xi^*} - Q_{k+1}^\Lambda \le \gamma\|Q^{\xi^*} - Q_k^\Lambda\|_\infty + \|\epsilon_k\|_\infty.$$

In particular, we have shown that

$$\|Q^{\xi^*} - Q_{k+1}^\Lambda\|_\infty \le \gamma\|Q^{\xi^*} - Q_k^\Lambda\|_\infty + \|\epsilon_k\|_\infty, \quad \text{for all } k\in[K-1]\cup\{0\}.$$

By recursion,

$$\|Q^{\xi^*} - Q_K^\Lambda\|_\infty \le \sum_{k=0}^{K-1}\gamma^{K-k-1}\|\epsilon_k\|_\infty + \gamma^K\|Q^{\xi^*} - Q_0^\Lambda\|_\infty.$$

By Corollary 2, $Q^{\xi^*} \in \mathcal{C}$. Together with $Q_0 \in \mathcal{C}$ and $\|\Lambda(\,\cdot\,;\eta^k)\|_\infty \le \eta^k B_\alpha$ for all $k \ge 0$, it holds that

$$-2B_\alpha \le -Q_0 - B_\alpha \le Q^{\xi^*} - Q_0^\Lambda \le Q^{\xi^*} + B_\alpha \le 2B_\alpha.$$

Therefore, $\|Q^{\xi^*} - Q_0^\Lambda\|_\infty \le 2B_\alpha$. Together with $\|Q_K^\Lambda - Q_K\|_\infty \le \eta^k B_\alpha$,

$$\|Q^{\xi^*} - Q_K\|_\infty \le \sum_{k=0}^{K-1} \gamma^{K-k-1}\|\epsilon_k\|_\infty + 2B_\alpha\gamma^K + B_\alpha\eta^K.$$

Finally, the choice of Q^k as in Algorithm 1, line 2 gives $\|Q_{k+1} - \mathcal{T}^* Q_k^\Lambda\|_\infty \le \varepsilon_m$. Together with (41) and (40), we obtain

$$\|\epsilon_k\|_\infty \le \|Q_{k+1} - \mathcal{T}^{\xi_k} Q_k^\Lambda\|_\infty + \|\Lambda(\,\cdot\,;\eta^{k+1})\|_\infty \le \varepsilon_m + \eta^k B_\alpha,$$

for all $j \in [K]$. Therefore, for $\eta < \gamma < 1$, we obtain

$$\sum_{k=0}^{K-1} \gamma^{K-k-1}\|\epsilon_k\|_\infty \le \sum_{k=0}^{K-1} \gamma^{K-k-1}(\varepsilon_m + \eta^k B_\alpha) = \frac{\varepsilon_m(1-\gamma^K)}{1-\gamma} + \frac{(\gamma^K - \eta^K)B_\alpha}{\gamma - \eta}.$$

It follows from $\eta = \max\{\gamma/2, 2\gamma - 1\} < \gamma$ that $\gamma - \eta = \min\{\gamma/2, 1-\gamma\} \le \gamma(1-\gamma)/2$, and

$$2\gamma^K + \eta^K + \frac{\gamma^K - \eta^K}{\gamma - \eta} \le 2\gamma^K + \frac{\gamma^K}{\gamma - \eta} \le \frac{2\gamma^{K-1}}{1-\gamma} + \frac{\gamma^K}{\gamma/2(1-\gamma)} \le \frac{4\gamma^{K-1}}{1-\gamma},$$

implying that

$$\|Q^{\xi^*} - Q_K\|_\infty \le (\varepsilon_m + 4\gamma^{K-1}B_\alpha)/(1-\gamma).$$

Recall that

$$B_\alpha = \frac{\bar{c}(m+n)x_\alpha}{1-\gamma} = \frac{3\bar{a}\bar{c}\bar{d}L_n(m+n)}{\alpha^2(1-\gamma)}.$$

The choice of K gives $4\gamma^{K-1}B_\alpha \le \varepsilon_m$. Together with (39), we have shown (29), finishing the proof.

D.6 The Coupling Arguments

In this section, we provide coupling arguments to prove Lemmas 8 and 9. The proof of Proposition 2 also appears in this section. Throughout the section, all demands are sampled from the unknown distribution F.

Proof of Lemma 8. We consider the following coupling between two systems, both with initial state $\boldsymbol{s}(0)$. System 1 is controlled by policy $\hat{\pi}^* = \{\hat{\boldsymbol{a}}(t) : t \ge 0\}$ and has no damping, where the controlled dynamics $\{\widehat{\boldsymbol{s}}(t)\}$ evolves as

$$\widehat{\boldsymbol{s}}(t+1) = f(\widehat{\boldsymbol{s}}(t), \widehat{\boldsymbol{a}}(t), \boldsymbol{d}(t+1)) \quad \text{for} \quad \widehat{\boldsymbol{s}}(t) = (\widehat{\boldsymbol{x}}(t), \widehat{\boldsymbol{u}}(t), \widehat{R}(t)), \quad t \ge 0.$$

System 2 has damping factor α and is controlled by policy $\pi_K = \{\boldsymbol{a}(t) : t \ge 0\}$, where $\boldsymbol{a}(t) = \xi_K(\boldsymbol{s}(t))$ and the controlled state $\{\boldsymbol{s}(t)\}$ evolves as

$$\boldsymbol{s}(t+1) = f_\alpha(\boldsymbol{s}(t), \boldsymbol{a}(t), \boldsymbol{d}(t+1)), \quad \text{for} \quad \boldsymbol{s}(t) = (\boldsymbol{x}(t), \tilde{\boldsymbol{u}}(t), R(t)), \quad t \ge 0.$$

We can compute that

$$V^{\hat{\pi}^*}(\boldsymbol{s}; F) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t C(\widehat{\boldsymbol{s}}(t), \widehat{\boldsymbol{a}}(t))\Big] \quad \text{and} \quad V_\alpha^{\pi_K}(\boldsymbol{s}; F) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t C(\boldsymbol{s}(t), \boldsymbol{a}(t))\Big],$$

and thus

$$V^{\hat{\pi}^*}(\boldsymbol{s}; F) - V_\alpha^{\pi_K}(\boldsymbol{s}; F) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t C(\widehat{\boldsymbol{s}}(t) - \boldsymbol{s}(t), \widehat{\boldsymbol{a}}(t) - \boldsymbol{a}(t))\Big]. \tag{42}$$

We next show that

$$\begin{aligned} \widehat{\boldsymbol{s}}(t) - \boldsymbol{s}(t) &= \alpha \sum_{\ell=0}^{t-1} \big(\boldsymbol{x}(\ell) - A^\top \boldsymbol{y}(\ell), \tilde{\boldsymbol{u}}(\ell) - \boldsymbol{y}(\ell), O\big) \\ \boldsymbol{r}(t) &= \widehat{\boldsymbol{s}}(t) - \boldsymbol{s}(t), \quad \text{and} \quad \widehat{\boldsymbol{a}}(t) = \boldsymbol{a}(t) \quad \text{for all } t \ge 0, \end{aligned} \tag{43}$$

where the sequence $\boldsymbol{r}(t)$ is defined in (3). Proof by induction. When $t = 0$, it is clear that $\boldsymbol{s}(0) = \widehat{\boldsymbol{s}}(0)$ and (3) implies that $\widehat{\boldsymbol{a}}(0) = \boldsymbol{a}(0)$. Now take the induction hypothesis that (43) holds for some t_0 and $t = t_0 + 1$. For

the first equality in (43), we can compute that

$$\begin{aligned}
\widehat{\boldsymbol{s}}(t) &= f(\widehat{\boldsymbol{s}}(t_0), \widehat{\boldsymbol{a}}(t_0), \boldsymbol{d}(t_0+1)) \\
&= f\Big(\boldsymbol{s}(t_0) + \alpha \sum_{\ell=0}^{t_0-1} (\boldsymbol{x}(\ell) - A^\top \boldsymbol{y}(\ell), \tilde{\boldsymbol{u}}(\ell) - \boldsymbol{y}(\ell), O), \boldsymbol{a}(t_0), \boldsymbol{d}(t_0+1)\Big) \\
&= f(\boldsymbol{s}(t_0), \boldsymbol{a}(t_0), \boldsymbol{d}(t_0+1)) + \alpha \sum_{\ell=0}^{t_0-1} f\Big((\boldsymbol{x}(\ell) - A^\top \boldsymbol{y}(\ell), \tilde{\boldsymbol{u}}(\ell) - \boldsymbol{y}(\ell), O), \boldsymbol{0}, \boldsymbol{0}\Big) \\
&= f_\alpha(\boldsymbol{s}(t_0), \boldsymbol{a}(t_0), \boldsymbol{d}(t_0+1)) + \alpha \sum_{\ell=0}^{t_0+1} (\boldsymbol{x}(\ell) - A^\top \boldsymbol{y}(\ell), \tilde{\boldsymbol{u}}(\ell) - \boldsymbol{y}(\ell), O) \\
&= \boldsymbol{s}(t_0+1) + \alpha \sum_{\ell=0}^{t_0} (\boldsymbol{x}(\ell) - A^\top \boldsymbol{y}(\ell), \tilde{\boldsymbol{u}}(\ell) - \boldsymbol{y}(\ell), O) \\
&= \boldsymbol{s}(t) + \alpha \sum_{\ell=0}^{t-1} (\boldsymbol{x}(\ell) - A^\top \boldsymbol{y}(\ell), \tilde{\boldsymbol{u}}(\ell) - \boldsymbol{y}(\ell), O).
\end{aligned}$$

Here, note that the second equality is the induction hypothesis, the third equality uses the linearity of f, the fourth equality uses the fact that

$$f(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d}) = f_\alpha(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d}) + \alpha(\boldsymbol{x} - A^\top \boldsymbol{y}, \tilde{\boldsymbol{u}} - \boldsymbol{y}, O) \quad \text{and} \quad \boldsymbol{f}(\boldsymbol{s}, \boldsymbol{0}, \boldsymbol{0}) = \boldsymbol{s},$$

for all $\boldsymbol{s} = (\boldsymbol{x}, \tilde{\boldsymbol{u}}, \boldsymbol{d})$ and $\boldsymbol{a} = (\boldsymbol{y}, \boldsymbol{z})$. As a result, the first equality in (43) holds for $t = t_0 + 1$, which also implies that

$$\alpha(\boldsymbol{x}(t_0) - A^\top \boldsymbol{y}(t_0), \tilde{\boldsymbol{u}}(t_0) - \boldsymbol{y}(t_0), O) = \widehat{\boldsymbol{s}}(t) - \boldsymbol{s}(t) - \widehat{\boldsymbol{s}}(t_0) + \boldsymbol{s}(t_0). \tag{44}$$

To show that the second equality in (43) holds for $t = t_0 + 1$, observe that the induction hypothesis gives

$$\widehat{\boldsymbol{y}}(t_0) = \boldsymbol{y}(t_0) \quad \text{and} \quad \boldsymbol{r}(t_0) = \widehat{\boldsymbol{s}}(t_0) - \boldsymbol{s}(t_0).$$

Using the choice of $\boldsymbol{r}(t)$ in (3),

$$\begin{aligned}
\boldsymbol{r}(t) &= (1-\alpha)\boldsymbol{r}(t_0) + \alpha\big(\widehat{\boldsymbol{x}}(t_0) - A\widehat{\boldsymbol{y}}(t_0), \widehat{\boldsymbol{u}}(t_0) - \widehat{\boldsymbol{y}}(t_0), O\big) \\
&= (1-\alpha)\widehat{\boldsymbol{s}}(t_0) - \boldsymbol{s}(t_0) + \alpha\big(\boldsymbol{x}(t_0) - A\boldsymbol{y}(t_0), \tilde{\boldsymbol{u}}(t_0) - \boldsymbol{y}(t_0), O\big) + \alpha(\widehat{\boldsymbol{s}}(t_0) - \boldsymbol{s}(t_0)) \\
&= \widehat{\boldsymbol{s}}(t) - \boldsymbol{s}(t).
\end{aligned}$$

Here, the first equality follows from the induction hypothesis and the second equality utilizes (44). As a result, the second equality in (43) holds for $t = t_0 + 1$. Also, the choice of $\widehat{\boldsymbol{a}}(t)$ in (3) gives

$$\widehat{\boldsymbol{a}}(t) = \xi_K(\widehat{\boldsymbol{s}}(t) - \boldsymbol{r}(t)) = \xi_K(\boldsymbol{s}(t)) = \boldsymbol{a}(t), \quad \text{for} \quad t = t_0 + 1.$$

To summarize, we have shown that (43) holds for $t = t_0 + 1$. By induction, (43) holds for all $t \geq 0$.

We can now compute that, for all $t \geq 0$,

$$\begin{aligned}
& C(\widehat{\boldsymbol{s}}(t) - \boldsymbol{s}(t), \widehat{\boldsymbol{a}}(t) - \boldsymbol{a}(t)) \\
=& C(\alpha \sum_{\ell=0}^{t-1} (\boldsymbol{x}(\ell) - A^\top \boldsymbol{y}(\ell), \tilde{\boldsymbol{u}}(\ell) - \boldsymbol{y}(\ell), O), \boldsymbol{0}) \\
=& \alpha \sum_{\ell=0}^{t-1} \boldsymbol{h}^\top (\boldsymbol{x}(\ell) - A^\top \boldsymbol{y}(\ell)) + \boldsymbol{b}^\top (\tilde{\boldsymbol{u}}(\ell) - \boldsymbol{y}(\ell)) \\
=& \alpha \sum_{\ell=0}^{t-1} C(\boldsymbol{s}(\ell), \boldsymbol{a}(\ell)).
\end{aligned}$$

Notice that $C(\boldsymbol{s}(\ell), \boldsymbol{a}(\ell)) \geq 0$ holds for all t due to the expression of C in (7) and $\boldsymbol{a}(t) \in \mathcal{A}_\alpha(\boldsymbol{s}(t))$. Together with (42), we have shown that

$$\begin{aligned}
V^{\hat{\pi}^*}(\boldsymbol{s}; F) - V_\alpha^{\pi_K}(\boldsymbol{s}; F) =& \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t C(\widehat{\boldsymbol{s}}(t) - \boldsymbol{s}(t), \widehat{\boldsymbol{a}}(t) - \boldsymbol{a}(t))\Big] \\
=& \alpha \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \sum_{\ell=0}^{t-1} C(\boldsymbol{s}(\ell), \boldsymbol{a}(\ell))\Big] \\
=& \alpha \mathbb{E}\Big[\sum_{\ell=0}^{\infty} C(\boldsymbol{s}(\ell), \boldsymbol{a}(\ell)) \Big(\sum_{t=\ell+1}^{\infty} \gamma^t\Big)\Big] \\
\leq& \frac{\alpha}{1-\gamma} \mathbb{E}\Big[\sum_{\ell=0}^{\infty} C(\boldsymbol{s}(\ell), \boldsymbol{a}(\ell)) \gamma^\ell\Big] = \frac{\alpha}{1-\gamma} V_\alpha^{\pi_K}(\boldsymbol{s}; F),
\end{aligned} \tag{45}$$

finishing the proof. □

Proof of Proposition 2. Assertion (i) follows from Proposition 7.33 in [3] and that Q_K is a convex function.

We next prove Assertion (ii). Using the induction in the proof of Lemma 8, we have shown that $\widehat{\boldsymbol{s}}(t)-\boldsymbol{r}(t) = \boldsymbol{s}(t)$ and $\widehat{\boldsymbol{a}}(t) = \boldsymbol{a}(t)$ for all t. It is known that $\{\boldsymbol{a}(t)\}$ is a feasible policy for the damped control problem, that is, $\boldsymbol{a}(t) \in \mathcal{A}_\alpha(\boldsymbol{s}(t))$ and $\boldsymbol{a}(t) \in \mathcal{F}_t$. As a result,

$$\widehat{\boldsymbol{a}}(t) \in \mathcal{A}_\alpha(\widehat{\boldsymbol{s}}(t) - \boldsymbol{r}(t)).$$

By recursion, it is easy to show that, for any $t \geq 0$, $\widehat{\boldsymbol{s}}(t)$, $\boldsymbol{r}(t)$, $\widehat{\boldsymbol{a}}(t) = \boldsymbol{a}(t)$ are all independent of $(\boldsymbol{d}(t+1), \boldsymbol{d}(t+2), \cdots)$. As a result, $\{\widehat{\boldsymbol{a}}(t)\}$ is an admissible policy. □

Proof of Lemma 9. First, recall that

$$V^*(\boldsymbol{s}; F) = \inf_{\pi\in\Pi} V^\pi(\boldsymbol{s}; F).$$

We consider a subfamily of $\Pi_S \subseteq \Pi$ such that any policy $\pi = \{\boldsymbol{a}(t)\} \in \Pi_S$, in period t, can order a maximum number of components that is just sufficient to assemble products that meet the unsatisfied demand at period t, plus L_n additional periods of maximum demand. Formally, we require $\boldsymbol{z}(t) \leq A^\top(\tilde{\boldsymbol{u}}(t) + L_n\bar{d}\mathbf{1})$, for all $t \geq 0$. As there is no need at period t to order components for assembly more than $t + L_n$ periods later, and thus there is no loss of optimality by restricting to the subfamily Π_S, namely,

$$V^*(\boldsymbol{s}; F) = \inf_{\pi\in\Pi_S} V^\pi(\boldsymbol{s}; F).$$

Our goal is to show that, for any policy $\pi \in \Pi_S$, there is a corresponding policy $\pi_\alpha \in \Pi_\alpha$, such that

$$V_\alpha^{\pi_\alpha}(\boldsymbol{s}; F) \leq V^\pi(\boldsymbol{s}; F) + \frac{\alpha\bar{c}m(L_n+1)\bar{d}}{(1-\gamma)^2}. \tag{46}$$

Once (46) is proved, the lemma follows from taking the infimum over $\pi \in \Pi_S$.

We again consider a coupling between two systems, both with initial state $\boldsymbol{s}(0)$. System 1 is not damped and controlled by a fixed policy $\pi = \{\boldsymbol{a}(t)\} \in \Pi_S$. Recall from (3) that $\{\boldsymbol{s}(t)\}$ is the controlled state of system 1, which is specified by (6). System 2 has damping factor α and is controlled by policy $\pi_\alpha = \{\boldsymbol{a}_\alpha(t) : t \geq 0\}$ that we specify later. Denote by $\{\boldsymbol{s}_\alpha(t)\}$ the controlled state of system 2, namely,

$$\boldsymbol{s}_\alpha(t+1) = f_\alpha(\boldsymbol{s}_\alpha(t), \boldsymbol{a}_\alpha(t), \boldsymbol{d}(t+1)), \quad \text{for} \quad \boldsymbol{s}_\alpha(t) = (\boldsymbol{x}_\alpha(t), \tilde{\boldsymbol{u}}_\alpha(t), R_\alpha(t)), \quad \text{for all } t \geq 0.$$

To specify the coupling between $\boldsymbol{a}(t)$ and $\boldsymbol{a}_\alpha(t)$, take

$$\boldsymbol{a}(t) = (\boldsymbol{y}(t), \boldsymbol{z}(t)) \quad \text{and} \quad \boldsymbol{a}_\alpha(t) = \{\boldsymbol{y}_\alpha(t), \boldsymbol{z}_\alpha(t)\}.$$

Take $\bar{\alpha} := 1 - \alpha$, I_α be an n by n diagonal matrix, such that $(I_\alpha)_{ii} = \bar{\alpha}^{L_i}$. Let

$$\boldsymbol{y}_\alpha(t) = \bar{\alpha}^t\boldsymbol{y}(t) + \boldsymbol{y}'(t) \quad \text{and} \quad \boldsymbol{z}_\alpha(t) = \bar{\alpha}^t I_\alpha \boldsymbol{z}(t) + \boldsymbol{z}'(t).$$

where $\{\boldsymbol{y}'(t)\}$ and $\{\boldsymbol{z}'(t)\}$ are specified as follows. For $t = 1, 2, \cdots L_n - 1$, we take $\boldsymbol{y}'(t) = 0$. For $t \geq L_n$, we take $\boldsymbol{y}'(t) := \big(\bar{\alpha}^{L_n} - \bar{\alpha}^t\big)\boldsymbol{d}(t - L_n)$. The intuition of $\boldsymbol{y}_\alpha(t)$ is that the production of $\bar{\alpha}^t\boldsymbol{y}(t)$ components can only handle no more than $\bar{\alpha}^t\boldsymbol{d}(t)$ out of $\boldsymbol{d}(t)$ demands that arrive at period t. To satisfy the rest $(1 - \bar{\alpha}^t)\boldsymbol{d}(t)$ demands, the system schedules a production in period $t + L_n$ and prepares the required components during period $[t, t + L_n - 1]$. Notice that the system only needs to assemble $\boldsymbol{y}'(t)$ products due to the presence of damping. The choice of $\{\boldsymbol{z}'(t)\}$ is such that the demand arriving at period t matches the assembly need of $\boldsymbol{y}'(t)$, for all $t \geq L_n$. Formally,

$$\boldsymbol{z}_i'(t) := (A^\top\boldsymbol{y}'(t + L_i))_i = (\bar{\alpha}^{L_n} - \bar{\alpha}^{t+L_i})\big(A^\top\boldsymbol{d}(t - L_n + L_i)\big)_i, \quad \text{for} \quad t \geq L_n - L_i.$$

Denote $\boldsymbol{s}_\alpha(t) = (\boldsymbol{x}_\alpha(t), \tilde{\boldsymbol{u}}_\alpha(t), R_\alpha(t))$ for all $t \geq 0$. We first show that

$$\boldsymbol{x}_\alpha(t) = \bar{\alpha}^t\boldsymbol{x}(t) + A^\top\boldsymbol{y}'(t) \quad \text{and} \quad \tilde{\boldsymbol{u}}_\alpha(t) = \bar{\alpha}^t\tilde{\boldsymbol{u}}(t) + \sum_{s=t}^{t+L_n} \bar{\alpha}^{t-s}\boldsymbol{y}'(s). \tag{47}$$

Again proof by induction. First, it is clear that the equality holds when $t = 0$. Take an induction hypothesis that the equality holds for some t_0 and let $t = t_0 + 1$. Expanding f_α, the choice of $\boldsymbol{z}'(t)$ gives that

$$\tilde{\boldsymbol{u}}_\alpha(t) = \bar{\alpha}(\tilde{\boldsymbol{u}}_\alpha(t_0) - \boldsymbol{y}_\alpha(t_0)) + \boldsymbol{d}(t).$$

In particular, combining the induction hypothesis with the second equation gives

$$\begin{aligned}\tilde{\boldsymbol{u}}_\alpha(t) =&\bar{\alpha}\Big(\bar{\alpha}^{t-1}\tilde{\boldsymbol{u}}(t-1) + \sum_{s=t-1}^{t+L_n-1}\bar{\alpha}^{t-1-s}\boldsymbol{y}'(s) - \boldsymbol{y}_\alpha(t-1)\Big) + \boldsymbol{d}(t)\\ =&\bar{\alpha}^t(\tilde{\boldsymbol{u}}(t-1) - \boldsymbol{y}(t-1) + \boldsymbol{d}(t)) + \sum_{s=t-1}^{t+L_n-1}\bar{\alpha}^{t-s}\boldsymbol{y}'(s) - \bar{\alpha}\boldsymbol{y}'(t-1) + \boldsymbol{d}(t)(1-\bar{\alpha}^t)\\ =&\bar{\alpha}^t\tilde{\boldsymbol{u}}(t) + \sum_{s=t}^{t+L_n-1}\alpha^{t-s}\boldsymbol{y}'(s) + \bar{\alpha}^{-L_n}\boldsymbol{y}'(t+L_n)\\ =&\bar{\alpha}^t\tilde{\boldsymbol{u}}(t) + \sum_{s=t}^{t+L_n}\alpha^{t-s}\boldsymbol{y}'(s).\end{aligned}$$

Here, the last equality uses the choice of $\boldsymbol{y}'(t-1)$. Therefore, the second equality of (47) holds for $t = t_0 + 1$. For the first equality, the choice of $\boldsymbol{z}_\alpha$ results in that the component replenishment received at period t has two parts. The first part corresponds to the procurement $\{\bar{\alpha}^t I_\alpha \boldsymbol{z}(t)\}$, which thus equals $\bar{\alpha}^t R(t-1)\boldsymbol{e}_1$. The second part corresponds to the procurement $\{\boldsymbol{z}'(t)\}$, which equals $A^\top\boldsymbol{y}'(t)$. As a result, the induction hypothesis again gives

$$\begin{aligned}\boldsymbol{x}_\alpha(t) =&\bar{\alpha}(\boldsymbol{x}_\alpha(t_0) - A^\top\boldsymbol{y}_\alpha(t_0)) + \bar{\alpha}^t R(t-1)\boldsymbol{e}_1 + A^\top\boldsymbol{y}'(t)\\ =&\bar{\alpha}^t(\boldsymbol{x}(t-1) - A^\top\boldsymbol{y}(t-1)) + \bar{\alpha}^t R(t-1)\boldsymbol{e}_1 = \bar{\alpha}^t\boldsymbol{x}(t) + A^\top\boldsymbol{y}'(t).\end{aligned}$$

Therefore, (47) holds for $t = t_0 + 1$, and thus for all $t \geq 0$.

We next show that π_α is an admissible policy for the damped control problem, namely, $\pi_\alpha \in \Pi_\alpha$. for $\boldsymbol{a}_\alpha(t) := \{\boldsymbol{y}_\alpha(t), \boldsymbol{z}_\alpha(t))$. We need to show that (i) $(\boldsymbol{y}_\alpha(t), \boldsymbol{z}_\alpha(t))$ is independent of $(\boldsymbol{d}(t+1), \boldsymbol{d}(t+2).\cdots)$; and (ii)

$$\boldsymbol{a}_\alpha(t) \in \mathcal{A}_\alpha(\boldsymbol{s}_\alpha(t)), \quad \text{for all } t \geq 0.$$

Here, (i) is clear due to the choice of $\boldsymbol{a}_\alpha(t)$ and the admissibility of policy $\pi = \{\boldsymbol{a}(t)\}$. To show (ii), it is known that $A^\top\boldsymbol{y}(t) \leq \boldsymbol{x}(t)$ and $\boldsymbol{y}(t) \leq \tilde{\boldsymbol{u}}(t)$. Using the choices of $\boldsymbol{y}_\alpha(t)$ and (47), it holds that

$$A^\top\boldsymbol{y}_\alpha(t) \leq \boldsymbol{x}_\alpha(t) \quad \text{and} \quad \boldsymbol{y}_\alpha(t) \leq \tilde{\boldsymbol{u}}_\alpha(t). \tag{48}$$

Therefore, $\boldsymbol{a}_\alpha(t) \in \mathcal{A}(\boldsymbol{s}_\alpha(t))$. It is also clear that $y_\alpha(t) \geq 0$ and $z_\alpha(t) \geq 0$ for all $t \geq 0$. The choices of $\{\boldsymbol{y}'(t)\}$ and $\{\boldsymbol{z}'(t)\}$ give $\|\boldsymbol{y}'(t)\|_\infty \leq \bar{d}$ and $\|\boldsymbol{z}'(t)\| \leq \bar{a}\bar{d}$. Due to the choice that $\alpha \leq \bar{d}/\|\boldsymbol{s}\|_\infty$,

$$\|\boldsymbol{y}(t)\|_\infty \leq \|\tilde{\boldsymbol{u}}(0)\|_\infty + t\bar{d} \leq (t + \alpha^{-1})\bar{d}.$$

Because $\pi \in \Pi_S$, the additional constraint on procurement results in

$$\|\boldsymbol{z}(t)\|_\infty \leq \bar{a}(\|\tilde{\boldsymbol{u}}(0)\|_\infty + (t+L_n)\bar{d}) \leq \bar{a}(t + L_n + \alpha^{-1})\bar{d}.$$

We have the inequalities

$$\begin{aligned}&(t+\alpha^{-1})\bar{\alpha}^t \leq (t+\alpha^{-1})/(1+\alpha)^t \leq (t+\alpha^{-1})/(1+\alpha t) \leq \alpha^{-1}, \quad \text{for} \quad t \geq 0,\\ &\|\boldsymbol{y}_\alpha(t)\|_\infty \leq \bar{\alpha}^t\|\boldsymbol{y}(t)\| + \|\boldsymbol{y}'(t)\|_\infty \leq \bar{\alpha}^t(t+\alpha^{-1})\bar{d} + \bar{d} \leq 2\alpha^{-1}\bar{d} = u_\alpha \quad \text{and}\\ &\|\boldsymbol{z}_\alpha(t)\|_\infty \leq \bar{\alpha}^t\|\boldsymbol{z}(t)\| + \|\boldsymbol{z}'(t)\|_\infty \leq \bar{a}\bar{\alpha}^t(\alpha^{-1} + t + L_n)\bar{d} + \bar{a}\bar{d} \leq 3\alpha^{-1}L_n\bar{a}\bar{d} = z_\alpha.\end{aligned}$$

Therefore, $\boldsymbol{a}_\alpha(t) \in [0, y_\alpha]^m \times [0, z_\alpha]^n$, implying that $\pi_\alpha \in \Pi_\alpha$ is an admissible policy.

Finally, we can expand the expression of V^π and $V_\alpha^{\pi_\alpha}$ by

$$V^\pi(\boldsymbol{s}; F) = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t C(\boldsymbol{s}(t), \boldsymbol{a}(t))\Big] \quad \text{and} \quad V_\alpha^{\pi_\alpha}(\boldsymbol{s}; F) = \mathbb{E}\Big[\sum_{t=0}^{\infty}\gamma^t C(\boldsymbol{s}_\alpha(t), \boldsymbol{a}_\alpha(t))\Big].$$

We can compute that

$$\begin{aligned}C(\boldsymbol{s}_\alpha(t), \boldsymbol{a}_\alpha(t)) =&\bar{\alpha}^t C(\boldsymbol{s}(t), \boldsymbol{a}(t)) + \boldsymbol{b}^\top\sum_{s=t}^{t+L_n}\alpha^{t-s}\boldsymbol{y}'(s)\\ \leq&C(\boldsymbol{s}(t), \boldsymbol{a}(t)) + \bar{c}\bar{d}m(L_n+1)(1-\bar{\alpha}^t).\end{aligned}$$

Here, the last inequality utilizes the fact that $\boldsymbol{y}'(s) = 0$ for $s < L_n$ and

$$\bar{\alpha}^{t-s}\boldsymbol{y}'(s) = (\bar{\alpha}^{L_n+t-s} - \bar{\alpha}^t)\boldsymbol{d}(s-L_n) \leq (1-\bar{\alpha}^t)\mathbf{1},$$

for $s \geq L_n$. Therefore,

$$V_\alpha^{\pi_\alpha}(\boldsymbol{s}; F) \leq V^\pi(\boldsymbol{s}; F) + \bar{c}\bar{d}m(L_n+1)\sum_{t\geq 0}^{\infty}\gamma^t(1-\bar{\alpha}^t) \leq V^\pi(\boldsymbol{s}; F) + \frac{\alpha\bar{c}\bar{d}m(L_n+1)}{(1-\gamma)^2},$$

giving (46) and finishing the proof. □

D.7 The Covering Arguments

We finally prove Lemma 7 using covering arguments. For function $v : \mathcal{S}_\alpha \to \mathbb{R}$, Markovian policy ξ, and $\boldsymbol{s} \in \mathcal{S}$, let

$$\mathcal{P}^\xi v(\boldsymbol{s}) := \mathbb{E}_{\boldsymbol{d}}[v(f_\alpha(\boldsymbol{s}, \xi(\boldsymbol{s}), \boldsymbol{d}))], \quad \text{and} \quad \widehat{\mathcal{P}}^\xi v(\boldsymbol{s}) := \mathbb{E}_{\hat{\boldsymbol{d}}}[v(f_\alpha(\boldsymbol{s}, \xi(\boldsymbol{s}), \hat{\boldsymbol{d}}))], \tag{49}$$

Clearly, $\mathcal{P}^\xi$ and $\widehat{\mathcal{P}}^\xi$ are state-transition operator with respect to Markovian policy ξ.

Lemma 10. *For Markovian policy ξ and function $v : \mathcal{S}_\alpha \to \mathbb{R}$ such that $\|v\|_\infty < \infty$ it holds that*

$$\|v\|_\infty \le (1-\gamma)^{-1}\|(I - \gamma\mathcal{P}^\xi)v\|_\infty \quad \textit{and} \quad \|v\|_\infty \le (1-\gamma)^{-1}\|(I - \gamma\widehat{\mathcal{P}}^\xi)v\|_\infty \tag{50}$$

where I is the identity operator, i.e., $Iv = v$ for all v. Moreover, if $\pi \in \Pi$ is the induced policy by ξ, then

$$(I - \gamma\mathcal{P}^\xi)V^\pi(\boldsymbol{s}; F) = C(\boldsymbol{s}, \xi(\boldsymbol{s})) = (I - \gamma\widehat{\mathcal{P}}^\xi)V^\pi(\boldsymbol{s}; \widehat{F}), \quad \textit{for all } \boldsymbol{s} \in \mathcal{S}_\alpha. \tag{51}$$

Proof. Proof of Lemma 10 We only provide the proof for operator $\mathcal{P}^\xi$. The results for operator $\widehat{\mathcal{P}}^\xi$ follow from similar arguments. It follows from (49) that $\|\mathcal{P}^\xi v\|_\infty \le \|v\|_\infty$. Therefore,

$$\|v\|_\infty = \|(I - \gamma\mathcal{P}^\xi)v + \gamma\mathcal{P}^\xi v\|_\infty \le \|(I - \gamma\mathcal{P}^\xi)v\|_\infty + \gamma\|\mathcal{P}^\xi v\|_\infty \le \|(I - \gamma\mathcal{P}^\xi)v\|_\infty + \gamma\|v\|_\infty.$$

Rearranging terms, $\|v\|_\infty \le (1-\gamma)^{-1}\|(I - \gamma\mathcal{P}^\xi)v\|_\infty$, finishing the proof.

Next, we take the induced policy π and the corresponding controlled state process $\{\boldsymbol{s}_\alpha(t) : t \ge 0\}$ through (23). It holds that

$$\begin{aligned}
V_\alpha^\pi(\boldsymbol{s}; F) &= \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t C\Big(\boldsymbol{s}_\alpha(t),\ \xi(\boldsymbol{s}_\alpha(t))\Big)\Big|\boldsymbol{s}_\alpha(0) = \boldsymbol{s}\Big] \\
&= C(\boldsymbol{s}, \xi(\boldsymbol{s})) + \mathbb{E}\Big[\sum_{t=1}^{\infty} \gamma^t C\Big(\boldsymbol{s}_\alpha(t),\ \xi(\boldsymbol{s}_\alpha(t))\Big)\Big|\boldsymbol{s}_\alpha(1) = f_\alpha(\boldsymbol{s}, \xi(\boldsymbol{s}), \boldsymbol{d})\Big] \\
&= C(\boldsymbol{s}, \xi(\boldsymbol{s})) + \gamma\mathbb{E}_{\boldsymbol{d}}[V_\alpha^\pi(f_\alpha(\boldsymbol{s}, \xi(\boldsymbol{s}), \boldsymbol{d}); F)] \\
&= C(\boldsymbol{s}, \xi(\boldsymbol{s})) + \gamma\mathcal{P}^\xi V_\alpha^\pi(\boldsymbol{s}; F),
\end{aligned}$$

which concludes the proof.

□

Proof. Proof of Lemma 7. We first prove (30). Rearranging term in (51) gives

$$(I - \gamma\mathcal{P}^\xi)(V_\alpha^\pi(\boldsymbol{s}; F) - V_\alpha^\pi(\boldsymbol{s}; \widehat{F})) = \gamma(\mathcal{P}^\xi - \widehat{\mathcal{P}}^\xi)V_\alpha^\pi(\boldsymbol{s}; \widehat{F}), \quad \text{for all } \boldsymbol{s} \in \mathcal{S}.$$

Using Lemma 10, we obtain

$$\|V_\alpha^\pi(\,\cdot\,; F) - V_\alpha^\pi(\,\cdot\,; \widehat{F})\|_\infty \le \frac{\gamma}{1-\gamma}\|(\mathcal{P}^\xi - \widehat{\mathcal{P}}^\xi)V_\alpha^\pi(\,\cdot\,; \widehat{F})\|_\infty. \tag{52}$$

To bound the right-hand side of (52), it follows from (49) that

$$\begin{aligned}
0 \le \mathcal{P}^\xi(V_\alpha^\pi(\boldsymbol{s}; \widehat{F}) - V_\alpha^*(\boldsymbol{s}; \widehat{F})) \le \big\|V_\alpha^\pi(\boldsymbol{s}; \widehat{F}) - V_\alpha^*(\boldsymbol{s}; \widehat{F})\big\|_\infty, \quad \text{and} \\
0 \le \widehat{\mathcal{P}}^\xi(V_\alpha^\pi(\boldsymbol{s}; \widehat{F}) - V_\alpha^*(\boldsymbol{s}; \widehat{F})) \le \big\|V_\alpha^\pi(\boldsymbol{s}; \widehat{F}) - V_\alpha^*(\boldsymbol{s}; \widehat{F})\big\|_\infty.
\end{aligned}$$

Subtracting the two inequalities and taking supreme over $\boldsymbol{s} \in \mathcal{S}_\alpha$,

$$\|(\mathcal{P}^\xi - \widehat{\mathcal{P}}^\xi)(V_\alpha^\pi(\,\cdot\,; \widehat{F}) - V_\alpha^*(\,\cdot\,; \widehat{F}))\|_\infty \le \big\|V_\alpha^\pi(\boldsymbol{s}; \widehat{F}) - V_\alpha^*(\boldsymbol{s}; \widehat{F})\big\|_\infty.$$

As a result,

$$\begin{aligned}
\|(\mathcal{P}^\xi - \widehat{\mathcal{P}}^\xi)V^\pi(\,\cdot\,; \widehat{F})\|_\infty \le& \|(\mathcal{P}^\xi - \widehat{\mathcal{P}}^\xi)(V^\pi(\,\cdot\,; \widehat{F}) - V^*(\,\cdot\,; \widehat{F}))\|_\infty + \|\gamma(\mathcal{P}^\xi - \widehat{\mathcal{P}}^\xi)V^*(\,\cdot\,; \widehat{F})\|_\infty \\
=& \big\|V_\alpha^\pi(\boldsymbol{s}; \widehat{F}) - V_\alpha^*(\boldsymbol{s}; \widehat{F})\big\|_\infty + \sup_{g \in \mathcal{G}} \big|\mathbb{E}_{\boldsymbol{d}}[g(\boldsymbol{d})] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g(\hat{\boldsymbol{d}})]\big|,
\end{aligned} \tag{53}$$

where the function class $\mathcal{G}$ is characterizes by all $g : \mathcal{D} \to \mathbb{R}$ such that

$$g(\,\cdot\,) = \gamma V_\alpha^*(f(\boldsymbol{s}, \boldsymbol{a},\ \cdot\ ); \widehat{F}) \text{ for some } (\boldsymbol{s}, \boldsymbol{a}) \in \mathcal{SA}.$$

Temporarily take $L_v := (1-\gamma)^{-1}\bar{c}(\bar{a}+1)(m + nL_n)$. It follows from Corollary 1 (taking $G = \widehat{F}$) that $V^*(\,\cdot\,; \widehat{F})$ is L_v-Lipschitz continuous. As f is $(\bar{a}+1)$-Lipschitz continuous, every element in $\mathcal{G}$ is a $(\bar{a}+1)L_v$-Lipschitz continuous function. Take

$$N_0 := \sqrt{2M}(\bar{a}+1)/\bar{d} \quad \text{and} \quad N = \lceil x_\alpha N_0 + 1\rceil^{p_S + p_A},$$

where $\lceil x \rceil$ is the ceiling operator that returns the smallest integer that compares no less than x, for $x \in \mathbb{R}$. Let the finite set $\{(\boldsymbol{s}_i, \boldsymbol{a}_i) : i = 1, 2, \cdots, N\}$ form a lattice on $\mathcal{SA}_\alpha \subseteq [0, x_\alpha]^{p_S+p_A}$ with unit length N_0^{-1}. In particular, $\{(\boldsymbol{s}_i, \boldsymbol{a}_i) : i = 1, 2, \cdots, N\}$ is an N_0^{-1}-covering of $\mathcal{SA} \subseteq [0, x_\alpha]^{p_S+p_A}$ with respect to the L^∞ norm, i.e.,

$$\sup_{(\boldsymbol{s},\boldsymbol{a})\in\mathcal{SA}} \inf_{i\in[N]} \|(\boldsymbol{s}, \boldsymbol{a}) - (\boldsymbol{s}_i, \boldsymbol{a}_i)\|_\infty \le N_0^{-1}.$$

Let $g_i : \mathcal{D} \to \mathbb{R}$ such that $g_i(\boldsymbol{d}) := V_\alpha^*(f_\alpha(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{d}); \widehat{F})$, for $i \in [N]$, the $(\bar{a}+1)L_v$-Lipschitz continuity of the function $(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d}) \mapsto V_\alpha^*(f_\alpha(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d}); \widehat{F})$ gives

$$\sup_{g\in\mathcal{G}} \inf_{i\in[N]} \|g - g_i\|_\infty = \sup_{(\boldsymbol{s},\boldsymbol{a},\boldsymbol{d})\in\mathcal{SA}\times\mathcal{D}} \inf_{i\in[N]} |V_\alpha^*(f_\alpha(\boldsymbol{s}, \boldsymbol{a}, \boldsymbol{d}); \widehat{F}) - V_\alpha^*(f_\alpha(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{d}); \widehat{F})| \le (\bar{a}+1)L_v/N_0.$$

Therefore, for any individual function $g \in \mathcal{G}$, there exists an $i \in [N]$ such that

$$\|g_i - g\|_\infty \le (\bar{a}+1)L_v/N_0.$$

In particular,

$$\Big|\mathbb{E}_{\boldsymbol{d}}[g(\boldsymbol{d}) - g_i(\boldsymbol{d})]\Big| \le (\bar{a}+1)L_v/N_0 \quad \text{and} \quad \Big|\mathbb{E}_{\hat{\boldsymbol{d}}}[g_i(\hat{\boldsymbol{d}}) - g(\hat{\boldsymbol{d}})]\Big| \le (\bar{a}+1)L_v/N_0,$$

implying that

$$\begin{aligned}\Big|\mathbb{E}_{\boldsymbol{d}}\left[g(\boldsymbol{d})\right] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g(\hat{\boldsymbol{d}})]\Big| \le& \Big|\mathbb{E}_{\boldsymbol{d}}[g(\boldsymbol{d}) - g_i(\boldsymbol{d})]\Big| + \Big|\mathbb{E}_{\boldsymbol{d}}\left[g_i(\boldsymbol{d})\right] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g_i(\hat{\boldsymbol{d}})]\Big| + \Big|\mathbb{E}_{\hat{\boldsymbol{d}}}[g_i(\hat{\boldsymbol{d}}) - g(\hat{\boldsymbol{d}})]\Big| \\ \le& 2(\bar{a}+1)L_v/N_0 + \Big|\mathbb{E}_{\boldsymbol{d}}\left[g_i(\boldsymbol{d})\right] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g_i(\hat{\boldsymbol{d}})]\Big|.\end{aligned}$$

Taking supreme over $g \in \mathcal{G}$, we have shown that

$$\sup_{g\in\mathcal{G}} \Big|\mathbb{E}_{\boldsymbol{d}}\left[g(\boldsymbol{d})\right] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g(\hat{\boldsymbol{d}})]\Big| \le 2(\bar{a}+1)L_v/N_0 + \sup_{i\in[N]} \Big|\mathbb{E}_{\boldsymbol{d}}\left[g_i(\boldsymbol{d})\right] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g_i(\hat{\boldsymbol{d}})]\Big|.$$

Next, we expand the second term on the right-hand side by

$$\mathbb{E}_{\boldsymbol{d}}\left[g_i(\boldsymbol{d})\right] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g_i(\hat{\boldsymbol{d}})] = \sum_{k=1}^{M} \Big(g_i(\boldsymbol{d}_k) - \mathbb{E}_{\boldsymbol{d}}[g_i(\boldsymbol{d})]\Big),$$

which, for each $i \in [N]$, is the sum of M copies of mean zero i.i.d. random variables that are bounded by

$$\begin{aligned}&\sup_{\boldsymbol{d}_1\in\mathcal{D}} \big|g_i(\boldsymbol{d}_1) - \mathbb{E}[g_i(\boldsymbol{d})]\big| \\ =& \sup_{\boldsymbol{d}_1\in\mathcal{D}} \big|V_\alpha^*(f_\alpha(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{d}_1), \widehat{F}) - \mathbb{E}_{\boldsymbol{d}}[V_\alpha^*(f_\alpha(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{d})], \widehat{F})\big| \\ \le& \sup_{\boldsymbol{d}_1\in\mathcal{D}} \mathbb{E}_{\boldsymbol{d}}\Big[\big|V_\alpha^*(f_\alpha(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{d}_1), \widehat{F}) - V_\alpha^*(f_\alpha(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{d}), \widehat{F})\big|\Big] \\ \le& L_v \sup_{\boldsymbol{d}_1\in\mathcal{D}} \mathbb{E}_{\boldsymbol{d}}\Big[\big\|f_\alpha(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{d}_1) - f_\alpha(\boldsymbol{s}_i, \boldsymbol{a}_i, \boldsymbol{d})\big\|_\infty\Big] \le L_v\bar{d}.\end{aligned}$$

Temporarily take $B_g := L_v\bar{d}$. For each $i \in [N]$, it follows from Hoeffing's inequality, (see, for example, Proposition 2.5 and (2.11) in [14]),

$$\mathbb{P}\Big(\Big|\mathbb{E}_{\boldsymbol{d}}\left[g_i(\boldsymbol{d})\right] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g_i(\hat{\boldsymbol{d}})]\Big| \ge \delta_0\Big) \le \exp\big(\frac{-2M\delta_0^2}{B_g^2}\big),$$

for all $\delta_0 > 0$. Here $\mathbb{P}$ is taken with respect to $\widehat{D}$. Taking $\delta = N\exp(-2B_g^{-2}M\delta_0^2)$,

$$\sup_{i\in[N]} \Big|\mathbb{E}_{\boldsymbol{d}}\left[g_i(\boldsymbol{d})\right] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g_i(\hat{\boldsymbol{d}})]\Big| \le \delta_0 = M^{-1/2}B_g(\log(N\delta^{-1})/2)^{1/2}$$

with probability at least $1-\delta$. Together with (52) and (53), we have shown that the event

$$\begin{aligned}\|V_\alpha^\pi(\,\cdot\,; F) - V_\alpha^\pi(\,\cdot\,; \widehat{F})\|_\infty \le& \frac{\gamma}{1-\gamma}\Big(\|V_\alpha^\pi(\,\cdot\,; \widehat{F}) - V_\alpha^*(\,\cdot\,; \widehat{F})\|_\infty \\ &+ 2(\bar{a}+1)L_v/N_0 + (2M)^{-1/2}B_g(\log(N\delta^{-1}))^{1/2}\Big)\end{aligned} \tag{54}$$

holds with probability at least $1-\delta$.

Finally, the choice of N_0 gives $(\bar{a}+1)L_v/N_0 \le (2M)^{-1/2}B_g$. It follows from the choice of x_α that $x_\alpha \ge \bar{d}$, and thus $N_0x_\alpha \ge \sqrt{2M}(\bar{a}+1) \ge 2\sqrt{2}$. For $\delta \in (0,1)$, we have

$$\log(N\delta^{-1}) \ge \log(N_0x_\alpha + 1) \ge \log(2\sqrt{2}+1) \ge 1,$$

giving

$$(\bar{a}+1)L_v/N_0 \le (2M)^{-1/2} B_g (\log(N\delta^{-1}))^{1/2}.$$

We also have

$$N_0 x_\alpha + 2 \le x_\alpha(\sqrt{2M}(\bar{a}+1)+2)/\bar{d} \le 3(\bar{a}+1)M/\bar{d},$$

giving

$$\log(\delta^{-1}N) \le (p_S+p_A)\log(\delta^{-1}(N_0 x_\alpha + 2)) \le (p_S+p_A)\log\frac{3(\bar{a}+1)Mx_\alpha}{\delta\bar{d}}$$

Therefore, the "constant" terms in the right-hand side of (54) can be bounded by

$$\begin{aligned}
&\frac{\gamma}{1-\gamma}\Big(2(\bar{a}+1)L_v/N_0 + (2M)^{-1/2}B_g(\log(N\delta^{-1}))^{1/2}\Big) \\
\le &\frac{\gamma}{1-\gamma}\Big(3L_v\bar{d}\sqrt{\frac{p_S+p_A}{2M}\log\frac{3M(\bar{a}+1)x_\alpha}{\delta\bar{d}}}\Big) \\
\le &\frac{3(\bar{a}+1)\bar{c}p_S}{(1-\gamma)^2}\sqrt{\frac{p_S+p_A}{2M}\log\frac{3M(\bar{a}+1)x_\alpha}{\delta\bar{d}}} = \mathcal{E}_1(\delta).
\end{aligned}$$

Together with (54), we obtain (52).

The proof of (31) is similar. Using Lemma 10, similar computation to (53) gives

$$\|V_\alpha^\pi(\,\cdot\,;F) - V_\alpha^\pi(\,\cdot\,;\widehat{F})\|_\infty \le \frac{\gamma}{1-\gamma}\big\|V_\alpha^\pi(\boldsymbol{s};\widehat{F}) - V_\alpha^*(\boldsymbol{s};\widehat{F})\big\|_\infty + \frac{1}{1-\gamma}\sup_{g\in\mathcal{G}'}\big|\mathbb{E}_{\boldsymbol{d}}[g(\boldsymbol{d})] - \mathbb{E}_{\hat{\boldsymbol{d}}}[g(\hat{\boldsymbol{d}})]\big|,$$

where the class $\mathcal{G}'$ is characterized by all functions $g : \mathcal{D} \to \mathbb{R}$ such that

$$g(\,\cdot\,) = \gamma V_\alpha^*(f(\boldsymbol{s},\boldsymbol{a},\,\cdot\,);F) \text{ for some } (\boldsymbol{s},\boldsymbol{a}) \in \mathcal{SA}.$$

Taking $G = F$ in Corollary 1, $V_\alpha^*(\,\cdot\,;F)$ is L_v-Lipschitz continuous on $\mathcal{S}_\alpha$. Again, similar arguments to the proof of (52) give (31). □

E Numerical Results

Hyperparameters	Value
Batch size	2^8
Buffer size	2^{12}
Soft update	2^{-8}
Reward scale	10^{-2}
Learning rate	$lr \in [10^{-5}, 10^{-3}]$
Network width	2^8
Delayed update	16
Learning Discount	0.95
Number of layers	3

Table 1: Hyperparameter settings for RL algorithms.

Example 1. *(*N-SYSTEM) There are two components and two products in the N-system. Assembling one unit of product 1 requires one unit of component 1. Assembling one unit of product 2 requires one unit of component 1 and one unit of component 2. The inventory holding and backorder costs are $(h_1, h_2) = (2, 3)$ and $(b_1, b_2) = (9, 6)$, respectively. The component lead times are $L_1 = L_2 = 1$. Product demands d_1 and d_2 have marginal uniform distributions on [0, 40]. We separately consider two scenarios where d_1 and d_2 are independent and correlated. When d_1 and d_2 are independent, it is known that the optimal ordering policy follows a base-stock policy with base-stock levels $s_1 = 51.0$ and $s_2 = 23.9$ (see [10]). When d_1 and d_2 are correlated, we consider that the conditional distribution of $[d_2|d_1]$ is $40 - d_1$ with probability 0.9 and an independent uniform random variable on [0,40] with probability 0.1. In particular, the two products become substitutes. Table 2 summarizes the numerical performance of the three policies in different sample sizes.

	Algorithm	Demand Sample Size		
		$M = 10$	$M = 100$	$M = 200$
		C (s.d.)	C (s.d.)	C (s.d.)
I.I.D. Demand (Optimality: 81.0)	CTD3	86.9 (0.04)	82.9 (0.06)	82.2 (0.04)
	PTD3	87.4 (0.04)	84.6 (0.05)	83.0 (0.04)
	NV-PRP	86.1 (0.05)	84.7 (0.03)	82.9 (0.04)
Correlated Demand	CTD3	66.1 (0.03)	60.2 (0.05)	58.1 (0.02)
	PTD3	67.3 (0.03)	62.6 (0.04)	61.1 (0.05)
	NV-PRP	79.2 (0.07)	72.8 (0.04)	68.2 (0.08)

Table 2: The performances of CTD3 policy, PTD3 policy, and NV-PRP policy on the N-system with different demand sample sizes and demand joint distributions.

From Table 2, we first observe that all three methods generate near-optimal policies when the demands of the two products are independent. With $M = 10$, the relatively high optimality gap of the two RL algorithms likely results from an overfit of the 10 observed demands. Note that the effect of overfitting is analyzed in [15]. As M increases, the performance of RL algorithms improves substantially, where the CTD3 algorithm generates a policy with 1.4% (82.2 compared to 81.0) optimality gap. Table 2 also suggests that the CTD3 policy outperforms the PTD3 policy by a statistically significant margin. To this end, Table 2 suggests that RL algorithms are capable of generating near-optimal policies with a relatively small sample of demands.

With independent demand, the near-optimal performance of the NV-PRP policy is expected from its design. Compared to the RL algorithms, the heuristic relies only on the mean demand and, therefore, does not overfit the small sample as much. From Table 2, the NV-PRP policy outperforms the two RL policies by a statistically significant margin when $M = 10$. However, due to the suboptimality of the NV-PRP policy, it is eventually outperformed by the two RL policies as M grows. With correlated demands, the optimal policy is unavailable, and the NV-PRP policy is expected to be costly (see [7]). Indeed, the policies from the two RL algorithms substantially outperform NV-PRP, with 17% cost reduction when $M = 10$ and around 14% reduction when $M = 200$. As a result, our numerical findings suggest that RL algorithms are robust to demand correlation, a common feature that practical ATO systems share.

For a more detailed comparison between the CTD3 algorithm and the PTD3 algorithm, Figure 1 (left and middle) summarizes the convergence rate of the algorithms with different choices of sample size M. As both CTD3 and PTD3 are iterative algorithms, the comparison is conducted for each iteration count K. Figure 1 (left and middle) shows that the convergence rate of the CTD3 algorithm is significantly faster than that of the PTD3 algorithm, for both $M = 10$ and $M = 200$, evidencing an improved computational efficiency from incorporating the shape constraint into the neural network architecture.

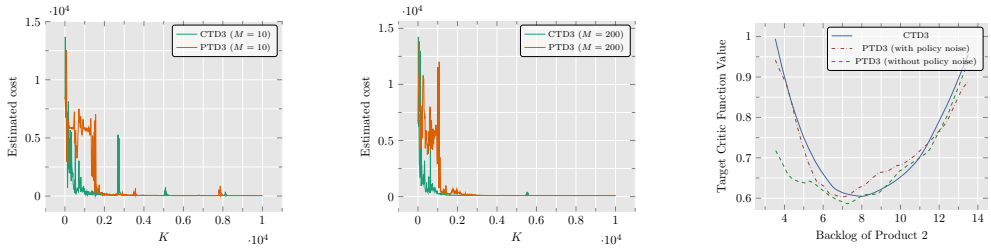


Figure 1: (Color online) Convergence of CTD3 and PTD3 algorithms (left and middle). A slice of actor-value function (right).

The acceleration of convergence can be partly illustrated by Figure 1 (right), which plots a slice of the actor-value function with different b_2, i.e., the backlog of product 2, while keeping the other state variables fixed. In the figure, the actor-value function computed by the PTD3 algorithm is not convex and is less smooth compared to that of the CTD3 algorithm. Indeed, the policy optimization in the PTD3 algorithm depends on the policy noise, a hyperparameter that smooths the computed value function, so that the optimization is less likely to be stuck in local minima, at the cost of an additional source of error. When policy noise is set to 0, the actor-value function is even less smooth compared to the default value of 0.01. In contrast, the actor-value functions are guaranteed to be convex for the CTD3 algorithm, so that the policy noise can safely be set to 0. To this end, CTD3 delivers a better fit of the actor-value functions compared to PTD3, partially contributing to convergence acceleration.

Sensitivity Analysis. Next, we consider how the performance of the two RL algorithms depends on two important tuning factors, the learning rate and the network width. In particular, the learning rate has the same effect on the CTD3 and PTD3 algorithms as the step size affects the convergence of a gradient descent iteration, i.e., a higher learning rate speeds up the initial convergence of the algorithms at the cost of numerical stability. The network width determines the neural network's size, namely, the number of parameters it contains, with the default network set to 256. For a fixed sample size $M = 200$ and independent demand, the simulated cost and the optimality gap $\Delta = [C(\pi) - C(\pi^*)]/C(\pi^*)$ are summarized in Table 3.

Algorithm	Learning Rate	Network Width (NW)					
		NW=128		NW=256		NW=512	
		C (s.d.)	Δ (%)	C (s.d.)	Δ (%)	C (s.d.)	Δ (%)
CTD3	lr = 5e-5	84.9 (0.02)	4.60	84.6 (0.03)	4.22	85.1 (0.04)	4.84
	lr =1e-4	82.3 (0.04)	1.53	82.9 (0.03)	2.27	83.1 (0.02)	2.53
	lr = 5e-4	84.8 (0.02)	4.51	82.7 (0.05)	2.07	82.7 (0.05)	2.06
	lr =1e-3	82.9 (0.03)	2.34	82.2 (0.04)	1.44	82.2 (0.03)	1.45
PTD3	lr =5e-5	86.2 (0.03)	6.01	86.5 (0.03)	6.32	85.9 (0.05)	5.70
	lr =1e-4	85.2 (0.02)	4.92	84.9 (0.03)	4.62	83.8 (0.04)	3.39
	lr =5e-4	83.7 (0.05)	3.27	83.0 (0.04)	2.41	83.1 (0.03)	2.53
	lr =1e-3	83.3 (0.02)	2.76	84.0 (0.04)	3.55	85.5 (0.04)	5.22

Table 3: Hyperparameter tuning for CTD3 and PTD3 algorithms.

From Table 3, we conclude that the performances of the CTD3 and PTD3 algorithms do not significantly depend on the network width, where both algorithms output near-optimal policies. Hence, in subsequent experiments, we keep this parameter at its default value. On the other hand, the performance of the PTD3 policy depends on the choice of learning rate when the network width is 128 or 256, while the same dependence of the CTD3 policy is not significant.

Example 2. (PC-SYSTEM) This example is a simplified PC system with six components and four products considered in [11]. The demand for product $i \in \{1, 2, 3, 4\}$ follows an independent Poisson distribution of intensity λ_i. We present the BOM matrix and related parameters of this medium-scale PC system in the appendix.

We start with the comparison of the cost among policies obtained from the CTD3 algorithm, the PTD3 algorithm, and the NV-PRP policy, and summarize the results for various sample sizes M by Figure 2. We first observe that the cost of the CTD3 policy again decreases as M increases from 10 to 1000. The decrease rate is much faster when $M \leq 200$, supporting our result in Theorem 3 that the effective learning of the optimal policy π^* does not require a very large demand sample. Second, Figure 2 (left) indicates that the CTD3 policy consistently achieves a substantially lower cost among all sample sizes compared to the PTD3 policy and the NV-PRP policy for any choice of sample size. For example, when $M = 100$, the cost reduction of the CTD3 policy is 3% compared to the PTD3 policy and 9% compared to the NV-PRP policy.

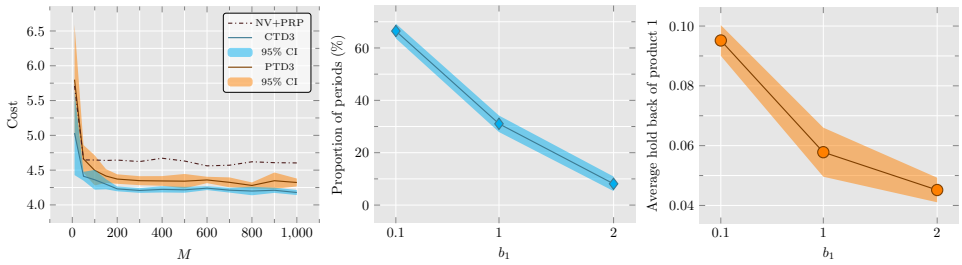


Figure 2: (Color online) Performances of CTD3 and PTD3 algorithms (left). Strategical holdback of product 1 (middle and right).

To analyze the effects of hyperparameter tuning, Table 4 presents the average and standard deviation of the simulated cost for both the PTD3 and CTD3 algorithms under various learning rates from {1e-5, 5e-5, 1e-4, 5e-4} and sample sizes $M \in \{10, 500, 1000\}$. We find that for all sample sizes, the CTD3 algorithm outperforms the PTD3 algorithm. Furthermore, the CTD3 algorithm converges for all learning rates, whereas the PTD3 algorithm fails to converge at lr =1e-5. The findings, therefore, support our insight that geometric properties of the ATO control problem help improve the performance, as well as the robustness of the general-purpose PTD3 algorithm.

Algorithm	Learning Rate	Demand Sample Size		
		$M = 10$	$M = 500$	$M = 1000$
		C (s.d.)	C (s.d.)	C (s.d.)
CTD3	lr =1e-5	5.78 (0.03)	4.22 (0.03)	4.35 (0.02)
	lr =5e-4	5.03 (0.04)	4.36 (0.02)	4.18 (0.03)
	lr =1e-4	5.26 (0.03)	4.52 (0.03)	4.33 (0.03)
	lr =5e-4	5.43 (0.03)	4.28 (0.03)	4.29 (0.03)
PTD3	lr =1e-5	-*	-	-
	lr =5e-5	5.93 (0.05)	4.90 (0.05)	4.32 (0.03)
	lr =1e-4	5.80 (0.03)	4.34 (0.03)	4.81 (0.03)
	lr =5e-4	6.15 (0.05)	5.47 (0.06)	4.76 (0.04)

* -: Algorithm fails to converge.

Table 4: Learning-rate tuning of PC-system.

Performance Analysis of the No-holdback Rule. The output policy of the highly efficient CTD3 algorithm facilitates a numerical analysis of the suboptimality of the no-holdback rule. During the 10,000-period simulation of the CTD3 policy with sample size $M = 1000$, the decisions intriguingly made in 812 periods violate the no-holdback rule by withholding the production of product 1. Indeed, by strategically withholding components required to assemble product 1, the system can save components 2 and 5 for future assembly of more expensive products. To illustrate this effect, we take $b_1 \in \{0.1, 1, 2\}$ while keeping the other parameters unchanged. Figure 2 (middle and right) summarizes the holdback of product 1, which is averaged over 100 simulation runs, accompanied by a 95% confidence interval by the shaded areas. In particular, the proportion of periods with holdback increased to 31.04% when b_1 decreased from 2 to 1 and to 66.46% when $b_1 = 0.1$. To this end, our numerical results support the observation that, when holding costs are relatively low, policies that involve strategic holdback may be more cost-effective than no-holdback policies.

Example 3. (LARGE SYSTEM) We finally analyze the performance of the CTD3 algorithm on the ATO system from the processed food company. There are 15 components and 8 products, where Table 5 provides the BOM matrix, lead times, and other cost parameter settings. As discussed, the actual demand data in this example contains 100 monthly demand points for each end product. As a result, the performance metric $C(\pi)$ is unknown. We report the performance of the PTD3 algorithm, CTD3 algorithms, and the NV-PRP policy by $\widehat{C}(\pi)$.

From Table 6, the CTD3 policy (tuned) yields an average cost of 32.0, outperforming the NV-PRP policy by 11.6%. We also observe that the CTD3 algorithm can achieve satisfying policies under the learning rates from {5e-5, 1e-4, 5e-4}, while the PTD3 algorithm fails to deliver a substantial cost reduction under all choices of the learning rate. It is essential to reiterate that the only difference between PTD3 and CTD3 lies in the network architecture, ICNN versus a general-purpose dense neural network. To this end, our numerical results suggest that the utility of ICNN, which originates from the convex structure of the ATO transition dynamics, becomes more evident as the size of the ATO control problem increases. In the N-system, the effect of ICNN is limited to convergence rate acceleration. For the slightly larger PC system, the ICNN starts to deliver better policies

Component	h_j	L_j	Product							
			1	2	3	4	5	6	7	8
1	0.163	2	1	1	1	0	0	0	0	0
2	0.318	1	0	0	0	1	1	1	0	0
3	0.104	1	0	0	0	0	0	0	1	0
4	0.206	2	0	0	0	0	0	0	0	1
5	0.167	1	1	1	0	0	0	0	0	0
6	0.575	1	0	0	1	1	1	0	0	0
7	0.126	1	0	0	0	0	0	1	1	0
8	0.166	1	0	0	0	0	0	0	0	1
9	0.275	1	1	0	1	0	0	0	0	0
10	0.303	2	0	1	0	0	0	0	0	0
11	0.191	1	0	0	0	1	0	1	0	0
12	0.307	1	0	0	0	0	0	0	1	1
13	0.473	1	0	0	0	0	1	0	0	0
14	0.239	1	1	1	0	0	0	0	0	0
15	0.129	3	1	0	0	1	0	1	0	0

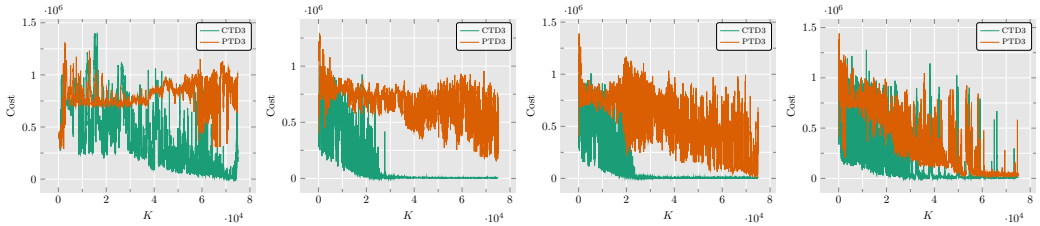
Table 5: Large scale PC example.

compared to dense neural networks. The ICNN eventually becomes the critical factor for an RL algorithm to generate cost-efficient policies that outperform the benchmark policies.

	PTD3	CTD3	NV-PRP
	$\hat{C}$ (s.d.)	$\hat{C}$ (s.d.)	$\hat{C}$ (s.d.)
lr =1e-5	-	50.0 (0.16)	36.2 (0.14)
lr =5e-5	-	32.1 (0.05)	
lr =1e-4	77.7 (0.09)	32.0 (0.04)	
lr =5e-4	118.7 (0.07)	33.2 (0.06)	

Table 6: Learning-rate tuning of processed food system.

Figure 3 compares the convergence of CTD3 and PTD3 under different learning rates. We observe that, as the learning rate becomes larger, the initial convergence of the CTD3 algorithm becomes faster but less stable. However, the learning curve of the PTD3 algorithm oscillates in the order of 10^6 throughout the iteration process under every choice of the learning rate, where the output policy significantly underperforms that of the CTD3 algorithm. We remark that training neural networks is demanding in computational resources, emphasizing the value of effective algorithmic design. To this end, our numerical results underscore the significance of integrating structural insights into RL algorithms, as seen in architectural variations like ICNN, which becomes increasingly valuable when confronted with intricate real-world control problems.



(a) lr =1e-5 (b) lr =5e-5 (c) lr =1e-4 (d) lr =5e-4

Figure 3: (Color online) Convergence of CTD3 and PTD3 algorithms in the processed food system.